A Bridge between Short-Range and Seasonal Forecasts: Data-Based First Passage Time Prediction in Temperatures

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt

der Fakultät Mathematik und Naturwissenschaften
der Technischen Universität Dresden

von

Dipl.-Phys. Anja Karen von Wulffen,
geb. Garber am 15.11.1984 in Berlin

Eingereicht am 15.11.2012
Verteidigt am 25.01.2013

Die Dissertation wurde von November 2008 bis November 2012
im Max-Planck-Institut für Physik komplexer Systeme angefertigt.

Acknowledgements

Many thanks to Prof. Dr. Holger Kantz for supervising this thesis and providing helpful suggestions, ideas and encouragement, as well as for finding time for long and insightful discussions.

I am very grateful to Stefan Siegert for generating first passage time forecasts from dynamical model ensemble output including sophisticated post-processing to allow a direct forecast skill comparison between my work and global circulation models (Sec. 7.2), for helping to develop an appropriate benchmark for the full probabilistic first passage time forecasts, as well as for providing interesting discussions of the actual state of dynamical weather models and their inherent uncertainties and errors, as well as the intricacies of forecast scoring.

I thank both Dr. Reik Donner and Jonathan Donges from the Potsdam Institute of Climate Impact Research for directing me to better data and providing some helpful comments on my early work.

I also wish to thank Dr. Nicholas Moloney for many interesting talks and explanations especially related to our joint work on the influence of multiplicative noise, but also on statistical hypothesis tests in general.

I am very grateful to Dr. Dagmar Lackschewitz, Dr. Abigail Klopper and Prof. Dr. Holger Kantz for providing much needed encouragement and convincing me to continue working towards a PhD when my confidence was flagging.

My thanks to the current and past members of the research group on Nonlinear Dynamics and Time Series Analysis at the Max Planck Institute for the Physics of Complex Systems for the friendly and supportive atmosphere and many interesting and fun lunch discussions, as well as to Dr. Abigail Klopper, Dr. Li Chen and Wladimir Tschischik for a nice and quiet but nevertheless supportive office atmosphere.

I am also very grateful to my husband Georg for his willingness to listen to many rants on current problems at all stages of this work, for providing interesting ideas and suggestions especially on the clarity of notation during the writing stage, and to Georg and my parents for their proof-reading of the manuscript and general support.

Referees

1. Gutachter / 1st referee: Prof. Dr. Holger Kantz
2. Gutachter / 2^{d} referee: Prof. Dr. Jürgen Kurths

Abstract

Current conventional weather forecasts are based on high-dimensional numerical models. They are usually only skillful up to a maximum lead time of around 7 days due to the chaotic nature of the climate dynamics and the related exponential growth of model and data initialisation errors. Even the fully detailed medium-range predictions made for instance at the European Centre for Medium-Range Weather Forecasts do not exceed lead times of 14 days, while even longer-range predictions are limited to time-averaged forecast outputs only.

Many sectors would profit significantly from accurate forecasts on seasonal time scales without needing the wealth of details a full dynamical model can deliver. In this thesis, we aim to study the potential of a much cheaper data-based statistical approach to provide predictions of comparable or even better skill up to seasonal lead times, using as an examplary forecast target the time until the next occurrence of frost.

To this end, we first analyse the properties of the temperature anomaly time series obtained from measured data by subtracting a sinusoidal seasonal cycle, as well as the distribution properties of the first passage times to frost. The possibility of generating additional temperature anomaly data with the same properties by using very simple autoregressive model processes to potentially reduce the statistical fluctuations in our analysis is investigated and ultimately rejected.

In a next step, we study the potential for predictability using only conditional first passage time distributions derived from the temperature anomaly time series and confirm a significant dependence of the distributions on the initial conditions. After this preliminary analysis, we issue data-based out-of-sample forecasts for three different prediction targets: The specific date of first frost, the probability of observing frost before summer for forecasts issued in spring, and the full probability distribution of the first passage times to frost.

We then study the possibility of improving the forecast quality first by enhancing the stationarity of the temperature anomaly time series and then by adding as an additional input variable the state of the North Atlantic Oscillation on the date the predictions are issued.

We are able to obtain significant forecast skill up to seasonal lead times when comparing our results to an unskilled reference forecast.

A first comparison between the data-based forecasts and corresponding predictions gathered from a dynamical weather model, necessarily using a lead time of only up to 15 days, shows that our simple statistical schemes are only outperformed (and then only slightly) if further statistical post-processing is applied to the model output.

Zusammenfassung

Aktuelle Wetterprognosen werden mit Hilfe von hochdimensionalen, numerischen Modellen generiert. Durch die dem Klima zugrunde liegende chaotische Dynamik wachsen Modellfehler und Ungenauigkeiten in der Modellinitialisierung exponentiell an, sodass Vorhersagen mit signifikanter Güte üblicherweise nur für eine Vorlaufzeit von maximal sieben Tagen möglich sind. Selbst die detaillierten Prognosen des Europäischen Zentrums für mittelfristige Wettervorhersagen gehen nicht über eine Vorlaufzeit von 14 Tagen hinaus, während noch längerfristigere Vorhersagen auf zeitgemittelte Größen beschränkt sind.

Viele Branchen würden signifikant von akkuraten Vorhersagen auf saisonalen Zeitskalen profitieren, ohne das ganze Ausmaß an Details zu benötigen, das von einem vollständigen dynamischen Modell geliefert werden kann. In dieser Dissertation beabsichtigen wir, am Beispiel einer Vorhersage der Zeitdauer bis zum nächsten Eintreten von Frost zu untersuchen, inwieweit deutlich kostengünstigere, datenbasierte statistische Verfahren Prognosen von gleicher oder sogar besserer Güte auf bis zu saisonalen Zeitskalen liefern können.

Dazu analysieren wir zunächst die Eigenschaften der Zeitreihe der Temperaturanomalien, die aus den Messdaten durch das Subtrahieren eines sinusförmigen Jahresganges erhalten werden, sowie die Charakteristiken der Wahrscheinlichkeitsverteilungen der Zeitdauer bis zum nächsten Eintreten von Frost. Die Möglichkeit, durch einen einfachen autoregressiven Modellprozess zusätzliche Datenpunkte gleicher statistischer Eigenschaften wie der Temperaturanomalien zu generieren, um die statistischen Fluktuationen in der Analyse zu reduzieren, wird untersucht und letztendlich verworfen.

Im nächsten Schritt analysieren wir das Vorhersagepotential, wenn ausschließlich aus den Temperaturanomalien gewonnene bedingte Wahrscheinlichkeitsverteilungen der Wartezeit bis zum nächsten Frost verwendet werden, und können eine signifikante Abhängigkeit der Verteilungen von den Anfangsbedingungen nachweisen. Nach dieser einleitenden Untersuchung erstellen wir datenbasierte Prognosen für drei verschiedene Vorhersagegrößen: Das konkrete Datum, an dem es das nächste Mal Frost geben wird; die Wahrscheinlichkeit, noch vor dem Sommer Frost zu beobachten, wenn die Vorhersagen im Frühjahr ausgegeben werden; und die volle Wahrscheinlichkeitsverteilung der Zeitdauer bis zum nächsten Eintreten von Frost.

Anschließend untersuchen wir die Möglichkeit, die Vorhersagegüte weiter zu erhöhen - zunächst durch eine Verbesserung der Stationarität der Temperaturanomalien und dann durch die zusätzliche Berücksichtigung der Nordatlantischen Oszillation als einer zweiten, den Anfangszustand charakterisierenden Variablen im Vorhersageschema.

Wir sind in der Lage, im Vergleich mit einem naiven Referenzvorhersageschema eine signifikante Verbesserung der Vorhersagegüte auch auf saisonalen Zeitskalen zu erreichen.

Ein erster Vergleich zwischen den datenbasierten Vorhersagen und entsprechenden, aus den dynamischen Wettermodellen gewonnenen Prognosen, der sich notwendigerweise auf eine Vorlaufzeit der Vorhersagen von lediglich 15 Tagen beschränkt, zeigt, dass letztere unsere simplen statistischen Vorhersageschemata nur schlagen (und zwar knapp), wenn der Modelloutput noch einer statistischen Nachbearbeitung unterzogen wird.

Contents

List of Notation and Abbreviations

$\langle \cdot \rangle$	sample mean
$\{ \cdot \}$	data set, time series
$\lvert \cdot \rvert$	absolute value
$\lfloor \cdot \rfloor$	entire value, i.e. largest previous integer
α	significance level of a statistical hypothesis test
ACC	anomaly correlation coefficient
AR(p)-process	autoregressive process of order p
ARMA(m,n)-process	autoregressive moving average process of orders m and n
BIC	modified Akaike (Bayesian) information criterion
BS (BSS)	Brier score (Brier skill score)
$\Delta(\cdot)$	estimated standard error
ΔP	normalised pressure
ΔT	temperature anomaly
$\Delta T^{(3)}$ ($\Delta T^{(10)}$)	temperature anomaly separated into three (ten) categories of equal size by terciles (deciles)
$\Delta T^{(3w,p)}$	temperature anomaly separated into three categories of diferent weights where the outer ones contain p% of all anomalies each
DWD	Deutscher Wetterdienst (German national weather service)
ϵ	residual after fitting
$\mathbb{E}[\cdot]$	expectation value
ECMWF	European Centre for Medium-Range Weather Forecasts
ENSO	El Niño - Southern Oscillation
e.g.	example given
F	cumulative distribution function
f	probability distribution function or frequency
fc	forecast value
γ	coefficients of an AR(p)-model
H_0	null hypothesis of a statistical hypothesis test
H_1	alternative hypothesis of a statistical hypothesis test
IPCC	Intergovernmental Panel on Climate Change
IQR	interquartile range
i.e.	*id est*, that is (Latin)
K	degrees Kelvin
MSE	mean square error
$N(\mu, \sigma)$	normal distribution with mean μ and standard deviation σ
NAO	North Atlantic Oscillation
NAOI	North Atlantic Oscillation Index
NAOI$^{(2)}$	North Atlantic Oscillation indices separated into two roughly equal groups according to their sign
NAOI$^{(3)}$ (NAOI$^{(10)}$)	North Atlantic Oscillation indices separated into three (ten) groups of equal size by terciles (deciles)
NCEP	National Centre for Environmental Prediction
NOAA	National Oceanic and Atmospheric Administration
n	total sample size (or length of a time series)
n_{sig}	number of days for which an effect is statistically significant
$n(\cdot)$	number of data points

obs	observed value, i.e. measurement data point
PC	proportion correct
P	measured pressure
$P(x)$	probability distribution of the variable x
$P(x, y)$	joint probability distribution of the variables x and y
$P(x\|y)$	conditional probability distribution of the variable x given y
$P_{\mathrm{data}}(t_{\mathrm{frost}}\|t_0, \Delta T, y, \mathrm{NAOI})$	first passage time probability distribution to frost gathered from the measured temperature data, conditioned on the initial date, initial temperature anomalies, initial year and initial NAO index
$P_{\mathrm{AR(1)}}(t_{\mathrm{frost}}\|t_0, \Delta T, y, \mathrm{NAOI})$	first passage time distribution to frost gathered from the AR(1) model process, conditioned on the initial date, initial temperature anomalies, initial year and initial NAO index
$P_{\mathrm{data}}(t_{T>13.8^{o}C}\|t_0)$	first passage time probability distribution to temperatures above 13.8 ℃, conditioned on the initial date and gathered from the measured temperature data
p	probability
p_{opt}	optimal anomaly bandwidth for $\Delta T^{(3w,p)}$
$\hat{\rho}_k$	sample autocorrelation for lag k
RMSE	root mean square error
RPS (RPSS)	ranked probability score (ranked probability skill score)
r	climatological rate (base rate) of an event
$\hat{\sigma}$	sample standard deviation
SNHT	standard normal homogeneity test
std	standard deviation
τ_D	decorrelation time
T	measured temperature
$\widetilde{T}$	climatological temperature
T^{detr}	detrended temperature
t	date
t_0	initial date
t_{frost}	first passage time to frost
t_{max}	maximal first passage time that is still fully resolved by a forecast
UTC	Universal Time, Coordinated
WMO	World Meteorological Organisation
w	width used for a kernel density estimate
w_t	time window width
y_i	data point i in a time series
y	calendar year in which the data points were recorded

1 Motivation

Current operational weather forecasts are usually based on three nested numerical models with different resolution that couple ocean, atmosphere and land surface dynamics: a global model whose output is given as lateral boundary conditions to a regional model which in turn provides the boundary conditions for a local model. The German national weather service 'Deutscher Wetterdienst'(DWD) operationally uses an average model grid point distance of 20 km, 7 km and 2.8 km respectively, with 40 to 60 vertical levels[1]. This means that the dimension of the operational weather model space is $\mathcal{O}(10^8)$.

While such models resolve a large amount of relevant physical processes, some happen on too short time or length scales, such as atmospheric turbulence and clouds for instance, and therefore only enter the equations as parametrisations. These unresolved processes lead to model uncertainty as do errors in the parameter values of the model equations and the numerical discretisation[1]. A further source of errors is the necessary data assimilation: Observations of the current state of the weather are made on a network of very non-uniformly distributed measuring stations that has a high density of sites over Europe and North America, but a low density over less populated areas and over the oceans. The current weather condition therefore needs to be interpolated to the model grid points and correct initialisations for the non-measurable model variables need to be found. This introduces further sources of error into the numerical weather prediction process.

In order to incorporate this uncertainty into the forecasts, usually an ensemble of model runs is initiated with slightly perturbed initial conditions[2]. As the underlying model dynamics are chaotic[3, 4], initial uncertainties grow rapidly during model runs causing the ensemble to spread until it covers the whole bandwidth of possible weather states. This limits the attainable forecast lead time which does not exceed 7 days in conventional weather forecasts[1, 5]. Nevertheless, within these limitations weather services claim to attain a good forecast quality of 90% correct forecasts for the first five days[6].

While the quality of current operational weather forecasts for the first week into the future is widely appreciated, there are many applications across a much larger number of time scales where weather forecasts could make a difference. It has been found, for example, that about $1/7^{\text{th}}$ of the US economy is weather sensitive[7].

The most obvious application of weather forecasts is in agriculture where the expected weather forecast value for longer lead times is quite large, especially when the climatology (i.e. the long-term mean value) shows a high variability[8]. This is especially true for rain-dependent agriculture in vulnerable areas of small subsistence farms, where high quality forecasts could increase food security and help with poverty reduction[9, 10], but also generally in cases of drought, severe frost or flooding where forecasts could improve preparedness in the case of upcoming disasters.

But also under normal weather conditions, accurate long-range forecasts could impact the choice of crops and crop cultivars, decisions on the total planted area and planting density, as well as the timing of seeding depending on the expected amount of solar irradiation and total rainfall, as well as soil water availability[8, 11, 12, 13]. Across the year, they would also help determine the choice of fertilizer and pesticide and the timing of their application as well as the intensity and timing of irrigation measures and tillage, the stocking rates for cattle and the beginning of the harvest[11, 12, 13, 14]. Those management decisions by farmers

that do not concern the exact timing are often made several weeks in advance[13], with crop and grazing systems planned about a month ahead, and crop rotation and fallowing as well as investments into new land or machinery decided up to several years in advance[14, 15]. Accurate seasonal weather forecasts could therefore improve both crop yield assessment and management significantly[2].

Through the farmers, such weather forecasts would also impact seed suppliers and the agro-alimentary processing industry, the supply, the prices and therefore the consumers, and thus eventually also employment statistics and currency exchange rates[14, 15, 16].

To mitigate some of the weather-related risks and increase the possibility of advance planning, farmers currently use a futures market[8] similar to the stock market, where they can sell produce rights to predetermined conditions already before the harvest. Weather derivatives in general can be bought and sold as insurance against adverse conditions. They are also used by insurance and re-insurance companies to vary their investments[17]. Long-term forecasts could help determine prices and conditions, but there is a large need for time-resolved and accurate data[18].

Originally, the weather risk market was created by energy companies. There is obviously a strong correlation between the energy demand and outside temperatures due to the need for house and water heating, air conditioning and lighting[17, 18, 19]. Additionally, the increasing focus on renewable energy sources further reinforces the weather dependence of the energy sector. Not only solar and wind-based power generation, but also biomass growth and hydroelectric power generation depend on solar irradiation and precipitation[2, 19]. Accurate forecasts would help not only with seasonal operating strategies for the power plants, but they could also form the basis of new capacity building projects[16, 19].

The health sector is another example of the impact accurate longer-term weather forecasts could have. Mortality and hospitalisation rates, for example, increase during severe winters, especially in connection to heart problems, so anticipatory reinforcement of medication supplies and hospital infrastructures could be implemented with adequate forecasts[20]. In Africa, where excess rainfall and flooding cause a sharp rise in malaria and epidemic meningococcal meningitis infections[21], specifically targeted experimental seasonal weather forecast systems are actually already operational[2]. Also for farmers, some seasonal forecasts have gone operational mostly in the tropical latitudes[13]. But despite this large need for seasonal predictions, the current state of forecasts beyond two weeks' lead time generally is still lacking, especially in Europe.

In fact, while a third of all member states of the World Meteorological Organisation (WMO) have implemented experimental seasonal predictions[13], the World Climate Research Programme acknowledges that there are still 'daunting challenges in terms of predicting land surface temperature and rainfall'[22]. A scientist from the UK MetOffice even went as far as to say that the experimental seasonal forecast was only a 'slight advantage over not knowing anything at all'[23]. Indeed, depending on forecast lead time and target, seasonal predictions over Europe improve forecast skill by only up to 5% in most of the cases[22, 24, 25], when compared to simply forecasting the long-term average. They often even perform worse[26]. Only in rare situations, the improvement reaches as much as 40%[27].

This rather dismal picture comes about even though most seasonal predictions only forecast probabilities for observing one of three or five equiprobable categories of monthly, 3-monthly or seasonal mean variables such as mean temperature or total rainfall with zero lead time between the issuing of the forecast and the start of the averaging period[13, 24, 28]. Fig. 1.1 shows two examples of such seasonal temperature forecasts as they are issued regularly by the DWD. They give the relative forecast probabilities of below-normal, average, and above-normal mean temperatures for the seasonal average. While the summer forecast (left panel) shows a rather high probability of warmer than average temperatures, the autumn forecast does not deviate much from the climatology, i.e. the long-term average, and thus carries little information despite

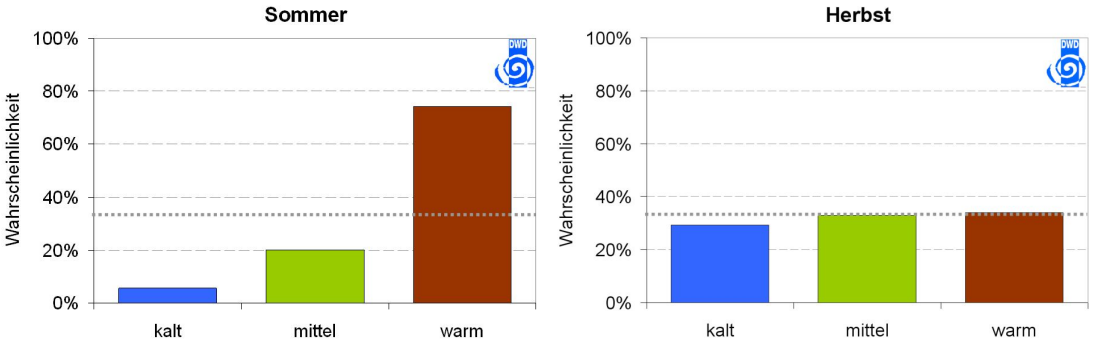


Figure 1.1: Two examples of seasonal temperature forecasts for Germany taken from the DWD website[1] in 2011 (left - summer) and 2012 (right - autumn) respectively. The columns represent the relative forecast probabilities for cold, averge and warm mean temperatures (left to right).

the heavy coarse-graining of the prediction target. This is rather typical as the forecast skill in spring and summer is generally better than in autumn and winter[1].

When forecasting weather conditions for the tropics, the quality for longer lead times is much better. Especially the sea surface temperatures and related phenomena such as the El Niño Southern Oscillation (ENSO) can be predicted very successfully at longer lead times[22, 29]. But over Europe, the seasonal weather forecasting skill is generally not very good. Here, purely atmospheric conditions have a larger influence, but the atmosphere varies due to baroclinic instability mostly on time scales of 5 to 10 days[4, 16], loosing memory after at most 15 days[26]. The longer term atmospheric variability is mostly due to the coupling with the oceanic heat bath and the interaction between the atmosphere and flow patterns over topography and large-scale heat anomalies[4, 13, 30].

Many other non-atmospheric variables with longer-term memory also contribute to the predictability on seasonal time scales. Some of them are external, such as solar forcing[29, 31], the influence of volcanic eruptions[32, 33], and greenhouse gas and aerosol forcing[29]. Others are internal, such as land surface temperatures[13, 34], soil moisture[22], snow cover and the related changes in the Earth's albedo[13, 16, 27]. Even the global warming trend contributes to the predictability of temperatures[24].

While every single factor usually only has a very modest impact on climate variability especially in the extratropics[29], the ocean's influence is very large not only through the sea ice cover[13, 22, 29], but mostly through the sea surface temperature anomalies that influence the atmosphere[2, 16, 22, 30]. Indeed, a strong ENSO signal that is forced through the sea surface temperature anomalies provides much of the seasonal predictability in the tropics[16, 24, 26, 35, 36].

But with the highly chaotic dynamics of the weather system especially in the extratropics[2, 4, 16, 30], the large number of variables that need to be assimilated into the model space mean a large initial error that is exponentially growing with time. Moreover, some of the influences such as soil moisture and snow cover but also the initial conditions of the ocean are especially difficult to assimilate into the model space[16]. Other climatic processes such as land-atmosphere interactions[22], but also ocean mixing, clouds and convection, and stratosphere-troposphere interactions that are important on seasonal time scales are not yet understood well enough to fully incorporate them[14, 16, 26].

The dynamical models not only have the problem of large and fast-growing initial errors, but also too large spatial scales for most specific applications[10, 13, 15].[1] This is to some extent remedied by dynamical downscaling, i.e. by nesting a local model into the global circulation model. However, downscaling does not only constitute a one-way flow of information from the global to more local scales that introduces inconsistencies, it also means that any errors in the global model output enter the local model through the boundary conditions, thus leading to even larger uncertainties on local scales[16].

The applications also need better-resolved time scales than averages over a month or even a whole season[10]. However, since the noise level on time scales between two weeks and two months is high and not yet averaged out, a higher temporal resolution of any model forecasting product would further lower the signal-to-noise ratio and result in even less predictability[13, 16]. On the scales that are needed for applications, the models therefore contain large errors and biases[29].

When faced with this extent of dynamical model problems on longer time scales, the justification for using the large computing resources and extensive data assimilation systems that are needed for the global circulation models[2] for seasonal prediction tasks and facing the ever growing expense due to added complexity to make them adequate for longer lead times has been questioned[24, 29].

Purely statistical data-based prediction schemes might provide a reasonable alternative for seasonal forecasts, following the principle of Ockham's Razor: If one has a choice, then the simplest method or model is best, if all other aspects are equal[37]. Statistical prediction schemes are usually based on the method of analogues as proposed by Lorenz[3]. Here, a situation that took place at the same time of the year (due to solar irradiation intensity) and had the same three-dimensional distributions of wind and pressure, as well as the same temperatures both of the land and sea surface, the same water vapor content, clouds, and snow cover, is identified as an analogue. Its temporal evolution is then taken as prediction. However, since this narrows the choice down to unacceptably few events, one usually has to be content with a smaller selection of similar conditions to define analogues, such as only the same pressure fields[3].

Apart from the problems caused by possible discrepancies between situations that are counted as analogues, statistical prediction also does not offer an understanding of the underlying dynamics[13], nor can it predict situations that have not yet occurred, be it because they are rare or because the climate and its variability are non-stationary[16, 29]. Nevertheless, considering the large operating expenses and errors of the dynamical models, purely statistical alternatives are usually fast, cheap and generally unbiased and can better be tailored to user needs thus making them potentially more valuable[29].

Using statistical methods in seasonal weather prediction is not a new approach: Even dynamical model output is generally modified by the so-called model output statistics, making it a hybrid prediction scheme. Every region also has its own country sayings or collections of purely observation-based weather rules such as the Old Farmer's Almanac, which is, however, generally not very skillful or simply based on long-term averages, i.e. the climatology[38]. Other prediction efforts for specific targets on a seasonal time scale have been implemented already with some skill even outperforming the dynamical models, such as for monsoon, rainfall and geopotential height[16], the 3-month mean surface temperature in Canada[35] or ENSO[39]. However, they are restricted to either very coarse-grained prediction tasks or more detailed forecasts valid only for tropical regions.

[1] The whole of Germany for instance is represented by only four grid points in the global model[2].

[2] One operational data assimilation and prediction run for a lead time of up to seven days of the global circulation model used by the DWD currently needs around 3,5 hours using a high-performance parallelised supercomputer[1].

In this thesis, we want to show that purely data-based statistical prediction schemes can offer significant skill also for specific time-resolved seasonal prediction tasks in the extratropics.

However, statistical prediction is often limited by the length of the observational data record[29]. One idea to remedy this drawback is to use an inexpensive, i.e. low-dimensionsal and computationally cheap, statistical model process to prolong the time series by mimicking its behaviour[40]. Such a procedure is often used as a stochastic mode reduction strategy in cases where a system contains fast degrees of freedom that are not fully resolved but only fitted with a discrete time stochastic difference scheme[40, 41].

Such an approach has been used to model the fast nonlinear atmospheric degrees of freedom such as turbulence[1, 42, 43, 44], but also temperature fields[41], soil moisture variability[45] and floods and precipitation (see [16] and refs. therein). The standard stochastic models used in climate are variations of seasonal autoregressive integrated moving average (SARIMA) models and thus use additive noise components[45, 41].

However, it has been found that the nonlinear coupling of modes in climate results in multiplicative noise[43, 44]. Indeed, even additive white noise forcing in the fast modes is transformed into multiplicative noise in the slow modes they influence[40]. Also the friction of a flow over topography is perturbed due to turbulent mixing in a state-dependent way[42]. Multiplicative noise therefore provides the appropriate closure for low-dimensional systems by representing the non-resolved degrees of freedom[43].

This is also due to the fact that multiplicative noise can introduce metastable states and the possibility of transitions between them into a system's dynamics[42, 45, 44]. It therefore has a significant impact on the qualitative structure of the probability density function of a system, where its effects are not only multimodality, but possibly also heavier tails[42, 43]. The introduction of multiplicative noise into a system therefore causes regime-like behaviour and enhances the probability that rarer states will be entered by the system, thus making it able to model large-scale flow regimes[42, 43] or also explain drought persistence as a noise-induced loss of stability of the deterministic non-drought equilibrium point[45].

Multiplicative noise has an influence on the underlying probability density functions, but two systems that have the same probability distributions do not necessarily have the same dynamical properties such as first passage times or mean residence times[43]. However, in another work we have shown exemplarily in a simple model system that first passage times arising from exclusively state-dependent noise-induced multistability can be adequately described by a corresponding system with purely deterministic hopping and additive white noise[46]. Therefore, to reproduce the dynamical behaviour of a system, one is not necessarily limited to multiplicative noise but needs to use an adequate combination of stochastic and deterministic properties in order to develop appropriate statistical prediction schemes.

As stated before, our goal is to demonstrate the predictive skill of statistical schemes on seasonal time scales in the extratropics. Such forecast methods have the advantage that they can be tailored to specific targets obtained from user needs. We will use the exemplary task of forecasting the first passage times until the occurrence of frost, i.e. for initial conditions above freezing we want to know the time it takes for the temperatures to drop below 0 °C for the first time, to illustrate the value of such prediction schemes. We want to show that for such a prediction target, it is possible to compete with and even outperform the full dynamical models operationally used in numerical weather prediction with data analysis methods. We will also try to improve our results by prolonging the time series using a very simple and computationally cheap autoregressive (AR) model process.

Therefore, after providing some necessary theoretical background in Chapter 2, we will first analyse the temperature data used in the thesis and look both at its first passage time properties and the possibility of modelling it using an AR process in Chapter 3. We will then proceed to analyse the potential for predictability using only conditional first passage time distributions

derived from the temperature time series and issue actual predictions of first frost in Chapter 4. In Chapter 5, we will look into enhancing the stationarity of the temperature anomaly time series the previous predictions were based on in a first effort at improving the predictive skill, before analysing the potential for further improvement by incorporating a second variable into our prediction schemes in Chapter 6. Chapter 7 then offers a direct comparison with the predictions from a dynamical weather forecast and the conclusions.

2 Theoretical background

2.1 Stochastic processes and time series analysis

2.1.1 Introduction

To define stochastic processes, we will first start with the concept of a *random variable*. This is a variable whose value cannot be predicted - there might be a lack of knowledge about initial conditions or the finer, small-scale details of its dynamics. A random variable is therefore defined on a set of possible outcomes together with a probability distribution that associates each outcome with a probability of occurrence.

A *stochastic process*[1] is a collection of random variables $\{X_t : t \in T\}$ indexed by a time set T.[2] All random variables have the same codomain, namely the *phase space* $\mathcal{S}$ of the stochastic process. The process is fully described by the infinite set of all joint probability distributions $p(x_1, t_1)$, $p(x_2, t_2; x_1, t_1)$, $p(x_3, t_3; x_2, t_2; x_1, t_1)$,...

Measurements obtain a specific realisation of an underlying stochastic process, i.e. a specific value in $\mathcal{S}$ for each random variable. Such a set of observations $\{x_t\}$ is also called a *time series*. To fully characterise the stochastic process, the goal is to analyse this time series and learn all joint probability distributions. But before we move on to some aspects of time series analysis, we will first define two specific stochastic processes that are often used as simple models of more complex dynamics.

The first and most basic one is the *random walk*. It is defined as $X_t = X_0 + \sum_{i=1}^{t} \xi_i$, where $\{\xi_t\}$ is *white noise*, i.e. a set of independent random variables with zero mean and constant variance that are normally distributed. In more mathematical terms, $\langle \xi_t \rangle = 0$, $\langle \xi_t \xi_{t'} \rangle = \hat{\sigma}^2 \delta_{tt'}$ and $\xi_t \sim N(0, 1)$.

The second example is a *Markov process*. Here, the knowledge of its future values is only determined by the present, not by the past. The Markov property therefore can be written as $p(x_n, t_n | x_{n-1}, t_{n-1}, x_{n-2}, t_{n-2}, ..., x_0, t_0) = p(x_n, t_n | x_{n-1}, t_{n-1})$.

Two important properties stochastic processes can have are stationarity and ergodicity. For a *stationary process*, the marginal probability distributions $p(x_i, t_i)$ are independent of the time t_i and therefore also the two-point joint probability distributions $p(x_i, t_i; x_j, t_j)$ depend only on the time difference $t_j - t_i$. As this is a rather restrictive definition that is difficult to verify when considering a single measured realisation, several weaker modifications of stationarity exist. The most notable one is the *weakly stationary random process* whose mean, variance and autocorrelation structures are time-independent.

A stochastic process is *ergodic*, on the other hand, if in the asymptotic limit its time average over one infinite realisation is the same as the ensemble average over all realisations.

2.1.2 Distributional analysis

As stated before, the full set of joint probability distributions is needed to fully characterise a stochastic process. To analyse a given time series, i.e. a realisation of a stochastic process one

[1] For a more extensive introduction to stochastic processes see for instance[47, 48, 49], on which this summary is based.

[2] In this thesis, we will only consider the special case of discrete sets with $T = \mathbb{N}$, as our observations x_t are daily sampled temperature measurements.

knows a priori nothing about, the most reasonable start would be to establish a probability distribution of measurement values. This can, however, prove to be quite difficult.

The easiest approach is to first use an array of summary measures to describe the underlying full probability distribution the observations are drawn from. The most common is the *sample mean*. It is, however, heavily influenced by outliers in the observations. Therefore, a more robust measure of location, for example the *sample median*, is often more useful. It is defined as $q_{0.5}$, i.e. that value which separates the lower half of the observations from the upper half[3].

To describe the spread of the distribution, the *sample variance* $\hat{\sigma}_X^2$, or its square root the *standard deviation* (std) are most commonly used. They are, however, also not robust with regard to outliers in the observations. A better measure also relies on quantiles in a similar way as the median, but in case of the spread, the *interquartile range* (IQR) is defined by $\mathrm{IQR} = q_{0.75} - q_{0.25}$.

Other summary measures include the *peak* of the probability distribution, i.e. the value x for which the probability distribution attains its maximum.

If summary measures are not sufficient, the simplest estimate of the underlying probability distribution is a *histogram*. A drawback of the histogram is its step-function-like nature. If one wants instead a smooth estimate of the underlying probability distribution, *kernel density smoothing* offers another nonparametric estimation possibility. To do this, for each value assumed in the time series, a normalised Gaussian kernel that is centred at the value is added to the estimate[4].

The specific choice of the kernel width is crucial: If one chooses a very small kernel width, the histogram contains large fluctuations due to the finite sampling. If on the other hand a large kernel width is employed, important details of the underlying probability distribution might be smoothed out.

A kernel density estimate thus is a generalisation of a histogram which can be considered as using rectangular kernels.

An important aspect of a time series is how closely the different observations are related to each other. This can be measured through the sample autocorrelation function ρ_k for lag k between observations, which is given by Eq.(2.1).

$$\rho_k = \frac{\sum_{t=1}^{n-k}(x_t - \langle x\rangle)(x_{t+k} - \langle x\rangle)}{\hat{\sigma}_X^2} \tag{2.1}$$

2.1.3 Spectra and frequency analysis

For some purposes when analysing a time series, it can be of interest to leave the time domain and instead look at the frequency domain to see how much oscillations with different frequencies contribute to the overall variance. In these cases, the *power spectral density* is a useful quantity to look at. The Wiener-Khinchine theorem states that it is the Fourier transform of the autocorrelation function ρ_k, whose sample estimate was introduced in Eq.(2.1), if one considers the time series to be a weakly stationary random process[5].

[3] A *quantile* q_p of a probability distribution $p(x)$ is defined such that $p(x \leq q_p) = p$.

[4] For kernel density smoothing, any functional shape can in principle be chosen as a kernel. However, we will limit ourselves to Gaussian kernels in this thesis. Of course, Gaussian kernels can distort the probability distribution estimate if the support is actually finite or in particular nonnegative as they will assign some nonvanishing probability to all real values.

[5] Sometimes the Wiener-Khinchine theorem is also based on the Fourier cosine transform. Since the autocorrelation function is even, these two versions are equal.

The power spectral density in discrete time is given by:

$$S(f) = \sum_{k=-\infty}^{\infty} \rho_k e^{-2\pi i f k} \tag{2.2}$$

In the case of a finite time series where the underlying dynamics are a priori unknown but the most important frequencies are of interest, the *periodogram* provides an asymptotically unbiased estimator of the power spectral density.

It is established as follows: First, the time series $\{x_t\}$ is fitted with a Fourier series as shown in Eq.(2.3), where $\alpha_0 = \langle x_t \rangle$ is the sample mean of all n sample points, ϵ_t is the residual noise caused by numerical approximations and $q = \frac{n-1}{2}$ for n odd and $q = \frac{n}{2}$ for n even[38, 50].

$$x_t = \alpha_0 + \sum_{i=1}^{q} [\alpha_i \cos(2\pi f_i t) + \beta_i \sin(2\pi f_i t)] + \epsilon_t \tag{2.3}$$

The frequencies $f_i = \frac{i}{n}$ represent the higher harmonics of the fundamental frequency and the coefficients α_i and β_i can be obtained through the orthogonality properties of the trigonometric functions as stated in Eqs.(2.4), with α_q and β_q calculated separately for even n as given by Eqs.(2.5).

$$\begin{aligned} \alpha_i &= \frac{2}{n} \sum_{t=1}^{n} x_t \cos(2\pi f_i t) \qquad \text{for } i = 1, 2, ..., q \\ \beta_i &= \frac{2}{n} \sum_{t=1}^{n} x_t \sin(2\pi f_i t) \qquad \text{for } i = 1, 2, ..., q \end{aligned} \tag{2.4}$$

$$\left.\begin{aligned} \alpha_q &= \frac{1}{n} \sum_{t=1}^{n} (-1)^t x_t \\ \beta_q &= 0 \end{aligned}\right\} \qquad \text{for } n \text{ even.} \tag{2.5}$$

The periodogram then consists of the q intensity values given by Eq.(2.6).

$$\begin{aligned} I(f_i) &= \frac{n}{2}(\alpha_i^2 + \beta_i^2), \qquad \text{for } i = 1, 2, ..., q, \\ I(f_q) &= n\alpha_q^2, \qquad \text{for } n \text{ even.} \end{aligned} \tag{2.6}$$

Even though the periodogram is a nice and widely used estimator whose expectation value converges to the true power spectral density for weakly stationary random processes, it has several drawbacks[38, 50]. Even though it is asymptotically unbiased and the estimates at adjacent frequencies are nearly uncorrelated, for finite samples it can have large biases. In fact, the finiteness of the time series is incorporated into the Fourier transform as a rectangular window in time, which abruptly turns the measurements on at $t = 1$ and off again at $t = n$[38]. Such a rectangular window has rather bad properties if it is Fourier transformed, leading to significant side lobes to the main peak in the spectrum.

This effect can be mitigated by *tapering* the initial time series, i.e. by using a windowing function that smoothly drops to zero at the edges and thereby decreases the magnitude of the side lobes. Of course, reducing the weight of a significant part of the time series - often around 10 to 20% - leads to a loss of information and one therefore needs more data points to get a similar amount of information as without the tapering. However, this is well worth it especially if sharp lines are suspected in the spectrum, because it avoids distributing the contribution of

a single frequency line to several non-adjacent frequency bins favored by the side lobes of a rectangular window.

Another problem is the discrete and finite frequency support of the power spectral density estimate. On the one hand, since the periodogram only calculates the higher harmonics of the fundamental frequency $f = \frac{1}{N}$, such lines whose frequencies do not correspond exactly to a higher harmonic contribute to several bins in the calculation of the periodogram, resulting in a smeared-out spectrum. On the other hand, the sampling frequency f_s of the original time series provides an upper bound $f_s/2$ for the frequency domain of the estimate, called the Nyquist frequency. This means that the estimate 'folds back' frequencies exceeding the Nyquist cutoff into the lower part of the spectrum, identifying them wrongly, an effect commonly called *aliasing*.

The periodogram as a power spectral density estimator is also not consistent, i.e. its variance does not converge to zero with a time series that becomes infinitely long. Variance reduction therefore also poses a problem. However, this can be addressed by computing periodograms for different segments of the original time series and then taking an ensemble average of all estimates. The number of segments the time series is separated into can be optimised for bias-variance tradeoff, as the larger the number, the smaller the variance in the average estimate but also the smaller the number of data points in each part, increasing the bias within each single estimate. A smaller number of data points N' in each individual periodogram also restricts the overall frequency resolution of the average, as the fundamental frequency changes to $f = \frac{1}{N'}$.

Considering all these different problems and attempts at resolving them, it is not surprising that there are many versions of implementing such an estimator. In this thesis, we will use Welch's method[51], where the time series is cut into segments of equal length L, which ought to be a power of 2, either directly or by padding the time series with additional zeros for maximum efficiency of the fast Fourier transform algorithm. The segments are then tapered by a Hann window - also called Hanning window, similar to a cosine arch spectral window - as given in Eq. (2.7).

$$w(i) = \frac{1}{2}\left(1 - \cos\left(2\pi \frac{i}{L-1}\right)\right), \qquad 0 \leq i \leq L-1. \tag{2.7}$$

For each of these segments[6], a finite Fourier transform is then calculated, resulting in a modified periodogram. Finally, these periodograms are averaged, leading to an estimate of the power spectral density.

2.1.4 Autoregressive processes

We will now turn to a special class of stochastic processes that is widely used as a very simple model process in time series analysis. The autoregressive processes are generally defined as follows[38]: $\{X_t : t \in \mathbb{Z}\}$ is an *autoregressive process of order p* (also referred to as *AR(p)-process*) if there exist real constants α_k, with $k = 0, ..., p$, $\alpha_p \neq 0$ and a Gaussian white noise process $\{\epsilon_t : t \in \mathbb{Z}\}$ such that

$$X_t = \alpha_0 + \sum_{k=1}^{\mathrm{p}} \alpha_k X_{t-k} + \sigma^2 \epsilon_t. \tag{2.8}$$

Although this class is not complete, they are widely used because any weakly stationary ergodic process $\{X_t\}$ can be approximated arbitrarily closely by an appropriately chosen au-

[6] Two consecutive such segments always have an overlap of 50% to reduce the information loss caused by the tapering procedure.

toregressive process[38].

AR(p)-processes are stationary if and only if all roots of their characteristic polynomial

$$P(x) = 1 - \sum_{k=1}^{\mathrm{p}} \alpha_k x^k, \tag{2.9}$$

where α_k represent the coefficients of the AR(p)-process, lie outside the circle $|x| = 1$.

To understand the nature of these processes a little better, we will look at the lowest order processes, as they are easier to interpret. The simplest case of the AR(p)-processes, the AR(0)-process, is equal to a white noise process, albeit one whose mean might differ from $\mu = 0$ depending on the parameter α_0.

The next order, the AR(1)-process as given by Eq.(2.10), corresponds to a red noise process. Its value X_t depends exclusively on the value X_{t-1} at the previous time step[7], which in the stationary case of $|\alpha_1| < 1$ is regressed towards the mean, resulting in a smoothing-out of short-term fluctuations. The spectrum is thus depleted in the shorter wavelengths which leads to a "red colour".

$$X_t = \alpha_0 + \alpha_1 X_{t-1} + \epsilon_t. \tag{2.10}$$

The AR(1)-process in fact describes the linear regression of X_t onto X_{t-1}, as already suggested by its name, where the α_i represent the regression coefficients, i.e. the intercept and the slope respectively, and ϵ_t represents the regression residuals.

A measure of the "memory" of the AR(1)-process can be given by the decorrelation time:

$$\tau_D = \frac{1 + \alpha_1}{1 - \alpha_1}. \tag{2.11}$$

As stated before, the AR(p)-processes are widely used as simple models of more complex and unknown dynamics. In this case, the relevant parameters first need to be estimated numerically from the time series. The most widely used methods of estimation for autoregressive processes are Yule-Walker, Burg and maximum likelihood estimation. While they are asymptotically equivalent, maximum likelihood and Burg are generally preferred for smaller sample sizes when the order p is known because the Yule-Walker estimate tends to have larger biases.

For overparameterised fits that appear in any problem where the optimal order first has to be determined, however, the Yule-Walker estimate of the variance σ^2 is more stable and reliable and this method thus does not tend to select overparameterized models. Any AR(p)-process parameter estimation in this thesis will therefore be based on the time series autocorrelation coefficients using the Yule-Walker equations. They are defined as follows:

$$\gamma_m = \sum_{i=1}^{\mathrm{p}} \alpha_i \gamma_{m-i} + \sigma_\epsilon^2 \delta_{m,0}. \tag{2.12}$$

γ_m denotes the autocorrelation of lag m of the time series.

When fitting an AR(p)-process to a time series, the optimal order p is not usually known a priori. If it is chosen to be too small, then the resulting model process will not be able to capture all the inherent correlations in the time series. A too large model order on the other hand results in overfitting, the model will be tuned excessively to the data points that are being used in the fitting procedure, thus picking up specifics of the noise realisation instead of representing the underlying dynamical system in general.

[7] The AR(1)-process thus possesses the Markov property as defined in Sec. 2.1.1.

Therefore, objective criteria to determine the optimal p are needed. They are usually based on the maximal log-likelihood which is used to characterize the expected variance of the prediction error when an AR model fitted to a specific time series is used for a different realisation of the same underlying process. However, they differ in the added penalty function for the number of parameters.

The best known procedure to determine the optimal model order was proposed by Akaike[52]. It involves the minimisation of the Akaike Information Criterion (AIC) as given in Eq.(2.13), where n denotes the sample size of the time series that is to be fitted, p the order chosen for the fit and $\hat{\sigma}_p^2$ the estimate of the variance.

$$\mathrm{AIC(p)} = \ln \hat{\sigma}_p^2 + \frac{2\mathrm{p}}{n} \tag{2.13}$$

However, it has been shown that the minimisation of the AIC does not converge to the optimal order p with probability 1, but overestimates p asymptotically[53]. Schwarz [54] proposed an alternative penalty function that remedies that problem, i.e. that is strongly consistent but imposes a greater penalty upon the number of parameters and therefore leans more towards lower-dimensional models than the AIC:

$$\mathrm{BIC_{Sch}(p)} = \ln \hat{\sigma}_p^2 + \frac{p}{n} \ln n. \tag{2.14}$$

Hannan and Quinn have refined the criterion provided by Schwarz using another penalty function that leaves the estimation strongly consistent and better than the AIC for larger sample sizes n, but will underestimate the optimal order less[53]:

$$\mathrm{BIC_{HQ}(p)} = \ln \hat{\sigma}_p^2 + \frac{2pc}{n} \ln \ln n, \qquad \text{with } c > 1. \tag{2.15}$$

2.2 Determining statistical significance and quantifying uncertainty

2.2.1 Introduction to statistical hypothesis tests

Statistical hypothesis tests are a method of determining whether any inference[8] made about sample data is the result of pure chance of the sampling or actually statistically significant. This is done by formulating the statement to be tested as a null hypothesis, usually in such a way that one expects it to be rejected if the effect observed is significant. Then one determines how likely the observed results in the data are if the null hypothesis is true. If this probability, which is also called *p value*, is smaller than a predetermined threshold, the *significance level* α of the test, the null hypothesis is rejected at the given significance level. Note that not rejecting a null hypothesis merely means that there is insufficient evidence to suggest it is false, not that it is necessarily true. Common choices for the significance level are $\alpha = 5\%$ or, to make it even less likely that the observed effect is due to chance, $\alpha = 1\%$.

The significance level is thus used to determine the probability of a so-called Type I error occurring in the testing, namely the probability that a true null hypothesis is falsely rejected. While one therefore has a good handle on this type of error and can reduce it by changing the significance level, this is not true for Type II errors, whose probability of occurring is often denoted as β. These are the errors where a false null hypothesis is not rejected. As Type II errors are inversely proportional to the significance level, they are in practice often rather

[8] In climate research, the frequentist approach is mostly preferred over the Bayesian inference[38] and therefore also the focus of this section.

large[55]. However, the *power* of a statistical hypothesis test is given by $(1-\beta)$, as low Type II errors are needed to reject the null hypothesis sufficiently often if it is indeed false.

The specific choice of the null hypothesis, as well as the choice of a *test statistic* to determine how likely the observed results are if the null hypothesis is true, significantly impact the results of a statistical hypothesis test. Also specifying an *alternative hypothesis* that differs from "the null hypothesis is not true" changes the test results. Accordingly, there is a wealth of hypothesis tests for different goals of testing and different underlying assumptions about the data such as e.g. its probability distribution or its correlation structure. We will therefore in the following only look into more detail at those tests that will be used later in this thesis.

2.2.2 A test for randomness of AR model fit residuals

The *Ljung-Box portmanteau test* was first proposed in 1978 [56] to test an autoregressive moving average (ARMA) model for goodness of fit to a time series. To do this, the residuals after the fit are tested for true randomness using their autocorrelations. This is done by computing the following test statistic Q:

$$Q = n(n+2)\sum_{k=1}^{N}\frac{\hat{\rho}_k^2}{n-k}, \tag{2.16}$$

where n denotes the total sample size of the residual time series, $\hat{\rho}_k$ the sample autocorrelation for lag k of the residuals, and N the number of lags being tested. This test quantity Q should then be distributed as χ^2_{N-p-q}, where $p+q$ is the total number of parameters in the ARMA model.[9]

Although the residuals of the model fit are assumed to be normally distributed, this test is insensitive to departures from normality as long as the variance remains finite[56].

Contrary to ordinary statistical hypothesis tests, a portmanteau test does not have a specific alternative hypothesis to test against. This makes it less powerful than more specific tests when the type of discrepancy is known, but very useful when it is not[56].

2.2.3 Tests for time series homogeneity

Especially in long-term measurements such as geophysical time series, measuring devices or even the measuring station location might have been changed in the interim. In these cases, it is of great interest to assess whether this has resulted in a systematic change in the measured values. This can be done using homogeneity tests. All of the following tests assume that the values y_i of the time series are independent and identically distributed[10].

The first of these homogeneity tests that we will look at is the *Standard Normal Homogeneity Test* (SNHT) proposed by Hawkins[58] and whose application was described in detail by Alexandersson[59]. Its purpose is to test a time series for a unique break in the mean value, i.e. a step-wise shift.

To this end, the n variables y_i are transformed to their normalised counterparts z_i as follows:

$$z_i = \frac{y_i - \langle y \rangle}{\hat{\sigma}} \tag{2.17}$$

with $\langle y \rangle$ the sample mean of the y_i and $\hat{\sigma}$ their sample standard deviation. The null hypothesis of this test is that these z_i are independent and belonging to $N(0,1)$, i.e. following a standard

[9] Ljung and Box showed that this asymptotic distribution is quite well approximated already for moderate sample sizes, contrary to an earlier similar test proposed by Box and Pierce[57].

[10] In climate research, the data are rarely truly independent since they usually come from a single observational record that even contains serial correlations. Keeping this in mind, one can nevertheless draw useful quantified analyses from statistical hypothesis tests.

normal distribution with mean 0 and standard deviation 1:

$$H_0 : \qquad z_i \in N(0,1) \qquad \forall i. \tag{2.18}$$

The alternative hypothesis is then formulated as follows:

$$H_1 : \begin{cases} \text{For some } 1 \leq \nu < n \text{ and } \mu_1 \neq \mu_2 : \\ \qquad z_i \in N(\mu_1, 1), \qquad \text{for} \quad i \leq \nu \\ \qquad z_i \in N(\mu_2, 1), \qquad \text{for} \quad i > \nu. \end{cases} \tag{2.19}$$

The test statistic is derived using the likelihood ratio method:

$$S_0 = \max_{1 \leq \nu < n} \{S_i\} = \max_{1 \leq \nu < n} \{\nu \langle z_1(\nu) \rangle^2 + (n-\nu) \langle z_2(\nu) \rangle^2\} \tag{2.20}$$

with $\langle z_1(\nu) \rangle = \sum_{i=1}^{\nu} z_i / \nu$ and $\langle z_2(\nu) \rangle = \sum_{i=\nu+1}^{n} z_i / (n-\nu)$.

If S_0 exceeds a critical level that can be obtained through simulation depending on the sample size and the desired confidence level[58], the time series should be classified as non-homogeneous.[11]

The advantage of this test is that it does not rely on other similar but definitely homogeneous time series to compare the data with. Moreover one can get an estimate of the time of the step-wise shift in the record from the index ν for which the maximum S_0 is reached. From the sample means of all values before and after this index, a magnitude of the break in the record can then be estimated.

Problems with this test can occur if the z_i are not normally distributed or if the standard deviation of the time series has some breaks as well. The most probable year of a break in a truly homogeneous time series according to this test is also more likely to be near the beginning or the end of the record[59].

The second such test that we will look at, the *Buishand range test*, also detects a single step-wise shift in the mean. But, unlike the SNHT, this test is more sensitive to breaks in the middle of the time series[60].

In terms of both null hypothesis and assumptions, the Buishand range test and the SNHT are very similar. The Buishand range test also employs the same transformation of variables as the SNHT (Eq.(2.17)), but then uses as test statistic R as given by Eq.(2.21), where $z_0 = 0$.

$$R = \max_{0 \leq i \leq n} z_k - \min_{0 \leq i \leq n} z_k. \tag{2.21}$$

For homogeneous records, the z_i should fluctuate around 0 and the range should stay smaller than some critical value depending on the total length n of the record and the desired significance level α. Tables of critical values can be obtained for example from Buishand[61].

The third test for homogeneity, namely the *Pettitt test*, uses the null hypothesis that all data points y come from a single distribution function f. As the Buishand range test, it is more sensitive to distribution changes that occur towards the middle of the time series[60]. However, as it is based upon the rank of the data points within the time series rather than upon their actual values, it is less sensitive to outliers and does not require an underlying normality of the distribution[60]. In fact, the test statistic of the Pettitt test[62] is equivalent to a Mann-Whitney statistic for testing that the data points in the two samples $y_1, ..., y_k$ and $y_{k+1}, ..., y_n$ are drawn from the same distribution. It tests for homogeneity against the alternative hypothesis that

[11] For many statistical hypothesis tests, the test statistic if the null hypothesis is true cannot be determined analytically. In these cases, it is often obtained through Monte Carlo sampling methods.

there exists a point y_τ such that all data points y_i with $i \leq \tau$ are drawn from one distribution f_1 and all points y_i with $i > \tau$ from the distribution f_2 with $f_1 \neq f_2$. The functional forms of f_1 and f_2 are not restricted, contrary to the tests mentioned previously, except to demand continuity in the version presented here[62].

The test statistic is given by:

$$X_E = \max_{1 \leq k \leq n} \left| \sum_{i=1}^{k} \sum_{j=i+1}^{n} \operatorname{sgn}(y_i - y_j) \right|, \tag{2.22}$$

with n the total number of sample points y.

The probability p that the null hypothesis $f_1 = f_2$ holds is then given approximately by Eq.(2.23). This approximation is accurate to two decimal places for $p \leq 0.5$ [62].

$$p \approx 2 \exp\left(\frac{-6 X_E^2}{n^3 + n^2} \right) \tag{2.23}$$

As the Pettitt test assumes underlying stationarity and independence of the data points y, problems can occur when there are long-range correlations in the time series. Indeed, the test is much more likely to issue high probabilities for the existence of a change point if correlations are present[63], as the resulting clustering of values leads to larger values of X_E.

The last test of time series homogeneity we are going to use in this thesis is the *Von Neumann ratio test*. It does not find the location of a step-wise shift in the mean of the time series as the previous ones but instead tests against the alternative hypothesis that the data points in the series are not randomly distributed. As such, it finds departures from homogeneity that are not strict step-wise shifts but does not give information on the time at which the break in homogeneity occurred[60].

The test was proposed by von Neumann[64, 65] and is based upon the following test statistic:

$$N = \frac{\sum_{i=1}^{n-1} (y_i - y_{i+1})^2}{\sum_{i=1}^{n} (y_i - \langle y \rangle)^2}. \tag{2.24}$$

In the homogeneous case, $\mathbb{E}[N] = 2$. If the sample contains a break in the mean, N tends to be smaller than the expected value[61], when there are rapid variations in the mean, it may become larger than the expected value[60].

This test is also rather sensitive to trends and correlations in the data that cause stronger deviations from the expected value of the test statistic[64].

2.2.4 Tests for normality of a distribution

Many statistical hypothesis tests have the underlying assumption of normally distributed data points. However, it is not always clear if this assumption is valid for the time series in question. It is therefore often a necessary first step to make sure that the data points do indeed follow a normal distribution because otherwise, the results of other statistical hypothesis tests using this assumption might be misleading.

The easiest method to determine whether a distribution can be considered normal is not a hypothesis test but rather a visualisation method: the *Normal probability plot*. It is less precise but much simpler than hypothesis tests and useful as often hypothesis tests based upon an assumption of underlying normality of the probability distribution still lead to valid results if the distribution is close enough to a true normal.

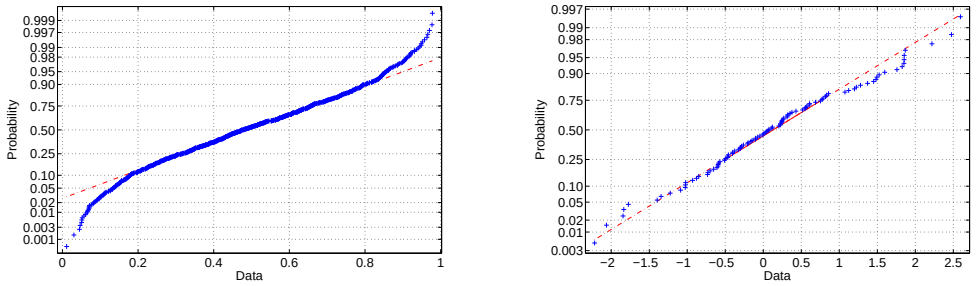


Figure 2.1: Normal probability plots of a Beta(2,2)-distribution (left panel) and of a N(0,1)-distribution sampled only with 100 data points (right panel). The red dashed lines represent the closest normal distribution.

The normal probability plot is a variation of plotting the cumulative distribution, where the vertical axis is scaled in such a way that the cumulative of a normal distribution would yield a straight line. The left panel of Fig. 2.1 shows such a normal probability plot exemplarily for a Beta distribution. It is clearly visible that the closest fitting normal distribution has heavier tails than the Beta distribution.

Deviations may, however, also occur on both ends of the sampled interval if the time series follows a truly normal distribution. The very rare extreme values might not have occurred yet within the total finite sampling duration or might have occurred comparatively often. An example of this is shown in the right panel of Fig. 2.1, where only 100 randomly sampled points from a normal distribution were used for the plot and the fluctuations in the tails are clearly visible.

Moving on to true statistical hypothesis tests, there are many versions that test for normality of the distribution of a data set. This is due to the large demand considering that normality is an underlying assumption of many other statistical tests. Moreover, when testing the goodness of fit of any model, normality in the residuals is also a sought-after indicator.

The specific hypothesis test we chose to use in this thesis is the *Jarque-Bera test*. It was first proposed by Jarque and Bera [66] and is based on Lagrange multipliers. It has been shown to perform best for larger sample sizes of $n \geq 100$ [66] at least for tests that do not specify an alternative probability distribution, making it one of the most used tests for this purpose[67].

Its test statistic is given by:

$$\begin{aligned} JB &= \frac{n}{6}\left(S^2 + \frac{1}{4}K^2\right) \\ \text{with} \quad S &= \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle y\rangle)^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle y\rangle)^2\right]^{\frac{3}{2}}} \\ \text{and} \quad K &= \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle y\rangle)^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \langle y\rangle)^2\right]^{2}} - 3. \end{aligned} \tag{2.25}$$

The Jarque-Bera test therefore measures deviations in sample skewness and kurtosis from those expected for a truly normal distribution. If the n sample points are from a normal distribution, JB follows a χ^2-distribution with two degrees of freedom.

2.2.5 Tests for equality of two probability distributions

The most used and well-known test for comparing two empirical data sets for equality of the underlying probability distribution is the χ^2-test. However, in this thesis, we will instead use the two-sample *Kolmogorov-Smirnov test*. While the χ^2-test compares probability densities, the Kolmogorov-Smirnov test instead relies on cumulative probability functions which can be estimated without binning or parametric fitting. Moreover, when binning the data, the χ^2-test introduces unordered categories and is then invariant to group permutations[68].

Given two independent data sets $\{y_i^{(1)}\}$, $i = 1, ..., n_1$, and $\{y_j^{(2)}\}$, $j = 1, ..., n_2$, with corresponding empirical cumulative distributions $F_1(y)$ and $F_2(y)$, the null hypothesis of this test is that both data sets were drawn from the same underlying continuous probability distribution $F(y)$.

The test statistic used in this case (see for instance [49]) is given by

$$D = \max_y |F_1(y) - F_2(y)|, \tag{2.26}$$

and the null hypothesis is rejected at the significance level α if

$$D > \left[-\frac{1}{2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \ln \left(\frac{\alpha}{2} \right) \right]^{\frac{1}{2}}. \tag{2.27}$$

The Kolmogorov-Smirnov test has two fundamental assumptions on the data sets in question. The first is independence. This means that if it is used to directly compare an empirical cumulative distribution to a specific theoretical one, no parameter estimation or distribution fitting should have been employed. If this is violated, the test will not necessarily reject the null hypothesis even when it is false, meaning that the critical values are overestimating the significance level[49].

The second assumption is the continuity of the underlying probability distribution. Only in this case the exact critical distribution is known[69] and distribution-free[68]. However, in the case where the underlying distribution is inherently discrete or has been made discrete through grouping of data points, the theoretical and distribution-free significance level obtained for the continuous case constitutes an upper bound for the significance level in the discrete case[70]. This makes the Kolmogorov-Smirnov test conservative if it is employed in the discrete case so that the null hypothesis is rejected less often than if the true distribution of the test statistic were known.

Instead of testing the full probability distributions of two data sets for equality, one might of course also focus on specific summary measures such as the mean or the variance. Statistical hypothesis tests for the equality of the mean or median usually are parametric, i.e. they require either an underlying normal distribution or knowledge about the specific probability distribution. This is not available in the cases analysed in this thesis, so we will not look further into these tests.

However, in another context, we want to test whether the variance in several subsets of a time series is the same, i.e. we are interested in the homoscedasticity of the data sets. Also in this case, most of the well-known hypothesis tests, such as the χ^2 test, the F test and Bartlett's generalisation to more than two samples are quite sensitive to departures from normality of the underlying distribution[71]. The F test is also not robust against outliers or in the case of dependent variables[38]. However, there exists a more robust alternative that performs best with mostly symmetric distributions with moderate tails, namely *Levene's test*[72]. While it is less powerful than Bartlett's test for normal distributions, it is more powerful when normality is violated[71].

The test statistic used by Levene's test to verify the null hypothesis that the population variances of all groups are equal is given by Eq.(2.28). Here, n is the total number of data points contained in k groups of size $n_i, i = 1...k$, and y_{ij} is the value of the j^{th} data point in group i. $y_{i\cdot}$ denotes the mean of all values in group i.

$$\begin{aligned} F &= \frac{n-k}{k-1} \frac{\sum_{i=1}^{k} n_i (z_{i\cdot} - z_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (z_{ij} - z_{i\cdot})^2} \\ \text{with} \quad z_{ij} &= |y_{ij} - y_{i\cdot}| \end{aligned} \tag{2.28}$$

2.2.6 Confidence intervals and the bootstrap

The determination of confidence intervals for estimated parameters or scores usually is parametric and uses assumptions about underlying distributions to make analytical calculations. If such assumptions are unlikely to be satisfied by the data that is analysed or if the analytic calculations prove to be intractable, then the *bootstrap* offers a non-parametric alternative[38, 73]. It relies solely on the assumption that the data are independent and identically distributed random variables.

The bootstrap is a resampling procedure based on Monte Carlo methods. If the original data set contains n data points, then n points will be randomly drawn from the set with replacement to form a new resampled data set of the same sample size. This ensures that the procedure asymptotically converges on the correct value[73]. From the resampled data set, the desired parameter is again estimated. This is repeated N times.

From the empirical distribution of the parameter estimates, one can proceed to obtain uncertainty estimates. One possibility is to calculate the standard error. It simply uses the sample standard deviation $\hat{\sigma}$ of the distribution, which corresponds to the non-parametric maximum likelihood estimate of the true standard error[73]. The corresponding interval of width $4\hat{\sigma}$ around the parameter estimate is then the standard interval which is roughly equivalent to a 95%-confidence interval if the distribution of the bootstrap estimates is close to a normal distribution. It can be quite inaccurate, however, if the distribution is skewed instead[73]. The next simplest way of obtaining confidence intervals through a bootstrap method is to use the percentiles of the bootstrapped sample distribution of the parameter[73, 74, 75].

An open question is the number N of bootstrap samples needed to obtain a good estimate of the uncertainty involved. To use all possible resampled data sets obviously carries a very high computational cost, especially if the calculation of the parameter to be estimated is lengthy. If one is only interested in the standard error, then 50 to 100 resampled data sets should already be enough[38, 73]. For the more involved percentile intervals, a larger number of samples is necessary as one then needs to estimate the tails of the sample parameter distribution. Efron et al.[73] propose using $N \approx 250$ in this case and $N \approx 1.000$ as a rough minimum for even more involved methods of estimating confidence intervals. Wilks even proposes $N = 10.000$[49]. To make sure that N was chosen large enough, one can run the estimation of the confidence interval twice with equal N and increase N if these two results still differ too much[76].

One problem with using the bootstrap is its sensitivity to a violation of the assumption of independent data points. Most data sets, especially those originating in the atmospheric sciences, contain serial correlations. These are destroyed in the resampling procedure of the bootstrap so that the resampled data sets have a quite different correlation structure from the original[38, 77]. These serial correlations lead to a smaller confidence interval than appropriate, an effect that increases with the strength of the correlations[77]. One possibility to avoid this problem is to use the bootstrapping procedure on blocks of data that encompass at least a correlation length instead of on single data points. The results will however depend sensitively on the choice of block size[49, 76].

2.3 Forecasts and their verification

2.3.1 Introduction to forecasting

Forecasting in general is an iterative procedure: First, a model for the true process to be forecast is chosen either through the formulation of a theory or by inferring it from historical data. Then, predictions are issued for the future values the process will take. Finally, the quality of the forecast will be evaluated and this analysis used to modify and improve the model.

Due to the vast number of applications of forecasts, there is a correspondingly large number of distinct forecasting systems. They can be classified by the type of target variables (continuous variables, categories, dichotomous outcomes) and the type of output (deterministic values or probabilities[12]). Other characteristics of forecasts are the *validity period* and the *lead time*. The validity period designates the time period for which the forecast is issued: The forecast might predict the process value on a specific day, or its average over a month, a 3-month-period or a whole season. The lead time, on the other hand, characterises the time between issuing the forecast and the start of the validity period, i.e. it characterises how much in advance a forecast is issued. The lead time is used in atmospheric sciences to separate forecasts into short-range or weather forecasts of lead times less than a week, medium-range forecasts issued from a week to a month in advance, long-range or seasonal forecasts of lead time between 30 days and 2 years, and climate forecasts beyond that[28].[13]

The particular choice of forecast for any given problem will be based on the specific requirements of the prediction task. These might be, among others, the lead time, the computation cost and the available data, but also the necessary goodness of the forecast or the cost of inaccuracy[37].

But what characterises the goodness of a forecast? Murphy proposed the following three main criteria to define it: consistency, quality and value[78]. While the *consistency* describes the correspondence between a forecaster's judgement and the forecasts he actually issues, the *quality* is defined by the correspondence between forecast and observation. Finally, the *value* describes the benefits realised by decision makers when taking the forecast into account. The value therefore differs from user to user according to their specific application needs. Of course, the value also depends on the quality of the forecast, but other aspects do play a role as well. Good quality therefore does not automatically imply good value[78]. Indeed, if one were to accurately predict an event that is almost certain to happen, then forecast and observation agree most of the time, leading to good forecast quality. The almost certainty of the event, i.e. the high *base rate*, however, does not allow for good value of the forecast. As value is highly specific to the user and some aspects influencing the value, such as the base rate, cannot be improved on in the forecasting process, forecasters are usually only concerned with the quality for which objective and general criteria exist. These are assessed in the forecast verification process.

2.3.2 Forecast verification

To compute the quality measures for the verification, which are also called *scores*, large amounts of data in the form of forecast-observation pairs are needed. In order to avoid waiting a long time for the observations to accumulate for the verification process, usually forecasters will not issue forecasts in the strictest sense, i.e. predictions of *future values* exclusively. Instead, they issue hindcasts (also called retroactive forecasts or ex post forecasts[79]) pretending to operate

[12] For a good overview, see for instance [76].

[13] Of course, this classification is not quite suited to first passage time prediction problems, which strictly speaking have lead time one day (minimum first passage time) and validity period one year (maximum first passage time), but without characterising a yearly average. Despite the lead time of one day, they are better classified as seasonal predictions.

at a different time in the past from which forecasts will be issued. The observations for the verification process are then already available.

This practice also takes care of the need to verify the forecasts *out-of-sample* while retaining an adequate number of observations for the verification. Indeed, if the same data is used for building the forecast model and verifying the resulting forecast, artificial skill is introduced by adapting the forecast model to the specific verification realisation of the process that is to be predicted. Observation data should therefore always be separated into two distinct parts: a training set for building the forecasts and a test set for the verification[38, 74, 37, 79]. With hindcasts, when special care is taken to really only use data for the model building that was recorded before the date on which the hindcast is issued, a separation into a training set before this date and a test set afterwards is inbuilt. Often, a yearly update for the model is then used, building a model on the training set up to a specific date, verifying on the next year of observations after this date (or another adequate number of observations depending on the forecast validity period), using this verification to develop forecast system improvements, moving the specific date forward by a year and restarting the process. In this way, inflated skill through the existence of a trend that skews future observation distributions is also avoided[80]. While it should in general lead to a realistic indication of forecast quality, it might even underestimate the forecast skill as actual forecasts made after the evaluation can be based on the whole data set, i.e. on a much larger sample than the one used for the earlier forecast verification [74].

As stated before, a general assessment of the value of a forecast is not possible. Coming back to the example given then, correctly forecasting an almost certain event correctly lacks value because the forecast does not really improve the general knowledge already available to the user. In order to better analyse forecast goodness and get at least a small handle on the value in addition to the quality, one therefore should compare the forecast to an unskilled reference that is also called *benchmark*. This ensures that it cannot be beaten by a very simple forecast[38, 79]. In meteorology, such a benchmark usually consists of either predicting the climatology, i.e. the long-time mean value, or predicting persistence, i.e. a continuation of the presently observed deviations from the climatology[76]. Other possibilities include random chance or damped persistence consisting of a regression of present anomalous conditions towards the long-term average modelled by an AR(1)-process[80]. Alternatively one can choose to always issue the same deterministic prediction not necessarily determined by the long-term average, such as always forecasting "no event", when the probability of the event occurring is in fact very low[79].

The comparison of the forecast quality to a benchmark can be summarised using a *skill score*, which can be computed as defined by Eq.(2.29).

$$\text{skill score} = \frac{\text{score}_{\text{forecast}} - \text{score}_{\text{benchmark}}}{\text{score}_{\text{perfect forecast}} - \text{score}_{\text{benchmark}}} \tag{2.29}$$

Such a skill score specifies the relative improvement over the benchmark obtained by the forecast and leads to a value of 0 for forecasts that have the same quality as the benchmark when measured by the score, and to a value of 1 for perfect forecasts. However, for small sample sizes, the skill score can be unstable[76].

To compute skill scores for a verification process, first it is necessary to make a specific choice for the score. However, one has a wealth of possible scores at one's disposal. Which score is chosen depends among other issues on the impact of the different kinds of possible forecasting errors[14]. Depending on the goals of the forecast, some errors might be worse than

[14] We focus here on the errors contained in the forecast as is usual in forecast verification. Errors in the verification data, i.e. in the observations, are usually ignored due to the overall high signal-to-noise ratio in atmospheric data[76].

others: A user might want small average deviations from the observations or an absence of large "outlier" errors, or maybe he needs the forecast to reproduce the observed probability distribution accurately[76]. It also might be more important to forecast the correct timing of an event than the exact magnitude, or vice versa. For most of these different objectives, there exist key statistics and summary measures for the forecast verification, some of which we will introduce later.

The general criteria used for selecting an appropriate score for the specific forecasting context can roughly be separated into the different aspects of forecast quality that are measured by the score on one hand and some desirable properties of the score itself that are independent of the specific forecast on the other.

We will first focus on the different aspects of forecast quality. Generally speaking, the joint probability distribution of forecasts fc and observations obs, $P(\text{fc,obs})$, contains all relevant non-time-dependent information for forecast verification. Two attributes of forecast performance commonly used relate directly to the full joint probability distribution: accuracy and association. In terms of forecast quality aspects, the *accuracy* denotes the average correspondence between the forecasts and the observations, whereas the *association* only focuses on the linear relationship between forecasts and observations[49].

The full joint probability distribution can, however, be difficult to analyse, especially for high-dimensional forecast problems. Murphy et al.[81] therefore proposed two decompositions of the joint probability distribution using the conditional and marginal probability distributions. The first one, as given by Eq.(2.30), focuses on the potential forecast quality as it centers on the observations, while the second one given by Eq.(2.31) characterises the actual quality.

$$P(\text{fc,obs}) = P(\text{fc}|\text{obs}) \cdot P(\text{obs}) \tag{2.30}$$

$$P(\text{fc,obs}) = P(\text{obs}|\text{fc}) \cdot P(\text{fc}) \tag{2.31}$$

These decompositions show other ingredients of forecast quality. Some of them are only based on the marginal probability distributions. This is the case of the *uncertainty*, which is characterised by the sample base rate $P(\text{obs})$ (climatological probability) and is therefore a property of the forecast situation exclusively: It measures how difficult it is to predict the observations[78]. Indeed, if all possible outcomes are equally likely, then the forecast task is much harder than if the outcome will almost always be the same. The corresponding attribute that is based only upon $P(\text{fc})$ is the *sharpness*. It measures the variability of the forecast, i.e. whether the forecast tends to always stay close to the mean or whether it also issues more extreme predictions[78]. As the sharpness does not take into account any information on the observations, it does not allow conclusions about the forecast accuracy[49]. Another attribute using only both marginal distributions is the *bias*. It measures the correspondence between the mean forecast and the mean observation[49].

The other forecast attributes that are in use to define the quality are based on the conditional probability distributions of forecasts and observations. The *discrimination* is based on $P(\text{fc}|\text{obs})$ and measures how much the forecasts for different observations differ on average. Its partner is the *resolution* which is based on $P(\text{obs}|\text{fc})$ and measures how much on average for different forecasts the observations differ[49]. The last commonly used forecast attribute is the *reliability* of probabilistic forecasts. It is also based on $P(\text{obs}|\text{fc})$ but measures the agreement between the forecast probability and the mean observed frequency[78], i.e. looking at all cases in which the forecast probability for an event was p, did the event actually happen with probability p?

There exists therefore a large number of different aspects of forecast quality that can be measured using scores. Depending on the requirements of the forecast tasks, some of them might be more important than others, leading to different choices of scores.

The scores themselves are characterised by several properties that are independent of the

forecasting scheme. The most important one is *propriety*. A proper score will be optimal when the issued forecast corresponds to the forecaster's best judgement[74]. It can therefore not be hedged after the fact to obtain a better score, thus not allowing the verification system to exert undesirable negative influence on the forecasts[82]. If it is strictly proper, i.e. if there exists a unique optimum, then the best forecast is the one that forecasts the probability distribution from which the observations are in fact drawn[83]. This is therefore a highly desirable property for measuring forecast goodness for probabilistic forecasts[15]. In fact, propriety of a score directly encourages forecast consistency - one of its main desirable properties[78]. However, one needs to keep in mind that propriety of a score does not necessarily also imply propriety of the corresponding skill score[84].

Another score property that is often sought after is *equitability*. This demands that all "naive" forecasting schemes such as random chance or constant deterministic forecasts should obtain the same score[76]. In contrast to propriety, this is mostly relevant for deterministic forecasts, since finding a probabilistic score that is both proper and equitable at once is impossible[84]. Moreover, naive forecasts for probabilistic schemes do not generally have equal reliability so one would rather have a proper score which rewards a more reliable forecast such as the climatology instead of treating it exactly the same as a less reliable forecast such as random chance[74].

Another score property for probabilistic forecasts is the locality. *Locality* of a score means that it only depends on the probability assigned to the verification value. Local scores measure the sharpness and the reliability[74]. However, this restriction is not necessarily desirable: If ordered categories are forecast, then it might be relevant whether the forecast was off only by one category or by many categories. This is not reflected in local scores. Moreover, local scores usually drop when the forecast gets more precise, i.e. when there are more categories included, as then on average the probability assigned to each category and therefore also to the category containing the verification drops[74].

The scores we choose in this thesis will therefore all be proper, but not necessarily equitable or local.

Of course, any verification score obtained from a data set of forecast-verification pairs is only a sample estimate of the true value. Forecast verification thus turns into a problem of the sample size: It might be that the forecasts fit the observations uncommonly well or uncommonly poorly on the small available verification sample. In this case, the score might be misleading - a measure of the uncertainty associated with the score value is therefore needed.

This uncertainty will be larger the smaller the sample size employed in the score calculation, but it will also grow when the data has a high intrinsic variability or if the base rate is small, leading to a high variation in the forecast scores. In addition to the scores themselves, their standard error or an estimate of a confidence interval associated with the scores should therefore be given (see e.g. [74, 75, 76])[16].

Most analytical means of obtaining an uncertainty estimate assume both independence of the verification samples and a specific underlying distribution - mostly normal or binomial. This is not usually reasonable in the context of meteorological forecast problems[76]. The most expedient way of obtaining confidence interval estimates is therefore the bootstrap method explained already in Sec. 2.2.6[26].

As stated before, there are many scores that are valid for different kinds of forecasts and focus on different aspects of forecast quality. However, using a single score value to characterise forecast performance is equivalent to collapsing the high-dimensional verification problem onto a one-dimensional number - there is a corresponding loss of information[49]. Choosing just one

[15] Note that propriety is irrelevant for deterministic forecasts as then, the correct value will always be best and thus no hedging is possible[84].

[16] See Mason[74] for an explanation of why statistical hypothesis tests and p values are not the best choice for forecast verification.

overall measure of skill will therefore give an incomplete and potentially misleading impression of the forecast performance[74, 81, 85]. Indeed, if one forecast performs best regarding one score, this does not imply that it will be the best forecast in all aspects. To cover several requirements, one should therefore use several key measures instead of restricting oneself to just one criterion.

2.3.3 Some specific scores

After discussing the important aspects when choosing a score for forecast verification, we will now take a closer look at some specific scores. The aim of this introduction is not to give a comprehensive overview, if that were even possible, but only to introduce the few scores that will appear in this thesis and that focus mainly on forecast accuracy[17].

We will start with the scores applicable to deterministic forecasts. The most widely used measure of accuracy for continuous variables is the mean squared error or its square root, the *root mean square error (RMSE)*[49, 5]. It is defined by Eq.(2.32), where n denotes the total number of data points available for the forecast verification.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{fc}_i - \mathrm{obs}_i)^2} \tag{2.32}$$

The RMSE measures the overall accuracy of the forecast, i.e. the correspondence of forecast and observation. As it is a measure of the average magnitude of forecast errors, a perfect forecast leads to RMSE = 0 (in the same units as the observations) and forecasts are progressively worse with increasing value of the RMSE.

Since it is a quadratic scoring rule, it gives a greater weight to very large errors. This means that it encourages conservative forecasting, i.e. forecasts near the climatological mean, and its value might be more representative of outliers than of the forecast as a whole[76, 5]. This outlier sensitivity also leads to a lack of stability of the score for small samples[85]. The lack of indication of the direction of any deviations between forecast and observation is another drawback of this score[76, 85].

When estimating the skill of a forecast, the RMSE does not rate all no-skill benchmarks equally, i.e. it is not equitable[74]. Livezey[80] advises against the use of climatology or persistence as a benchmark in this case as their RMSE usually is very high, and recommends damped persistence instead. The World Meteorological Organisation on the other hand proposes exactly climatology or persistence as reference forecast for the RMSE[28].

If the deterministic forecasts are discrete rather than continuous, the most direct, simple and intuitive measure of forecast accuracy is the *proportion correct (PC)*[49]. It is sometimes also directly called accuracy, or percent correct (when multiplied by a factor of 100)[49, 76, 87]. The PC is defined in Eq.(2.33), where n denotes the total number of verification observations, K the total number of different discrete categories the observations can fall into and $n(\mathrm{fc}_i, \mathrm{obs}_i)$ the number of forecast-observation pairs where both belong to category i.

$$\mathrm{PC} = \frac{1}{n}\sum_{i=1}^{K} n(\mathrm{fc}_i, \mathrm{obs}_i) \tag{2.33}$$

The possible values the proportion correct can assume are between 0 and 1, with a perfect forecast leading to PC = 1 and the worst possible forecast to PC = 0.

[17] As it is generally not clear which score is best suited for which application[86], we chose to focus on the accuracy as the main forecast attribute.

The PC treats all categories equally and is heavily influenced by the most common one. This leads to problems when some of them are markedly less probable than others[49]. It should therefore always be compared to a benchmark, especially one that always forecasts the same category deterministically, as the PC does not score all references equally (and is thus not equitable)[76].

For probabilistic binary forecasts, i.e. when an event either occurs or not and its probability of occurrence is forecast, the most commonly used verification measure is the *Brier score (BS)*[49, 82, 88]. It measures the accuracy of the forecast through the mean squared probability error and is given by Eq.(2.34), where fc denotes the forecast probability for the event to happen instead of a deterministic forecast value as before, and $o = 1$ if the event subsequently happened or $o = 0$ otherwise[49, 82, 86][18]. n again denotes the total number of verification events.

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(\mathrm{fc}_i - o_i)^2 \tag{2.34}$$

The Brier score can assume values between 0 and 1, with BS = 0 denoting perfect forecasts and BS = 1 the worst possible forecasts. In operational meteorological forecast verification, score values of BS > 0.30 are taken to denote poor forecasts, BS $\in [0.10, 0.25]$ are the usually obtained score values, unless the event forecast is rare leading to BS < 0.1[85].

This differentiation between usual forecasts and those for rare events reflects the score's sensitivity to the climatological frequency of the event, i.e. to the inherent uncertainty of the forecast: The rarer the event to be forecast, the easier it is to get a good score without any skill[49, 74, 85].

Generally, the Brier score is the analogue to the mean square error, but in probability space. Both the MSE and the BS can be decomposed into three components measuring the reliability, the resolution and the uncertainty respectively[74, 89]. As with most quadratic scores, the BS makes no difference between under- and overforecasting and favours conservative forecasts, i.e. those with high reliability but low resolution[85]. This implies that the Brier score is not equitable either, indeed it favours any benchmark that relies on the climatological frequency instead of random chance or a deterministic forecast of always the more probable case, event or no event, happening[82]. It is, however, a strictly proper score[83].

The Brier score is most often compared to the climatology. The corresponding skill score and its decomposition are given by Eq.(2.35), with $\mathrm{BS}_{\mathrm{ref}} = r(1-r)$ if r is the base rate of the event[86, 90].

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}} = \frac{\text{resolution} - \text{reliability}}{\text{uncertainty}} \tag{2.35}$$

Contrary to the Brier score itself, this skill score is not always strictly proper - it needs to contain the climatology as reference forecast to display propriety[76, 90]. It can also be unstable when applied to small data sets due to its nonlinearity: The rarer the event, the larger the sample size needed for an accurate score estimation[76, 85].

Brier skill scores using the climatology as a benchmark often are very low compared to other measures, even going so far as to be negative. This is mostly due to the fact that the climatology is perfectly reliable and reliability is a major component of the Brier score[22]. They are therefore considered to be a "harsh standard" for forecasts and it has been suggested that negative scores might conceal the fact that there actually still is some skill in the forecast, especially if the forecast sharpness is high[90].

[18] Note that Brier[82] originally summed also over the complementary case of non-events, leading to exactly double the score as given in Eq.(2.34). The version presented here is therefore also known as the half-Brier score[49].

For discrete probabilistic forecasts with a larger number of categories, the *ranked probability score (RPS)* is the generally preferred proper scoring rule[49]. It measures the forecast accuracy in terms of the squared distance between forecast and observation in probability space[90], but uses cumulative probabilities and therefore takes the ordering of the different categories into account[49, 91]. Its definition is given by Eq.(2.36), where $\mathrm{fc}_{k,i}$ denotes the forecast probability of category k for the i-th forecast-observation pair and $o_{k,i} = 1$ if the i-th observation falls into category k and $o_{k,i} = 0$ otherwise[49, 91].

$$\mathrm{RPS} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{K-1}\sum_{j=1}^{K}\left[\left(\sum_{k=1}^{j}\mathrm{fc}_{k,i}\right) - \left(\sum_{k=1}^{j}o_{k,i}\right)\right]^2 \tag{2.36}$$

The perfect score again is RPS = 0, and its values range from 0 to 1.

The RPS reduces to the Brier score for $K = 2$[91], when the ordering becomes unimportant. It is therefore a generalisation of the BS to ordered multi-categorical events and the two scores and their respective skill scores have the same characteristics[85]. In particular, the RPSS with respect to the climatological frequencies is also considered a "harsh standard" for forecast comparison[90].

The last forecast score that will be explained here is not calculated directly in this thesis as we are only concerned with forecasts made for one station. However, the *anomaly correlation coefficient (ACC)* is widely used for the spatial verification of operational numerical weather forecast fields, especially over long lead times[49, 76] and thus mentioned several times. Its definition is given by Eq.(2.37), where the primed quantities denote anomaly fields, i.e. deviations from the long-time mean, with the index i in this case denoting the grid point, the overline a spatial mean and $\hat{\sigma}$ the sample standard deviation[92].

$$\mathrm{ACC} = \frac{1}{n}\frac{\sum_{i=1}^{n}(\mathrm{fc}'_i - \overline{\mathrm{fc}'})(\mathrm{obs}'_i - \overline{\mathrm{obs}'})}{\hat{\sigma}_{\mathrm{fc}'}\cdot\hat{\sigma}_{\mathrm{obs}'}} \tag{2.37}$$

The anomaly correlation coefficient measures the centered[19] correlation between the forecast field and the corresponding observed anomalies[92], i.e. the pattern correlation between forecast and observations[49]. As the ACC is not sensitive to any form of bias, it reflects potential rather than actual skill and rewards correct patterns over correct magnitude[49, 76].

The forecasts are better, the larger the anomaly correlation coefficient. While ACC = 0.5 corresponds to a forecast equal to the long-term climatological average, ACC = 0.6 is usually taken as a reasonable lower limit for spatial forecast fields that still contain some synoptic skill at least in the largest patterns, while ACC = 0.8 is desired for some skill also in the slightly smaller synoptic patterns[49, 93].

2.4 The North Atlantic Oscillation

Most of the first passage time forecasts in this thesis are based solely on measured temperature data and thus require no further explanation. However, in Chapter 6, we will use a second variable in addition that is less widely known, namely the North Atlantic Oscillation (NAO).

The NAO is the most apparent atmospheric pattern over the extratropic latitudes of the northern hemisphere and the only teleconnection pattern in this region evident throughout the year[94, 95]. As such, its influence on European surface temperatures can be expected to be substantial. Indeed, the NAO directly influences the probability of occurrence of the different

[19] The uncentered version is also sometimes used, however there are some issues with it so the major weather forecasting services employ the ACC as given here[92].

"Großwetterlagen" (general weather types) in Europe[96]. It has also been stated previously that over the northern parts of Central Europe and over Scandinavia a statistical forecast of the surface air temperature anomalies ought to be possible at least towards the end of February if the state of the NAO were known accurately[97].

However, this is quite difficult to achieve in advance. It has been found that the NAO has an almost flat power spectrum close to white noise[98]. It is indeed well-known that the extratropical climate system in general is chaotic, making the NAO much harder to predict than other atmospheric oscillations such as the El Niño - Southern Oscillation (ENSO) for example[30]. Also many fundamental mechanisms and climatic processes governing the NAO are still not well understood[94, 99]. Even though some climate models were able to predict the NAO a season ahead with some success[98], the NAO variability is generally not well reproduced. Most climate models tend to overemphasize the correlations between the NAO and atmospheric patterns over the Pacific[99], even though the link between NAO and ENSO for instance is rather weak[98]. The lack of correct incorporation into the full global circulation models together with the importance of its influence on European weather makes the NAO a prime candidate for a second variable to use in a data-driven approach towards seasonal temperature prediction.

Research into the NAO started very early on: The opposing mean winter temperatures in Greenland and Denmark caused by the NAO had already been known to Scandinavian seamen, but Danish missionary Hans Egede Saabye's diaries from 1745 are generally credited to be the first detailed description of the phenomenon[100]. While the research stayed mostly descriptive for a long time, the topic underwent a large increase in popularity starting in the nineties of the last century with the goal of developing seasonal climate forecasts, as evidenced also by several special publications dedicated to the topic such as a review paper[101], a whole Geophysical Monograph (no. 134) and a special issue of the journal *promet* published by the DWD in 2008. However, there are many open questions remaining.

In fact, there still exists no unique way of defining or describing the NAO[94, 96]. The common basis is that it is the main mode of sea-level atmospheric pressure in the North Atlantic[102] that consists of a low-pressure system located somewhere around Iceland or Greenland and a high-pressure system around the Azores or even as far as southern Spain. The NAO oscillates between two main phases, namely a lower than average low-pressure system together with a higher than average high-pressure system (positive phase, sometimes also called 'Greenland' below[103]) and a higher than average low-pressure system together with a lower than average high-pressure system (negative phase or 'Greenland' above)[101, 103]. This is illustrated in the schematic in Fig. 2.2. Very rarely, a true reversal, i.e. an Icelandic high pressure system with a low pressure system over the Azores, can be observed[101]. The fundamental time scale of the oscillation is around 10 days, in winter there are additional oscillatory components of time scales around 2 years and between 5 and 8 years. The oscillation even shows some decadal variability[94, 96, 104]. The change from the negative to the positive regime is usually quicker than back[105] and the negative cycles are on average less frequent but longer lasting than the positive ones[106]. The oscillation also shows significant seasonal changes: The mean pressure difference between the northern and southern centers in winter is around 18 hPa, in summer both the pressure difference and the variability are weaker[96]. Also in winter, there are actually two separate southern centers which merge in summer, when both the low and the high pressure system move westward[95].

So how does the influence of the NAO on European surface air temperatures as depicted in Fig. 2.2 come about? During the positive phase of the NAO, there is a large pressure difference between Iceland and the Azores, signifying a net displacement of air over the Atlantic. This gives rise to wind from the high to the low pressure areas, trying to equilibrate the gradient. Due to Earth's rotation and the resulting Coriolis force, this current gets deflected and results in stronger than average Westerlies (west winds) across the Atlantic[96, 98]. Any change in

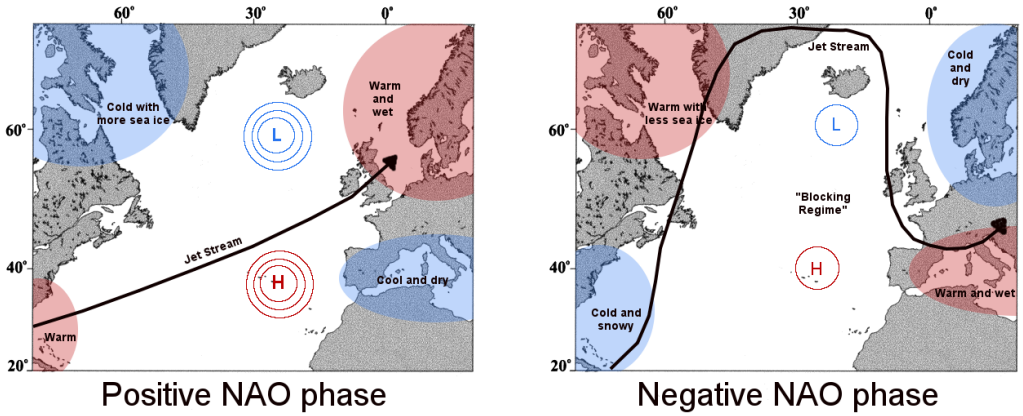


Figure 2.2: Schematic illustrating the two different phases of the North Atlantic Oscillation (NAO) and their general consequences on the weather of different regions around the Atlantic Ocean.

the phase of the NAO therefore means a change in the mean wind speed and direction and therefore also the heat and moisture transport across the Atlantic[94, 107]. Storm tracks, their intensity and number, sea surface temperatures and currents as well as sea ice cover are also affected[94]. The NAO can therefore serve as a measure of the strength of west winds over much of Europe[108]. Stronger west winds mean more moderate oceanic weather such as milder and wetter winters. Therefore the European climate variability both on the year-to-year and on decadal time scales is strongly correlated with the NAO[96]. While it influences both temperatures and precipitation[96, 109], it is especially important in determining hemispheric temperature anomalies[110], where winter temperatures gain around 1 ℃, locally up to 2 ℃, during a positive NAO phase[30]. In fact, the NAO explains around 75% of the total squared covariance between mean sea-level pressure anomalies and the European surface temperature[111].

Several detailed studies of the influence of the NAO on a more local scale have also been made, showing that most of the consequences for European surface temperatures are only valid for the Northern part of Europe[101]. Changes of the NAO phase are strongest reflected in surface temperatures over the Baltic states and northwestern Russia, but they are still quite pronounced over northern Germany, with a marked gradient from northeast to southwest[110]. Chen et al. have found that the correlation between the NAO and surface temperatures in Sweden is very high except in summer and that the NAO is an important factor in determining Swedish temperature variability[102]. Malberg et al. have analysed the temperatures in Berlin and found that the severe winters are due to a negative NAO phase, as then the inflow of cold air masses from a low pressure system over Eastern or Northern Europe is not hindered by the predominance of the Westerlies over Germany. There was only one notable exception when a severe winter occurred during a positive NAO phase: In the winter of 1953/54, the Icelandic low pressure system had moved quite far to the west, reducing its influence over Germany[105].

This strong relationship between European temperatures and the NAO seems to have evolved in time, with a stronger influence on winter temperatures during the years 1951 to 2000 than during 1901 to 1950[108, 112]. In addition to this trend-like component, there is also a seasonal effect: Even though the influence of the NAO on surface temperatures always exists, it is largest in winter, i.e. between December and March, when the NAO is the most important mode of spatial variability for European temperatures[94, 108, 113]. Tinz found that the correlations between European surface temperatures and the NAO are largest in February[96], while the NAO's greatest influence on Swedish temperatures is in March[102]. The coupling between surface temperatures and the NAO in summer, however, is rather weak[96, 105, 107], going

so far as to show almost vanishing correlations over Germany in July[96]. In fact, summer temperature anomalies are very correlated over the whole hemisphere, not just over Europe and the Atlantic region. They are therefore less influenced by the NAO than the more regional winter temperature anomalies[113].

An important question when looking at the influence of the NAO on surface temperatures is a possible lag between cause and effects. Indeed, a change in the NAO in winter precedes the corresponding change in sea surface temperatures by about a month[30]. However with land surface temperatures it has been found that in general simultaneous correlations explain most of the variance[110], and on a more local scale in Sweden there is also practically no delay of the NAO impact on the temperatures[102].

One problem has been the need to adequately describe the different phases and magnitude of the NAO to evaluate correlations with surface temperatures and quantify the different effects. Just as there is no unique definition of the atmospheric pattern, there also is no unique index to quantify it. Li et al. compiled a good overview[95] which shows that the indices in use can be separated into two main categories: Those that are based on instrumental measurement records of different quantities and those based on more complex calculations usually also involving climate model output or at least data assimilation into climate model space.

The simpler indices from instrumental records are based on sea level pressures, surface air temperature or sea surface temperature[95]. While van Loon and Rogers used the surface air temperature difference from Jakobshavn (Greenland) and Oslo (Norway) following the historic discovery of the phenomenon[103], a quantification of the NAO based on the sea level pressure difference between the two action centers of the Azores High and the Icelandic Low has become the most widely used version[95, 96]. Different measuring stations are favoured for this index, most notably Reykjavik, Stykkisholmur or Akureyri on Iceland and Ponta Delgada (Azores), Lisbon (Portugal) and Gibraltar (southern Spain)[95, 96]. Some rarer combinations include Gibraltar and Bergen (Norway) in winter or the coast of the Baltic Sea (for example Stockholm in Sweden) and Iceland[96]. In all these cases, the pressure time series should be normalised by subtracting the mean and dividing by the standard deviation over some reference period in order to avoid the index being dominated by the northern station data whose variability is about twice as large as for the southern stations [94, 108, 109].

Some of the more involved indices also use sea level pressures but not from only two land-based stations. Instead they calculate pressure differences between fixed latitudes[96] or quadrants, or between averaged values along defined latitudes across the Atlantic Ocean[95], needing the whole pressure field. Some also consider rotated principal components of the mean 700hPa height on the Northern hemisphere or $100 - 700$ hPa thickness differences between two stations[95], evidently needing model output or assimilated data.

Many studies have been made to evaluate the respective advantages and drawbacks of each index. The measurement-based indices do not involve the large numerical effort of model-analysis-based indices[95] and are both easy to implement and easy to interpret[101]. They also have the added advantage of being available for earlier years and therefore for longer time spans[107]. However, they do have several drawbacks. Any NAO index based on local station data will contain more noise since local, small-scale and transient meteorological phenomena that are not related to the NAO are not averaged out[112]. Another disadvantage of the local scale of station-based indices is that they might not be able to capture the spatial and temporal variations adequately[95, 101, 109]. Indeed, the NAO action centers wander significantly between seasons, moving westward in summer when the two distinct southern centers present in winter merge, so two fixed chosen stations might suddenly be quite far from the NAO centers[95, 101]. For seasonal forecasts, indices based on empirical orthogonal functions (or principal components) containing more spatial details would therefore in principle be preferable[107].

The specific choice of index made in this thesis will be discussed further in Sec. 6.1.

3 Temperature data analysis

3.1 Introduction

Our goal is to demonstrate the predictive skill of statistical schemes on seasonal time scales in the extratropics. As stated in Chapter 1, we chose to use the example of forecasting the first passage time to frost, thus basing our analysis upon temperature data.

Climate in fact comprises many different variables that are all interconnected and influence each other, contributing towards the actual state of the weather. This makes the choice of which variable to study for the chosen goal a priori an extremely subjective one. However, the most accessible and directly meaningful variables especially in applications are air temperature and precipitation[16]. As already seen in Chapter 1, they are also the most widely needed in terms of seasonal forecasts.

Temperature usually has a higher signal-to-noise ratio than precipitation[114]. While it is subject to small-scale variability and noise with fluctuations of up to 2 ℃ in a few seconds, their influence on the time series can be significantly reduced by choosing measurement devices with time constants of around 20 seconds, i.e. a 'reaction time' to changes in temperature, as is operationally done by the national weather services[115]. Undesirable effects caused by the sensitivity of temperature measurements to local environmental details such as vegetation, buildings and the specifics of the radiation screen around the measuring device are also easily reduced and altogether avoided by careful choice of the measuring site. This leads to a high accuracy of 0.2 ℃ for timed measurements, and 0.3 ℃ for measurements of daily extrema[115].

Precipitation on the other hand is slightly more problematic. While it is recorded to the nearest millimeter, measurements often underestimate the true value due to wind, evaporation (especially in dry areas) and equipment design. This can amount to a relative error of up to 50% in extreme cases with estimates saying that amounts on a global basis are underestimated by about 15%[116].[1] Precipitation measurement accuracy also depends on wind speed and rain intensity. Moreover, the actual amounts vary significantly on very local spatial scales, not only due to specific topography, especially in showery weather[116]. As a result, precipitation has not only a much greater spatial variability than temperature, but also a very large year-to-year variability in the total amounts[114]. Moreover, the precipitation time series are intermittent, having persistent dry and wet spells. It has also been claimed that they are not long-range correlated[117]. This makes for a much more complicated prediction task that needs to take into account many other meteorological variables to be better than simple persistence forecasts[118].

Since the analysis in this thesis is intended as a proof of concept for predictability using the time series only, we chose the simpler case of data that has less dependencies on other variables, longer intrinsic correlations and also less measurement errors, i.e. surface air temperatures. Moreover, it has previously been claimed that statistical forecasts of surface air temperature anomalies ought to be possible[97] - a claim for which we want to provide more solid backing.

As already explained in Chapter 1, fully time-resolved predictions on seasonal timescales, while highly desired in applications, prove to be very difficult. To nevertheless obtain some predictive skill, the common approach up to now has been to only forecast time averages.

[1] These figures reflect the state of the art in precipitation measurements almost two decades ago and several improvements have been made in the meantime. However, since the analyses presented in this thesis will take into account time series that start much earlier, these problems are very relevant.

An alternative that provides forecasts of non-averaged timing without trying to deliver a level of detail that is unattainable are predictions of first passage times to a threshold crossing. They are an important macroscopic quantity that characterises the dynamics of a system and have thus been increasingly studied with respect to applications as diverse as finance[119, 120], integrate-and-fire neurons or biological population strengths[120].

A forecast of the first passage time to some event is in fact most important when you need to be prepared before it happens but acting too early is costly. This is, for instance, the case with hurricane landfall and the resulting need for evacuations.

When considering surface air temperatures, a special threshold is given by frost, i.e. 0 °C. Knowing when it will next freeze is important for many applications: In agriculture, advance knowing of the first frost of the year helps for example with the timing of the wine harvest which needs to be as late as possible to profit from the sun but early enough for the grapes not to freeze to death (except for the production of ice wine). For transportation, the stocks of salt for de-icing roads need to be replenished on time and snow removal machines maintained. Over the course of the winter, logistics enterprises need to schedule large transports during times when the roads are least treacherous and construction work could take place in winter after all, if there is no frost expected for a long enough time to allow concrete to harden fully for instance.

In spring, knowing whether frost is still expected to happen before the summer is crucial in many sectors. In agriculture, planting or shearing sheep too early could lead to the crops or lambs freezing to death. Another well-studied problem is bud freezing in orchards[121, 122]. Here, frost protection measures such as wind machines or overhead sprinklers are costly and the equipment takes up space outside and is error-prone. Installation and maintenance could therefore profit from longer-term advance warning.

Forecasting the first passage time to frost is therefore a very relevant task for many applications. But before we start to prepare issuing actual predictions, we will first choose and analyse the temperature data we will base the forecasts on.

3.2 Data preparation

3.2.1 Provenance and Quality

The German national weather service 'Deutscher Wetterdienst'(DWD) provides large quantities of meteorological measurement data through their website[1]. It contains long time series with measurements of many variables such as atmospheric pressure, wind direction and speed, relative humidity, and precipitation, to name but a few. Especially for air temperatures it has a wealth of information, among others the daily minimum and maximum temperatures and thrice-daily measurements, roughly the morning, noon and evening temperatures. All of these are measured according to international standards at 2 metres above ground level to avoid the influence of soil temperature and animal interference[115].

The time series have very different lengths and originate from 79 different weather stations in Germany, which are located in many different environments such as larger cities like Berlin, mountains like the Fichtelberg at an altitude of 1213 m, or the sea coast on islands like Norderney or aboard lightvessels. This results in an embarras de richesse when it comes to choosing specific measuring stations to work with. The ones near the sea usually show a smaller variance in the temperature due to the large heat capacity of the ocean and the resulting persistence of its temperature[123, 124]. The stations in mountainous regions possibly show a larger dependence on the specific topology of the mountain as well as on the precise altitude of the stations. This means that stations located towards the center of Germany which have sufficiently long time series should be chosen.

The Potsdam station fulfills these requirements especially well. It is situated on top of the

	1893 - 1961	1962 - 1966	1967 - 1990	1991 - March 2001	April 2001 - 2010
morning temperature	7:00 MST ≈ 6:08 UTC	7:00 CET ≡ 6:00 UTC	6:00 UTC = 6:00 UTC	7:30 CET ≡ 6:30 UTC	6:00 UTC = 6:00 UTC
minimum temperature	21:00 MST ≈ 20:08 UTC	21:00 CET ≡ 20:00 UTC	18:00 UTC = 18:00 UTC	21:30 CET ≡ 20:30 UTC	0:00 UTC = 0:00 UTC

Table 3.1: Measuring time over the years for two time series of the Potsdam station as given by the DWD[1]. MST indicates Mean Solar Time, CET Central European Time and UTC Universal Time, Coordinated.

Telegrafenberg, a hill of 81 m altitude above mean sea level on the outskirts of the city. This station has been measuring without interruptions since January 1893, resulting in time series of almost 120 years length, or 43.098 data points with daily sampling frequency.

Since the objective of the next chapters is to predict the occurrence of first frost, the daily minimum temperature and the morning temperature as the time closest to the daily minimum are the two most relevant variables. We therefore have narrowed down the choice of data to two distinct time series: Daily minimum and morning temperatures at the Potsdam station.

Before actually using this measurement data, the quality of the time series should be assessed. The DWD offers some rudimentary information stating that the data is mostly without systematic checks or corrections but with single points that were corrected or confirmed; for some limited time spans, however, systematic checks were carried out, especially after the introduction of new measuring devices. However, these quality assessments are very vague and do not cover the more systematic errors that might arise from changing environmental conditions.

The common problem of increasing urbanisation around a weather station almost always leads to a marked rise in measured temperatures over time[125]. This should, however, not be so relevant for the chosen station since it is even today still located in a green and elevated area on the outskirts of the city. Gradual changes through instrument drift are usually corrected directly through regular recalibration. The existence of possible inhomogeneities in the data due to station measurement device, thermometer housing or observer changes, elevation or location changes, or observing time changes on the other hand should be investigated. This is especially true with such long time series which span periods of war and upheaval[116, 125].

A first general look at the data does not reveal any obvious inhomogeneities. However, the metadata indicates that while the measurements avoid introducing a measuring time change due to summer or daylight saving time, the Potsdam weather station has repeatedly changed the measuring time, which tends to produce systematic biases[125]. The specific measuring times[2] as provided by the DWD [1] are given in Table 3.1.

For the daily minimum temperatures, such a change is not expected to result in any abrupt jumps in the measured values. After all, the interval still spans 24 hours with the time change only shifting the end point of this interval. As the time most likely to register minimum temperatures is around 3:00 am, well away from the time range in which the changes in measurement time occurred, the consequences of the shift should be minimal. Looking more closely at the time series around the specific times listed in Table 3.1, one can see that there are indeed no jumps visible.

The morning temperatures, however, are a different case: Here a shift in measuring time with respect to the daily solar cycle might indeed matter, since after such a change the solar irradiation at the moment of measuring is different from before. However, this time series does not show any such jumps around the specific times listed in Table 3.1 either. Fig. 3.1

[2]For the minimum temperature, this time denotes the moment from which the measurements are counted for the next day.

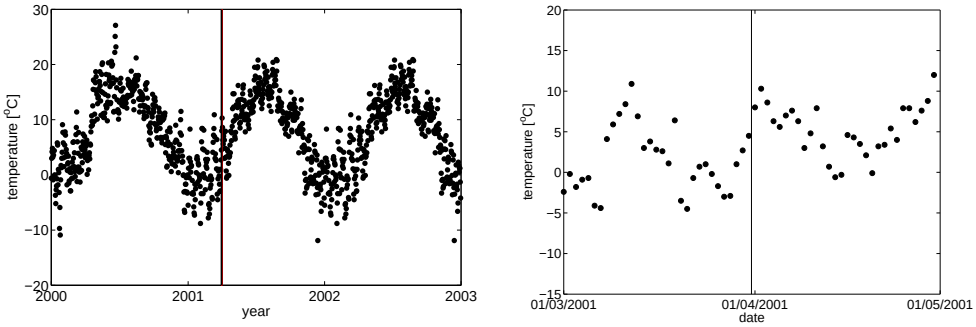


Figure 3.1: Two sections of the morning temperature time series from the Potsdam station. The vertical lines mark the biggest change in measurement time within the time series (from 6:30 UTC to 6:00 UTC).

illustrates this continuity in the Potsdam morning temperature time series around the last shift in measuring time, where one would expect lower temperature measurements after the time change when they are recorded earlier in the morning.

Therefore, at least the differences in measuring time are small enough not to have any discernible impact on the data and can be ignored.

Since we are looking at daily data in this thesis, which usually have larger homogeneity problems than, say, monthly averages[126], and since daily temperature extremes are particularly sensitive to changes in measurement[125], we should not rely solely on the metadata being complete to identify possible inhomogeneities. Indeed, Wijngaard et al.[60] found that 61% of the 158 European time series of daily temperature extremes analysed in the European Climate Assessment spanning the years 1946 to 1999 are of doubtful homogeneity. About a fifth of the apparent inhomogeneities could not be attributed to either known climate variations or changes documented in the metadata[60].

Since we do not have time series from suitable neighbouring weather stations, we cannot exploit the correlations between neighbouring stations following Karl et al.[127] to test the homogeneity. Statistical homogeneity tests applied directly to temperature time series are error-prone in the presence of trends or long-time correlations. It is therefore better to analyse the homogeneity of the time series variability for timed measurements or of the yearly average of the diurnal temperature range in the case of extremal temperature data[126, 60]. As can be seen in Fig. 3.2, these quantities fluctuate significantly for the Potsdam station temperatures, but no abrupt breaks appear.

We verified this following Wijngaard et al.[60] by using four different statistical tests that all assume as null hypothesis that the values are independent and identically distributed, namely the Standard Normal Homogeneity test, the Buishand range test, the Pettitt test and the von Neumann ratio test[3]. None of these tests rejected the null hypothesis of homogeneity at a significance level of $\alpha = 1\%$.

However, doing the transformation from extremal temperatures to diurnal temperature range, we discovered a slight problem unrelated to homogeneity in the Potsdam data set: The diurnal temperature range becomes negative on three occasions, i.e. the recorded minimum temperature is larger than the corresponding recorded maximum temperature. Comparing the temperature extrema to the timed measurements reveals that these seem to be single false recordings - twice the minimum temperature is larger than a timed measurement, once the maximum temperature is too small.

[3]For a more detailed explanation of these tests see Sec. 2.2.3

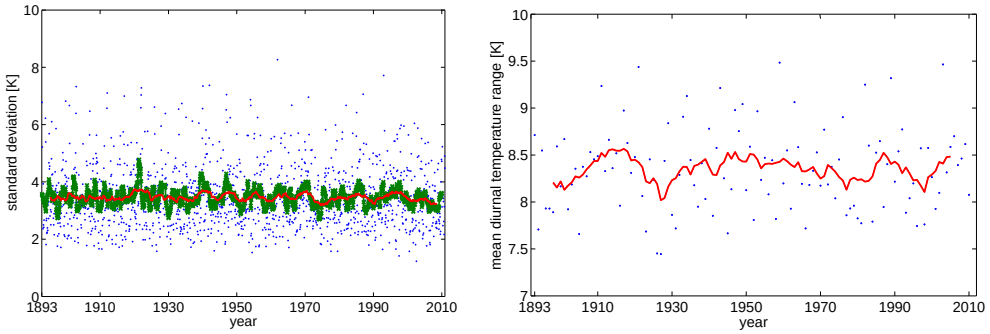
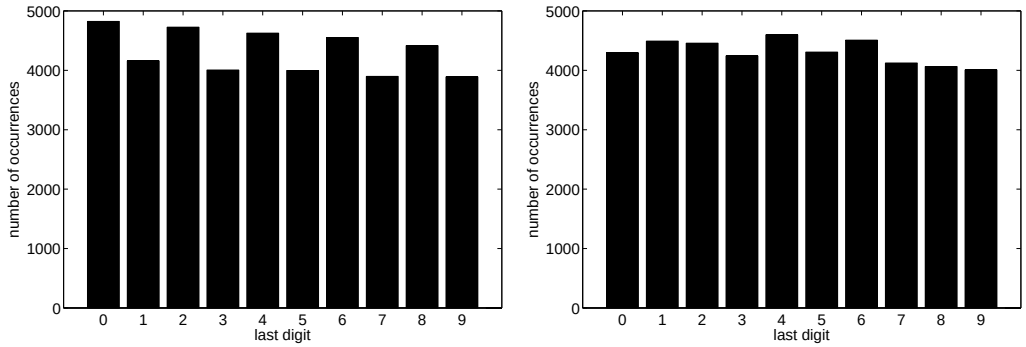


Figure 3.2: Left panel: Standard deviation of the morning temperatures recorded during one month (blue dots), as well as the moving average over twelve months (thick green crosses) and sixty months (red line). Right panel: Yearly average of the diurnal temperature range for the Potsdam station (blue dots), as well as the moving average with a window width of 10 years (solid red line).



(a) Potsdam morning temperatures

(b) Potsdam daily minimum temperatures

Figure 3.3: Histograms of the number of occurrences of the different possible last digits, i.e. the decimal places of the measured temperatures.

One other problem that often occurs in time series whose data collection is not yet automated is an interference of the psychology of the observer with the measurement. This often leads to a bias, where some values especially in the last digits - in this case the decimal place of the temperature - are preferred over others.

To check for this effect, Fig. 3.3 shows the incidence rates of the last digits for both time series, which should be very close to a uniform distribution for unbiased measurements. The daily minimum temperatures show a uniform distribution with statistical fluctuations. However, the morning temperatures show a systematic preference for even numbers in the last digit. This is almost certainly due to the use of the most common measuring device which is a mercury-in-glass thermometer with scale markings in increments of 0.2 K[115]. The measured morning temperatures are therefore only accurate to ± 0.1 K at most.

Thus the quality of the two time series is very good, a fact that is supported by the analysis of May et al.[123] who endorsed the quality of the Potsdam measurements in their study. We did find two isolated false data points in the Potsdam minimum temperature series. Since the timed measurements are less error prone in general than the extremal ones, we choose the daily morning temperatures from the Potsdam station for our predictability analysis. The only real issue encountered in this time series is a psychological one which slightly limits the accuracy

parameter	$\langle T \rangle$	a_1	b_1	a_2	b_2
value	6.60	-2.82	-8.47	-0.09	0.68

Table 3.2: Linear regression estimate in ℃ of the parameters for the climatology $\widetilde{T_j}$ (see Eq. (3.1)) for the Potsdam morning temperature time series.

of the data by preferring even numbers in the decimal place. However, this problem is smaller than usual in such time series, where half degrees are favoured markedly[38], and also smaller than the official accuracy of ±0.2 K.

3.2.2 Climatology and construction of anomalies

Due to seasonal solar forcing, the yearly component of the temperature spectrum representing the seasonal cycle causes the largest part of the variance in the time series, with sometimes also a semi-annual component[38]. Any temperature predictions therefore would already score quite well by always predicting the mean annual cycle, while the truly interesting information are the actual fluctuations around it. We will thus subtract the mean annual cycle, also called *climatology*, from the time series and in the following only consider the fluctuations around it, which are then called *anomalies*.

This has an additional benefit in our case. Conditioning first passage times on both the initial date and the initial anomaly reduces the amount of available data drastically, considering that there are only 118 years in our time series. It is therefore advisable to consider ways of grouping data together. In this way, better statistics can be obtained, but it also results in a more coarse-grained approach for our prediction task. Converting the time series to anomalies should make it pseudo-stationary[128] - at least over longer time spans than for the original temperature time series since there are still polynomial trends and unresolved smaller frequencies to consider. We could therefore relax the requirement for a specific initial day and instead take an initial anomaly regardless of the date on which it was recorded and then add the climatology for the initial date we want to consider.

First though, we need to determine the actual form of the seasonal cycle. We will proceed by calculating the average temperature for each calendar day, denoted in the following by $\langle T_j \rangle$ where j marks the number of the day within a calendar year. Then we fit a smoothed sinusoidal model of the climatology containing the two main frequency components of the spectrum (see also Sec. A.3.2), where $\omega = \frac{2\pi}{(365.2425 \text{ days})}$ denotes the yearly frequency in the Gregorian calendar and ϵ_j the residuals of the model fit[4]:

$$\begin{aligned} \langle T_j \rangle &= \langle T \rangle + a_1 \sin j\omega + b_1 \cos j\omega + a_2 \sin 2j\omega + b_2 \cos 2j\omega + \epsilon_j \\ &= \widetilde{T_j} + \epsilon_j \end{aligned} \tag{3.1}$$

The parameter values $c_i, i = 0, 1, ..., 4$, that were obtained using linear regression are listed in Table 3.2.

Fig. 3.4 shows the resulting climatology for the Potsdam morning temperatures T as well as the original measurements, illustrating their fluctuations around the seasonal cycle. These anomalies ΔT are computed as follows:

$$\Delta T_j = T_j - \widetilde{T_j}. \tag{3.2}$$

We will from now on focus on the resulting time series of temperature anomalies. However,

[4]For a more detailed discussion of the different definitions of the climatology used throughout the literature as well as the choice made here see Appendix A.

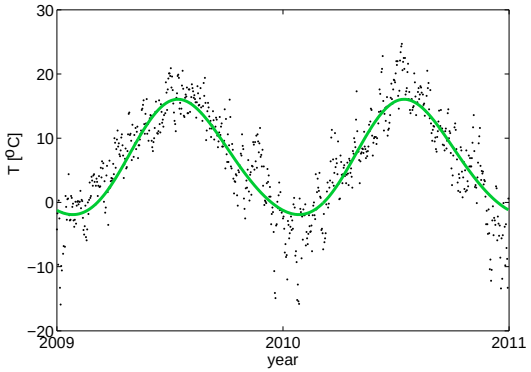


Figure 3.4: Morning temperature data of the Postdam station for the last two years of the time series (black dots), as well as the corresponding climatology (green line).

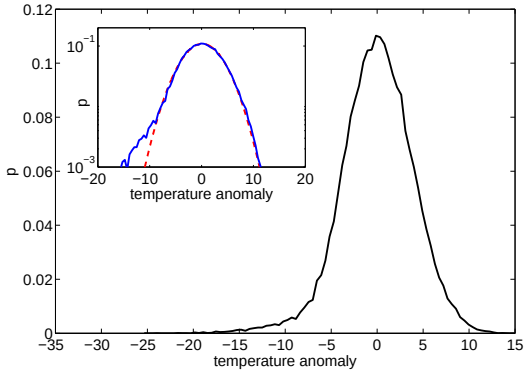


Figure 3.5: Probability distribution of the anomalies derived from the Potsdam morning temperature time series. The inset shows the data on a semilogarithmic scale with the closest fitting normal distribution represented by a red dashed line.

generating anomalies in this way does not eliminate slow trends such as gradual shifts in a time series' mean[32], implying a need to analyse the properties of the resulting anomalies time series. To determine whether the temperatures are well represented by this decomposition into seasonal cycle and anomalies, we will therefore take a closer look at the time series of the temperature anomalies in the following.

3.3 Time series properties

Since the anomalies could be considered as residuals after fitting the temperature measurements with the seasonal cycle, we would expect their probability distribution to be close to normal. Fig. 3.5 shows this to be the case, however there remains a substantial deviation from a normal distribution for the very negative anomalies that are much more frequent than expected.

Considering that we want to make use of the expected stationarity of the anomaly time series in order to enhance the statistics available for the calculation of conditional first passage times in the following, we need to make sure that these assumptions are met. We will therefore start by having a close look at potential time-dependences in the time series that would mean a significant violation of the stationarity assumption.

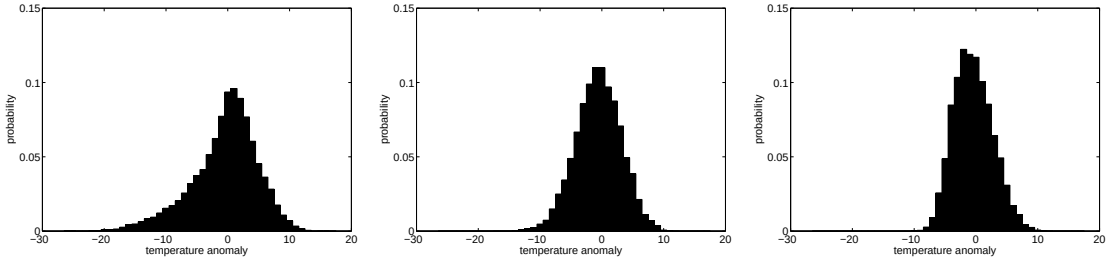


Figure 3.6: Probability distribution of the temperature anomalies recorded between December and February (left panel), in March and between September and November (center panel) and between April and August (right panel), using a bin width of 1 ℃.

3.3.1 Stationarity issues

Starting our stationarity analysis by looking at the full anomaly distribution for only a specific calendar month, we can already distinguish three general shapes: slightly left-skewed, almost symmetric, and very slightly right-skewed. Fig. 3.6 shows these different shapes using only anomalies from December to February (left panel), March and September to November (center panel), and April to August (right panel). This preliminary analysis shows that there are significant changes in the anomaly distribution with time, so that its stationarity is rather doubtful.

In the following, we will analyse the anomaly time series stationarity more thoroughly. As already defined in Sec. 2.1.1, a time series is considered wide-sense stationary if its first and second moments, i.e. its mean and variance, as well as its autocorrelation structures remain constant in time. We will therefore start by looking at the evolution of the first moment with time.

As anyone reading current newspapers knows, the question of global temperature trends is a generally very controversial topic. The latest report from the Intergovernmental Panel on Climate Change (IPCC) states that the existence of global warming especially over land is 'an unequivocal finding that holds up under a variety of methods and models'[129], with the only controversy the extent of the contribution of anthropogenic influences.

The magnitude of the trend is also generally unclear. Since the determination of any trend depends crucially both on the start and end dates of the analysis and for global ones also on the spatial averaging procedure, the actual value varies between publications.

For the Northern hemisphere it is believed to be around 0.13 ℃/decade for the period from 1951 to 1990[130], and 0.33 ℃/decade for the period from 1979 to 2006 [129]. Those two numbers already illustrate the apparent increase in the trend since 1995, with the 12 warmest years since 1850 being the years 1990, 1995 and 1997 to 2006[129]. Even though the warming is largest over continental North America and Asia[129] and trends are generally more significant in global data than in local data[131], we should expect to see its influence also locally in the German temperature series, introducing a non-stationarity.

As the conversion from temperatures to an anomaly time series simply shifted the mean and took out yearly and twice-yearly frequency components, any possible trends should have remained unchanged. Therefore, even though the origin and exact magnitude of the temperature trend are not of interest here, we need to get an impression of the importance of these nonstationarities in the data.

Using linear regression as illustrated by Fig. 3.7, we arrive at the trend magnitudes detailed in Table 3.3, where the only differences between the two approaches lie in the width of the confidence intervals due to the different sample size. It follows that over the 118 years in our time series, the mean temperature anomaly rose by about 1.15K.

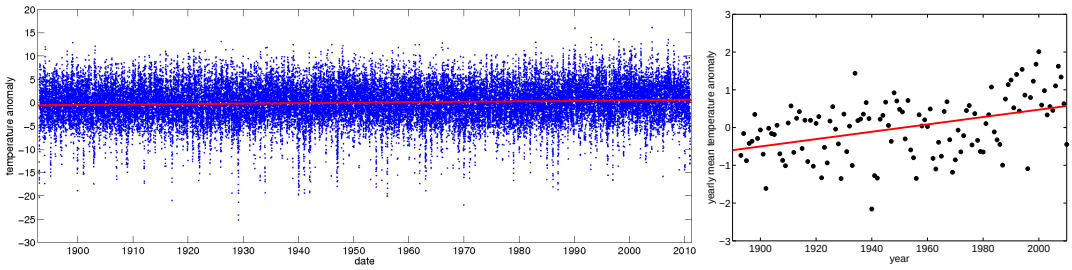


Figure 3.7: Left panel: Temperature anomalies (blue dots) and their trend obtained through linear regression (thick red line). Right panel: Mean temperature anomaly for each year (black dots) and their trend obtained as before (thick red line).

time series	trend	95% confidence interval	total warming	total warming range
measured data	$2.665 \cdot 10^{-5}$ K/day	$[2.366 \cdot 10^{-5}, 2.963 \cdot 10^{-5}]$ K/day	1.15 K	$[1.02, 1.28]$ K
yearly mean	0.009746 K/year	$[0.005918, 0.01357]$ K/year	1.15 K	$[0.70, 1.60]$ K

Table 3.3: Table detailing the linear trend present over the whole time period in the Potsdam morning temperature anomaly time series, obtained through linear regression.

One might speculate whether temperatures have evolved differently for the different seasons and whether there is therefore a marked change in the estimated linear trend when one considers only anomalies from a specific month or season. However, the range of estimated trends for different months and different seasons falls well within the 95% confidence interval for the general trend so there is no evidence for such seasonality in the linear trend.

We can therefore assume that the total shift in mean over the whole period covered by the Potsdam morning temperature time series is about 1.15K, or less than 0.01K per year, the time frame over which we will consider first passage times in our work. It is therefore only a very small and at first glance insignificant change. However, if we condition first passage time probability distributions on specific initial conditions, i.e. anomaly bands, then such a trend might change the relative weight of different years in the time series. For very cold anomalies, the earliest years in the time series might contribute many data points while these very cold anomalies almost do not occur anymore towards the end of the series. We will therefore investigate how large an effect the possible trend in the data has.

If the trend did not influence the selection of anomalies through different criteria, the fraction of anomalies that exceed a given threshold ΔT_{thr} selected from the first half of the time series, $p(\Delta T > \Delta T_{\mathrm{thr}} | y < 1952)$ with y the year the anomaly was recorded, should fluctuate around 50% for any given criterion. As can be seen from Fig. 3.8 for the more extreme anomalies, however, it is significantly lower than 50%, going as far down as 0% if the most extreme few anomalies are considered.[5] Therefore the trend in the anomaly time series does indeed influence the relative weight of different years when looking at extreme values.

Since we are going to look at anomaly bands, we also analyse in the following the relative influence of different years for all deciles of the anomaly distribution, not only for the most extreme cases.

As can be seen in Fig. 3.9, all deciles are influenced by the trend. This means that even though the magnitude of the trend is small enough not to be relevant over the time periods of

[5] The different appearance of both curves in the left panel of Fig. 3.8 is due to the asymmetry of the underlying anomaly distribution - a fraction of 0.0023 of all anomalies (i.e. 100 anomaly data points in total) corresponds to a threshold of 10.5K for warm anomalies and -15.3K for cold anomalies, causing the horizontal shift that disappears if one considers equal fractions of anomalies (as shown in the right panel).

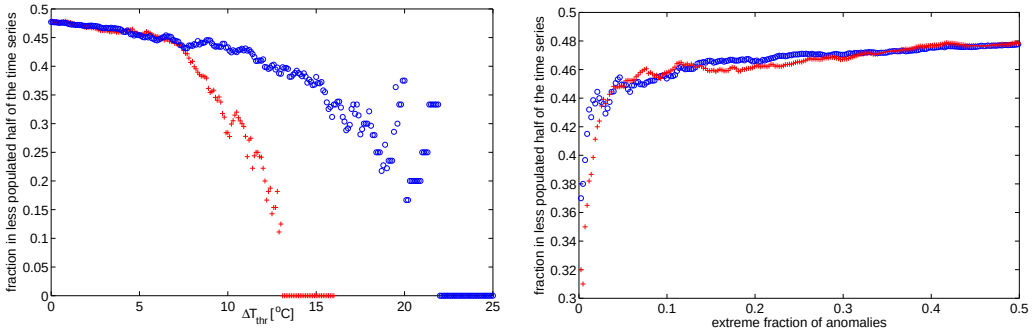


Figure 3.8: Dependence of $p(\Delta T > \Delta T_{\mathrm{thr}} | y < 1952)$ (left panel, red crosses), as well as $p(\Delta T < -\Delta T_{\mathrm{thr}} | y > 1951)$ (left panel, blue circles) on ΔT_{thr}. The right panel shows the same but the threshold is given in terms of a fraction of all anomalies.

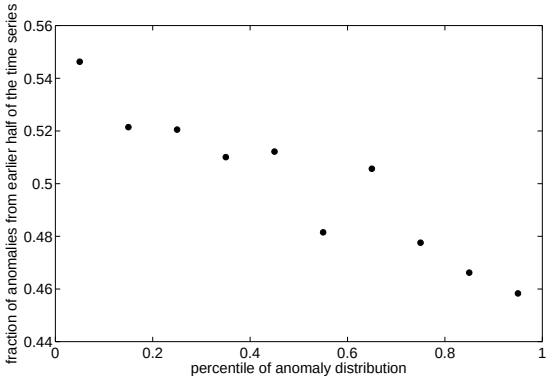


Figure 3.9: Fraction of anomalies in each decile that was recorded in the earlier half of the time series, i.e. between 1893 and 1951.

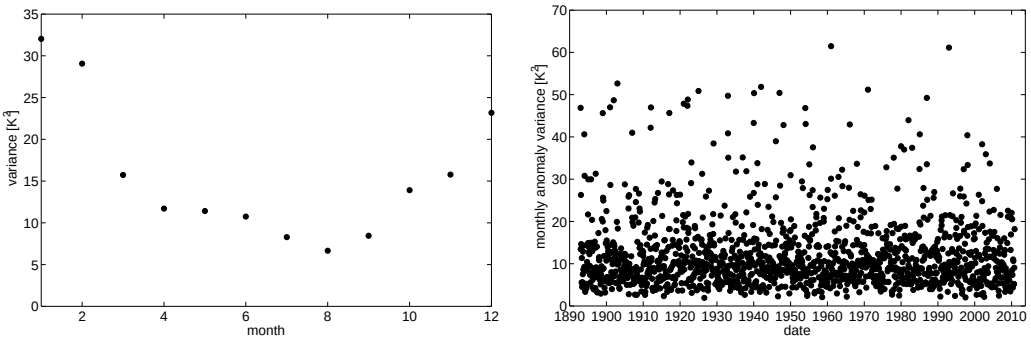


Figure 3.10: Variance of the anomalies grouped together for each calendar month (left panel), as well as for each consecutive month in the time series (right panel).

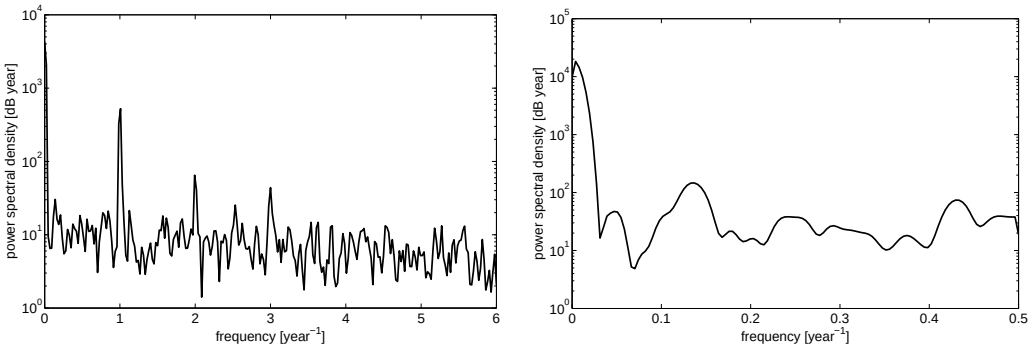


Figure 3.11: Power spectral density estimate of the monthly anomaly variances on a semilog scale, obtained using Welch's periodogram method with 5 segments of length 512 data points (left panel) and of the yearly anomaly variances using three segments of length 64 data points (right panel), both with a 50% overlap between consecutive segments.

less than a year that are of interest in this thesis, it will nevertheless play a significant role in the selection procedure of anomalies for the conditional first passage times considered later on in Chapter 4 - a caveat that needs to be kept in mind throughout the following analysis.

After looking at trends in the mean of the time series, we will now turn to examine any nonstationarities in the variance[6]. As the climatology was not rescaled by the seasonal variance, some seasonality remains, as shown in the left panel of Fig. 3.10 and as discussed also in Fig. A.3. Both show that the variance is larger in winter than in summer. Applying Levene's test - a more powerful alternative to the F test for nonnormal underlying probability distributions (see also Sec. 2.2.5) - to the anomaly variances grouped by calendar month confirms that these differences are indeed statistically significant. This phenomenon was already noted before for temperatures on the Northern hemisphere[132], and it is visible here even though it has been found to be much less pronounced in Central Europe than elsewhere[133]. So there is definitely some nonstationarity in the variance across different calendar months. Computing the variance instead for each consecutive month in the anomaly time series, there is no trend or oscillation immediately visible across the different years, as can be seen in the right panel of Fig. 3.10.

To better pick up any possible oscillations, we also look at the power spectral density estimate of these anomaly variances. The left panel of Fig. 3.11 shows that there are three distinct peaks at periods of twelve, six and four months, although the last one is barely distinguishable from the larger fluctuations in the estimate[7].

Since the different seasons seem to influence the anomaly variance heavily, we also group the data by years to analyse the smaller frequencies that might be overshadowed in the more finely split data. The right panel of Fig. 3.11 shows the power spectral density estimate for this yearly data. As can be seen, there are additional periods of around 2.3, 8 and 20 years (frequencies of 0.43, 0.125 and 0.05 per year) visible here. Therefore the anomaly variances still contain oscillations and are therefore not entirely stationary.

[6] Especially for time series with outliers or quite skewed distributions, employing order statistics such as the interquartile range instead of the variance to characterise the second moment of the underlying distribution is more robust. However, the following results were not changed qualitatively in any way by using the interquartile range instead, so we chose to include the more common measure of the two in this thesis.

[7] This estimate has high variance since it only averages over five segments, but varying the total number of segments did not lead to any new insights.

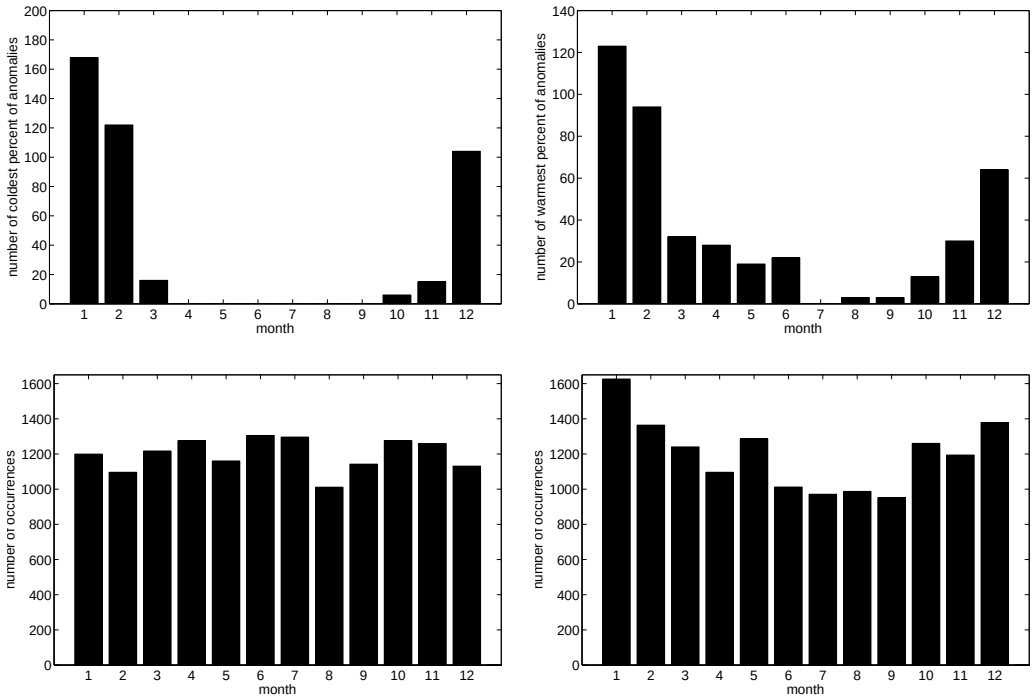


Figure 3.12: Histogram showing the occurrence of the coldest percent (top left panel) and the coldest tercile (bottom left panel) of temperature anomalies across the calendar year. The right panels show the occurrence of the corresponding warmest anomalies.

One resulting problem, just as with the slight trend observed before, is of course the influence a change in variance has on the selection of anomalies into bands, when the time series is treated as if it were stationary. Fig. 3.12 illustrates just how much certain months are favoured and others severely under-represented if only the coldest or warmest percent of anomalies were considered in an analysis. Even if one does not condition on the most extreme anomalies but considers the coldest and warmest terciles, there are still large deviations from the uniform distribution expected for a time series whose variance has no seasonal component at least for the warmest tercile.

Thus a selection of anomaly bands made regardless of the date on which the anomalies were recorded would result in an ever higher influence of winter measurements the more extreme the selection. Any prognoses made for initial dates in the other seasons would therefore be doubtful. A way to remedy this while still retaining better statistics would be to select anomalies that were recorded within a certain time window around the starting date under consideration but in any year within the record. This constitutes a compromise between high variance due to low statistics if one were to condition on the exact date, and high bias because the difference in variance on different dates would still allow effects that are only valid on other dates than the one in question. We should, however, bear in mind that while this approach is a good compromise there are still inherent stationarity problems with regards to the first few moments.

The chosen time series of daily morning temperatures recorded at the Potsdam station therefore contains both a positive trend as expected because of global warming and a yearly seasonal cycle, but also a twice-yearly frequency component. The temperatures approximately follow a normal distribution, but with a slight bimodality in the peak and a larger tail for negative temperatures.

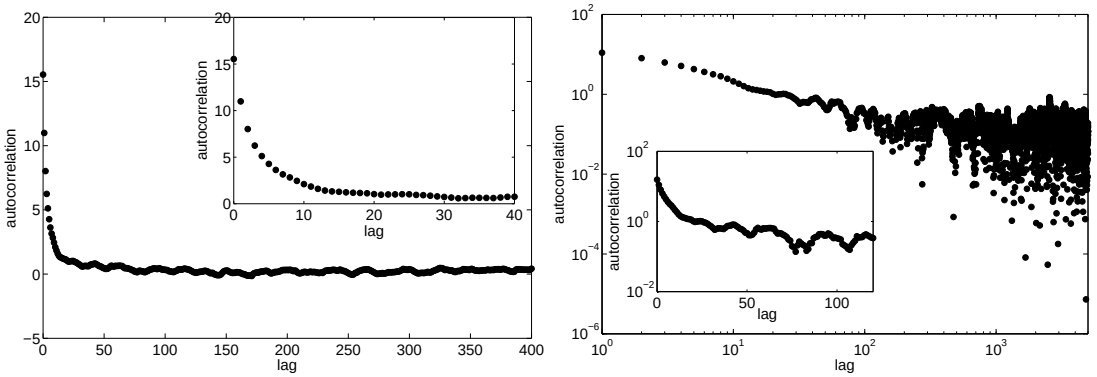


Figure 3.13: Autocorrelation function of the anomaly time series for two different zoom depths in the figure and inset of the left panel, as well as on a double-logarithmic scale (right panel figure) and a semilogarithmic scale (corresponding inset).

3.3.2 Correlations

Another very important aspect of any time series, and not only in terms of its stationarity, is its autocorrelation structure. It contains information on the extent of the memory of the underlying dynamical process and thus also on the potential for predictability of future measurement values.

If the climatological model truly fitted the temperature time series well, the anomalies would simply be the residual measurement noise realisation. As such their autocorrelation ρ_k is expected to decay exponentially and then show no further deterministic variations. However, it is well-known that temperatures show persistence, i.e. a large positive autocorrelation and a 'red' spectrum that is not incorporated into the climatological model, at least on very short time scales of several days (see for example [134]). Moreover, also long-term memory has been found in temperature data sets[117, 124, 32, 133, 34, 135, 33] and recreated quite well in the atmospheric models by incorporating volcanic forcing[33].

Looking at the autocorrelations in the Potsdam anomaly time series, we would therefore expect to see a power law decay. However, the left panel of Fig. 3.13 shows a very rapid decay of the autocorrelation function towards zero[8]. The right panel shows the same data on a double-logarithmic scale, where the power law decay associated with long-range memory should translate into a straight line. As can be seen, it does not seem truly linear. We therefore also plotted it on a semilogarithmic scale in the inset, where a straight line would indicate an exponential decay associated with an absence of long-range correlations in the data. Here, the linearity assumption is also only a rough approximation for approximately the first 12 lags before breaking down completely.

To get a rough idea of the characteristic correlation times of the temperature anomalies, we assume an exponential decay for the autocorrelation at very small lags. Using linear regression on the logarithm of the autocorrelation function, we extract an exponent of roughly $\lambda \approx -0.17$ from the data, corresponding to a characteristic correlation time on short time scales of five to seven days.

One problem with this direct calculation approach to the autocorrelation is the mixture of nonstationarities in the data such as the slight trend (see Sec. 3.3.1). The finiteness of the time series also causes large fluctuations at large lags, as can be seen in the right panel of Fig. 3.13. This leads to a sometimes incorrect and rather difficult evaluation of the true correlation structures[133].

[8] It stays above zero even at very large lags in this case as the warming trend adds long-range correlations.

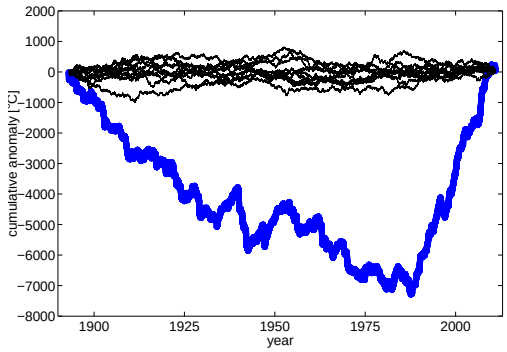


Figure 3.14: Cumulative anomalies as defined in Eq. (3.3) for the anomaly time series (thick blue line) and ten different surrogate realisations (thin black lines).

Several methods have been developed to better deal with these issues, most notably Detrended Fluctuation Analysis (DFA - see e.g. Rybski et al.[133], and references therein). One very simple but illustrative possibility, however, is to convert the anomalies to their cumulative values, as defined in Eq. (3.3) following Koscielny-Bunde et al.[135], and look at the resulting fluctuation landscape Y.

$$Y_j = \sum_{i=1}^{j} \Delta T_i \tag{3.3}$$

For uncorrelated or short-range correlated data, one would expect that the cumulative anomalies fluctuate around zero with a fluctuation amplitude directly related to the correlation time. Fig. 3.14 shows that the actual fluctuation landscape deviates very far from the zero line. For comparison, the fluctuation landscapes of ten different surrogate realisations, where any correlations were destroyed by random shuffling of the temperature anomaly time series, are also included in the figure. As can be seen, there are indeed significant long-range correlations in the anomalies.

Moreover, one can also see that the anomalies tend to be too negative at the beginning of the time series and then rapidly rise again towards the end. This directly reflects the trend in the mean that was not removed with the climatology (see also Sec. 3.3.1).

We have therefore seen quite a few stationarity issues in the Potsdam morning anomaly time series. In fact, there is a slight trend in the mean of the anomaly time series that is at least partly responsible for the presence of long-range correlations in the data. Moreover, the variance still contains a distinct seasonality in its magnitude and the skewness of the anomaly distribution also changes with the calendar months.

3.4 Modeling with an autoregressive process

The main problem with the data-based predictions, even if we treat the anomaly time series as if it were stationary enough to use the data irrespectively of the calendar day on which they were recorded, will remain the scarcity of data points and the correspondingly poor statistics once we condition on several distinct quantiles of different variables. Therefore it would be advantageous to find a very simple model that can reproduce as many of the characteristics of the temperature anomaly time series as possible without the computation costs of the large dynamical models presently used for prediction in weather and climate. Using such a simple model, we could compute as many simulated data points as needed, also for the rarer initial conditions.

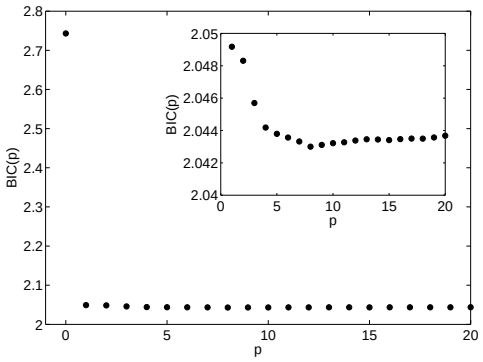


Figure 3.15: Modified BIC criterion by Hannan and Quinn[53] to determine the optimal AR(p) model order for the Potsdam morning temperature anomalies. The inset shows the same data as the full figure but excluding $p = 0$ in order to make the differences between the criterion values for the higher model orders visible.

One of the simplest models that contain some memory - necessary because of the persistence in temperature data (see Sec. 3.3.2) - are autoregressive processes (as described in Sec. 2.1.4). They are widely used for their ease of computation as well as their comparability to discretized ordinary differential equations with noise. It has been found that indeed many atmospheric variables are well-approximated by low-order autoregressive (AR) models[38, 32], at least when looking at high-frequency transients[134] and that predictions with short lead time of temperature anomaly threshold crossings have some skill when only relying on an AR(1)-process[136]. We will therefore explore the possibilities of reproducing the time series using models from this class.

The short-term correlation time of 5 to 7 days found in the previous section hints at a model order between AR(5) and AR(7) for our data. One problem we need to keep in mind, however, is that the autocorrelations in our data are not actually decreasing exponentially beyond approximately lag 12 as would be the case for an AR model of any order. This puts the adequacy of an autoregressive model especially on longer time scales somewhat into question[50], making a thorough analysis of the model fit necessary.

In order to determine the actual optimal order for an AR(p)-process to represent the Potsdam morning temperature anomalies, we fitted the time series using the Yule-Walker equations and then employed several well-established objective selection criteria, namely the Akaike Information Criterion, as well as the Schwarz and the Hannan and Quinn versions, and the estimated AR coefficients themselves.[9] The dependence on the model order p of these criteria is shown in Fig. 3.15 using the Hannan and Quinn criterion as a representative example, since the results are qualitatively equal for all criteria: They agree on an optimal model order of $p = 8$.

However, as can be seen in Fig. 3.15, the largest improvement by far occurs by changing the model order from $p = 0$, i.e. white noise, to $p = 1$. Any additional changes have a very small impact on the criteria, so that the simplest non-trivial AR model should represent the temperature anomalies already quite well.

Applying an AR(1) fit to the temperature anomalies, we arrive at the following model:

$$\Delta T_n = \gamma \Delta T_{n-1} + \epsilon_n, \tag{3.4}$$

with $\gamma = 0.707 \pm 0.005$ and a noise variance of $\sigma_\epsilon^2 = 7.76K^2$. The error estimate for γ represents the 95%-confidence interval which is therefore very narrow indeed.

[9] For more details on these procedures, see Sec. 2.1.4.

i	1	2	3	4	5	6	7	8
α_i	0.680 $\pm$0.001	-0.006 $\pm$0.007	0.024 $\pm$0.008	0.024 $\pm$0.007	0.009 $\pm$0.006	0.006 $\pm$0.008	0.005 $\pm$0.004	0.021 $\pm$0.002
rel. change [%]	0.2	117	33	29	67	133	80	10

Table 3.4: Parameter values for an AR(8) model (see Eq. (3.6)) of the Potsdam morning temperature anomaly time series estimated using the Levinson-Durbin recursion on the Yule-Walker equations, with confidence bands reflecting the change over time of the parameters, estimated by fitting different parts of the time series separately, as well as the corresponding relative change in %.

To check for a possible change in the parameter values over time, we fitted the AR(1) model to different distinct parts of the same temperature anomaly time series. The resulting parameter values varied within the following intervals: $\gamma \in [0.705, 0.710]$, well within the 95%-confidence interval for the fit over the whole time series length, and $\sigma_\epsilon^2 \in [7.63K^2, 7.89K^2]$. This suggests relative changes of the parameters over time of 0.4% for γ and 1.7% for σ_ϵ^2. The time dependence is therefore negligible for the AR(1)-model.

The AR(1)-model chosen here gives us a good feeling for the decorrelation time in the time series, defined as follows[38]:

$$\tau_D = \frac{1+\gamma}{1-\gamma}. \tag{3.5}$$

For the coefficient $\gamma = 0.707$ we obtain a decorrelation time of $\tau_D = 5.8$ or slightly less than six days. This means that there is hardly any information about the initial condition of the anomaly left in the model on the time scales relevant to our first passage time problem.

$$\Delta T_n = \sum_{i=1}^{8} \gamma_i \Delta T_{n-i} + \epsilon_n, \tag{3.6}$$

For the statistically preferred AR(8)-process as given by Eq. (3.6), the estimated parameters γ_i are given in Table 3.4. The sample variance of the noise is $\sigma_\epsilon^2 = 7.71K^2$, slightly smaller than for the AR(1)-process.

We again checked for a possible time dependence of the parameters. The error bars in Table 3.4 reflect the resulting change when fitting different parts of the time series. The sample variance of the noise varied as $\sigma_\epsilon^2 = (7.71 \pm 0.13)K^2$, corresponding again to a relative change of 1.7% over time. As can be seen, almost all higher order parameters are very small and their change over time is comparatively large, a confirmation of the assumption that the reduction to an AR(1)-process would make more sense in this context.

Since the objective determination yielded an optimal order of $p = 8$ but an AR(1)-process also looks very promising, we should test the model fit carefully for both cases to get a feeling for the adequacy of each. Box and Jenkins[50] propose a thorough testing method for the case of autoregressive models that we will adopt in the following.

Looking at the distribution of the residuals, we expect a normal distribution if the models fit perfectly. Fig. 3.16 shows the normal probability plots (see Sec. 2.2.4) for the residuals after fitting with an AR(1) and an AR(8) model. The probability plots show significant deviations from the straight line expected for a normal distribution for both model orders. While revealing some model inadequacies, the distributions therefore do not serve to show preference of one model order over the other.

Apart from the distribution of the residuals ϵ_n itself, their autocorrelations also offer significant insight into the goodness of fit of the AR model. In fact, if the model were perfect and

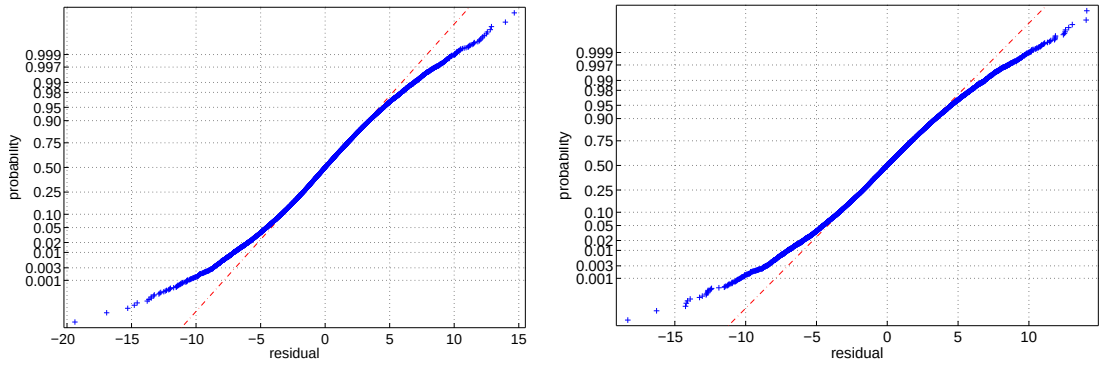


Figure 3.16: Normal probability plot of the residuals after fitting the Potsdam morning temperature anomaly time series with an AR(1) model (left) and an AR(8) model (right).

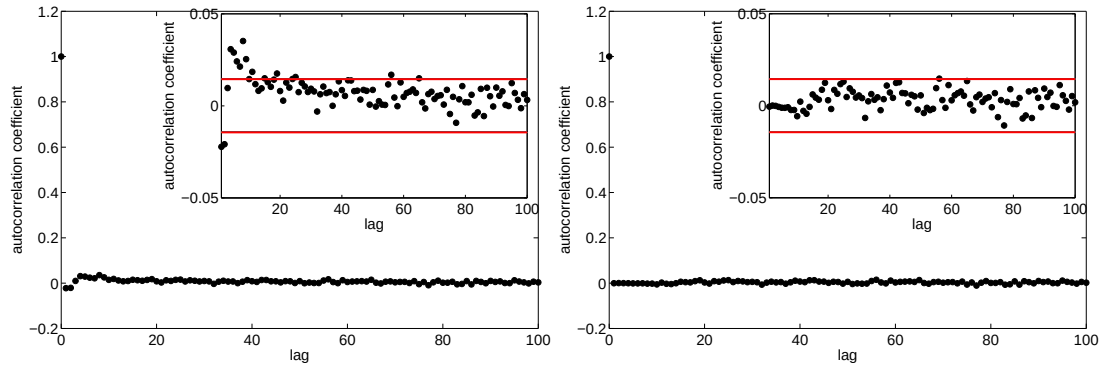


Figure 3.17: Autocorrelation coefficients of the residuals after fitting the Potsdam morning temperature anomaly time series with an AR(1) model (left) and with an AR(8) model (right). The insets show the same data as the full figures but without lag 0 and with added confidence bands for 3σ (red lines).

contained the correct parameters instead of their estimates, we would expect the residuals to be uncorrelated. The estimated autocorrelations of the residuals should therefore be distributed normally with mean 0 and standard deviation $1/n$, where n denotes the total number of sample points in the data[50].

Fig. 3.17 shows the autocorrelations of the residuals both after a fit with an AR(1) model and with an AR(8) model. As can be seen, the autocorrelation coefficients for small lags are encouragingly small for both models, but are indeed smaller for an AR(8) model[10]. However, the addition of the 3σ confidence bands in the insets shows that while the AR(8)-process describes the anomalies adequately and indeed all but one of the points lie within the confidence bands, there are significant deviations for an AR(1)-process. They are all the more remarkable as they occur for small lags, where constant confidence bands are larger than the actual confidence bands should be so that they underestimate the significance of departures from normality there[50]. From these plots we can infer that the AR(8)-model indeed seems to be significantly better than the AR(1)-model for the temperature anomaly time series.

In order to make the importance of such deviations from the expectation of uncorrelated

[10]The deviations from zero for lags up to the order of the fitted model are due to a seasonality in the model fit.

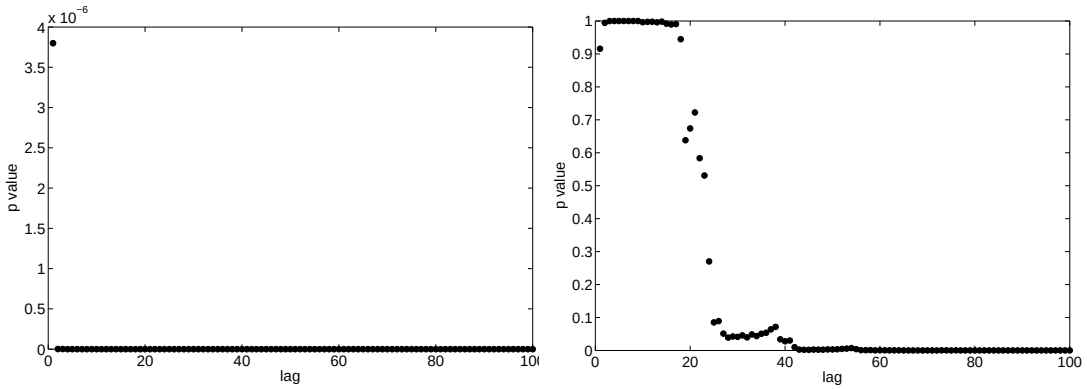


Figure 3.18: P value of the Ljung-Box statistic applied to the residuals of the Potsdam morning temperature anomaly time series fit with an AR(1) process (left panel) and an AR(8) process (right panel) for different cutoff values of the lag in the autocorrelation function.

residuals more immediately accessible, Ljung and Box proposed a portmanteau statistic as a summary measure to test for lack of fit[56]. It checks whether the deviations from zero of the first few non-zero lags of the autocorrelation function of the residuals are compatible with the assumption of purely statistical fluctuations (for more details on the procedure see Sec. 2.2.2). Fig. 3.18 shows the resulting p value, i.e. the probability that the observed residuals are indeed uncorrelated, plotted against the upper cutoff value of the lag used for the evaluation in the test. As can be seen, randomness of the residuals is rejected clearly for every lag in the case of the AR(1) model. For the AR(8) model, the fit is almost perfect up to lag 20, but then the goodness of fit starts to decay rapidly, until randomness of the residuals is rejected consistently from lag 45 onwards. This means that the autocorrelation structure of the temperature anomalies is not well represented by either model order, even though the AR(8) model is markedly better at small lags.

The possible problem of some periodicity not accounted for by the chosen simplistic model is not necessarily visible in the autocorrelation function, since it might be diluted over several different lags. A slightly better measure is provided by a periodogram (see also Sec. 2.1.3), where any such periodicities are amplified by the corresponding sine or cosine wave of the same frequency. If the model were perfect and the residuals indeed white noise, the power spectral density should be uniform over all possible frequencies.

Fig. 3.19 shows the estimates both for the residuals after fitting with an AR(1) process (left panel) and with an AR(8) process (right panel). As can be seen, both estimates are rather noisy but seem to show a reasonably uniform distribution, although the AR(8) process has less deviations for smaller frequencies.

A sometimes better way of checking this, since it significantly reduces the fluctuations still present in the periodogram estimates, is to switch to the cumulative distribution, i.e. the power spectrum itself[50]. In Fig. 3.19 we have chosen the normalised version as stated in Eq. (3.7), where $C(f_j)$ is the power up to frequency f_j, $S(f_i)$ is the estimated power spectral density at frequency f_i, f_S is the sampling frequency ($f_S = 365.2425\ \text{year}^{-1}$), L is the segment length used in the periodogram estimation and $\hat{\sigma}^2$ is the sample variance of the residuals.

$$C(f_j) = \frac{f_S}{L\hat{\sigma}^2} \sum_{i=1}^{j} S(f_i) \tag{3.7}$$

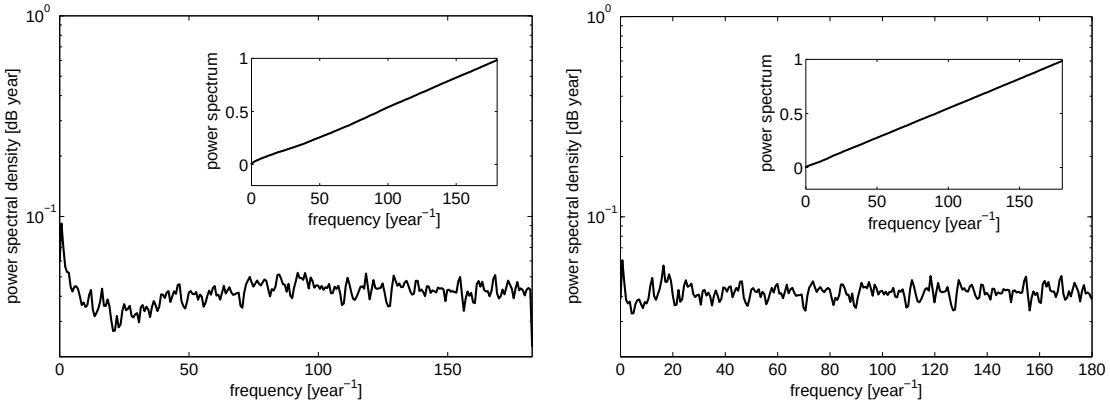


Figure 3.19: Power spectral density estimates of the residuals of the Potsdam morning temperature time series fit with an AR(1) process (left) and an AR(8) process (right) on a semilog scale, obtained using Welch's periodogram method with around 170 segments of length 512 data points with a 50% overlap of consecutive segments. The insets show the same data, but summed to give the cumulative and then normalised as given by Eq. (3.7).

For white noise, the normalised cumulative periodogram will be a straight line joining the points $(0, 0)$ and $(\frac{f_S}{2}, 1)$. As can be seen in the insets of Fig. 3.19, the estimates for both the AR(1) model and the AR(8) model are very close to the expected straight line, especially considering that the model parameters were only estimated from the data, leading to additional errors. This means that the models seem to represent the oscillations of the underlying dynamics quite well, even though, again, there is a slight improvement for the AR(8)-process over the simpler AR(1)-process.

Since most of the goodness of fit tests do not penalize the AR(1)-model overmuch when compared to the higher order version and since the estimated higher order parameters are vanishingly small but with large errors, it makes sense to stay with the simpler model. However, the model checking procedure has revealed that the autocorrelation structure of the temperature anomalies is not well represented by an autoregressive model. To further see how large a drawback this is, we will check how well the sample autocorrelation function of the temperature anomalies is fitted by the theoretical autocorrelation function of the AR(1)-process with the estimated parameters. As can be seen in Fig. 3.20, the anomaly autocorrelation decays more slowly than for the AR(1)-process and there seems to be some correlation left even at the largest lag shown in the figure.

While the difference in quality between an AR(1) and the statistically preferred AR(8) process is small so that the AR(1) process as the simpler model can be chosen from this class of model processes, autoregressive processes in general do not capture the autocorrelation structure of the temperature anomalies well. Indeed, the power spectrum of the anomalies is weakly red and therefore shows a long memory of the underlying process contrary to the model, as previously found by Caballero et al.[134]. One other problem is that the anomalies have a fatter tail towards the negative side (see Fig. 3.5) but that the AR(1)-process can only generate a time series that is distributed normally. Therefore, while the AR(1)-process might be taken as a first approximation, there are some model deficiencies that need to be kept in mind.

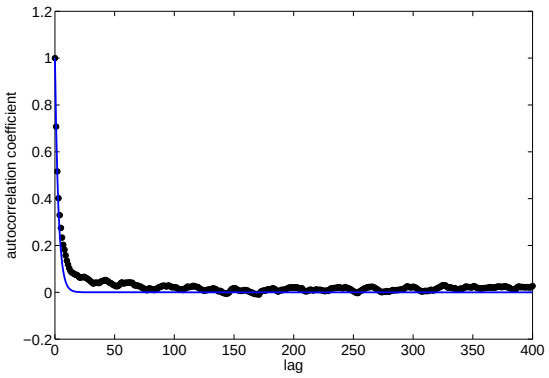


Figure 3.20: Sample autocorrelation coefficient of the Potsdam temperature anomaly time series (black dots) and the autocorrelation function of the corresponding AR(1) model process (blue line) for the first 400 lags.

3.5 First passage time distributions

3.5.1 First passage time to frost

In the next chapter, we will analyse the probability distributions $P(t_{\text{frost}}|t_0, \Delta T)$ of the first passage time to freezing temperatures t_{frost} conditioned on the initial temperature anomaly ΔT recorded on the initial date t_0. Before conditioning on ΔT, however, we will first look at the corresponding probability distributions $P(t_{\text{frost}}|t_0)$, which we will call *unconditioned* in the following, and how they change throughout the calendar year.

Since our time series only comprises 118 years (from 1893 to 2010), such distributions have to be reconstructed from at most 118 values for each t_0. Moreover, to avoid cut-off artefacts from the ending of the time series, we should only take initial anomalies ΔT from the years 1893 to 2009 so first passage times t_{frost} of up to 365 days are always possible, reducing the available data to 117 values for each initial date.

To look at the first passage time distributions, we will also limit our analysis to initial temperatures $T > 0$ °C to consider only true threshold crossings.[11] However, this additional exclusion of data points results in even less data especially in winter and therefore also in very poor statistics for a distribution estimate.

To avoid overly large statistical fluctuations, we will consider the anomaly time series to be pseudo-stationary at least over shorter time intervals as already discussed in Sec. 3.3.1. We will therefore use all anomalies recorded within time windows of a predefined width around the desired initial date t_0 as starting points, resulting in a repeated shifting of the anomaly time series with respect to the corresponding climatology. In detail, if we are interested in the first passage time from date t_0, $t_0 \in \{1, ..., 366\}$, to temperatures $T \leq 0$ °C using a window width of w_t days around t_0, then we consider all anomalies ΔT_j with $|t_0 - j| \leq \lfloor w_t/2 \rfloor$ as possible starting values[12]. Then, for each such j which also satisfies $\Delta T_j + \widetilde{T_{t_0}} > 0$ °C, the first passage time to frost t_{frost} is defined as the smallest number of days k, $k \in \mathbb{N}$, for which $\Delta T_{j+k} + \widetilde{T_{t_0+k}} \leq 0$ °C.

[11] Including initial temperatures already below the threshold would mean considering also persistence effects. These could, however, be reformulated as first passage time problems to an upper temperature threshold of low absolute value.

[12] Of course, $j = 366$ (or $j = 367$ in the case of a leap year) is identified with $j = 1$, leading to windows that can stretch across the boundary between two consecutive calendar years. Also, for simplicity, we will only consider odd values of w_t, so the desired initial date will be exactly at the center of each time window. The special case of $w_t = 0$ which recovers the unwindowed exact calculation with poor statistics will also be included.

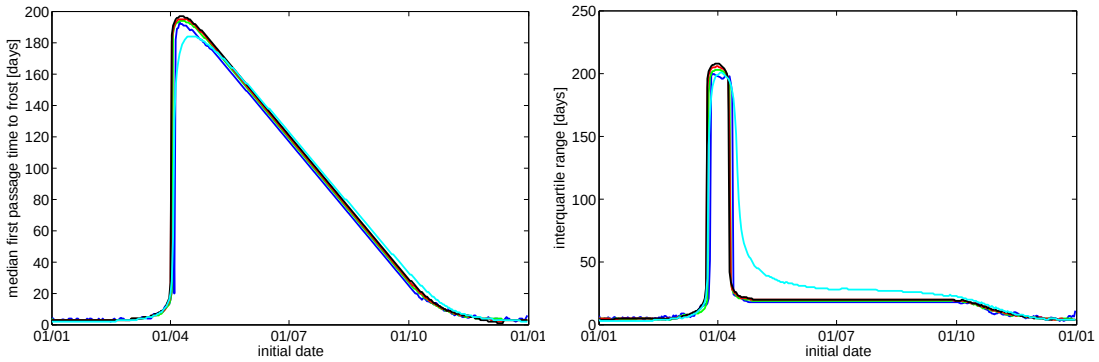


Figure 3.21: Median (left panel) and interquartile range (right panel) of the probability distribution $P(t_{\text{frost}}|t_0)$. The conditioning on t_0 was done using only anomalies recorded precisely on the date in question (dark blue), within windows of width 31 days (green), 61 days (red) and 91 days (black) around t_0, as well as using all recorded anomalies (cyan).

Fig. 3.21 shows the change of the median and the interquartile range (IQR) of the resulting first passage time distributions across the calendar year for different values of the window width w_t[13]. In the following, such summary measures for the location will always give the number of days until first frost, unless directly specified otherwise. Comparing the results for different window widths, it can be seen that both for the median and the IQR the curves for $w_t = 31$ days, $w_t = 61$ days and $w_t = 91$ days agree almost perfectly. However, a marked difference between the window widths occurs around the maximum of the median of the first passage time distributions, i.e. in early April. Here, the windows of one to three months in width seem to overestimate the median slightly, but only by about a week on time scales of half a year in total. However, simply using all anomalies independently of the date on which they were recorded leads to a rather marked underestimation. This emphasizes the lack of stationarity in the anomaly time series across the calendar year.

As with the location of the distribution, the spread also shows that using all anomalies irrespective of their recording date makes some serious changes to the first passage time distribution. In fact, the spread is then seriously overestimated in summer and autumn and underestimated in winter. This is consistent with the finding that the anomaly variance is much larger in winter than in summer (see Fig. 3.10), a fact that is disregarded completely when considering the time series as stationary.

Looking further at the IQR, one can also see that there is a hint of bimodality in the peak in early April that disappears when using windowing. However, it is not clear whether this is an artefact of this particular realisation or a serious effect that is lost when the anomaly time series is considered to be stationary[14].

Window widths of up to 3 months around the initial date therefore lead to a very good

[13] This choice was made to reduce the influence of outliers in the data. Indeed, looking at a single time series, i.e. a single realisation of the underlying dynamical process, means that unusually long first passage times affect many consecutive starting dates, since especially in winter the climatology is almost constant over longer time periods. Windowing around the initial date enhances the influence of outliers even further. This way of generating first passage times also leads to the relative smoothness of the curves despite the small number of data points used for the calculation.

[14] Using the first passage times gathered from the initial anomalies with window width $w_t = 91$ days and randomly extracting a similarly reduced amount of data points to calculate the IQR with the same lack of statistics only slightly increases fluctuations along the curve but fails to reproduce this bimodality effect or any other fluctuations of similar magnitude. It should therefore not be due to the lack of statistics.

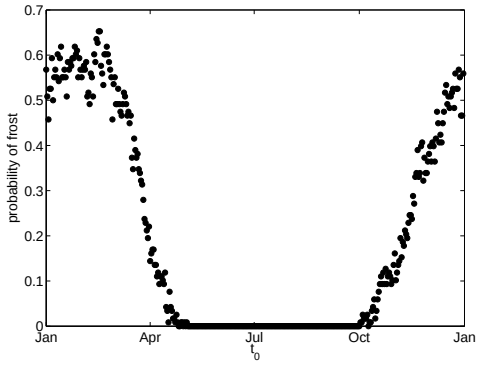


Figure 3.22: Probability of observing freezing morning temperatures in Potsdam depending on the date.

reproduction of the median and the IQR of $P(t_{\text{frost}}|t_0)$. A window width of 3 months therefore seems a reasonable compromise between introducing a bias and reducing statistical fluctuations.

Looking once more at Fig. 3.21, one can see that there are four main characteristic regimes for $P(t_{\text{frost}}|t_0)$. First, from mid November to the end of February, i.e. throughout winter, there is a region where the distribution is very narrow with constant spread of around a week and also constant median very close to 0 days, i.e. almost certain frost in the next few days. This is confirmed by Fig. 3.22 that shows that in winter, the probability of a first passage time to frost of one day is roughly between a half and two thirds.

The second and shorter region is in March, when the distribution acquires a large spread. There, as soon as there is a large probability of first frost only after the summer, the median and the IQR jump abruptly to times of around half a year. The third region is the plateau of constant IQR of around a month and linearly decreasing median that extends from the beginning of May to the beginning of October. This indicates a vanishing influence of t_0 on $P(t_{\text{frost}}|t_0)$ during summer with a first day of frost around November 1$^{\text{st}}$. The fourth characteristic region is the slow decrease of the IQR from the beginning of October to mid November, where the decrease of the median is slowing down.

Taking initial dates representative of each of these four characteristic regions, we will now look at kernel density estimates (see Sec. 2.1.2) of the full first passage time distributions for these distinctive cases[15]. As can be seen in Fig. 3.23, for both cases with low median and low interquartile range, i.e. for the initial dates February 14$^{\text{th}}$ and November 1$^{\text{st}}$, $P(t_{\text{frost}}|t_0)$ has a peak almost immediately after the initial date and then decays rapidly[16]. This was to be expected when starting during a time when temperatures in general are rather low. Starting far away from cold periods, namely on July 1$^{\text{st}}$, $P(t_{\text{frost}}|t_0)$ resembles a normal distribution with a peak around November 1$^{\text{st}}$ which is centered on the climatological date of first frost and, judging from the previous figures, does not change in either location[17] or spread with moderate changes of t_0. Finally, for t_0 in spring around the time of the jump both in median and interquartile range, one can see that $P(t_{\text{frost}}|t_0)$ becomes bimodal with a clear separation between next frost

[15] The theoretical support of the first passage time distributions is, of course, neither continuous nor unbounded. However, if the support is restricted to positive values and the data are log-transformed and back-transformed after the kernel density estimation, artificial oscillations will be introduced for small values. We therefore dispensed with this more complicated scheme.

[16] Note that there is a significant difference in spread between these two examples.

[17] A median of the first passage time linearly decaying with increasing initial date corresponds to a fixed date for the peak.

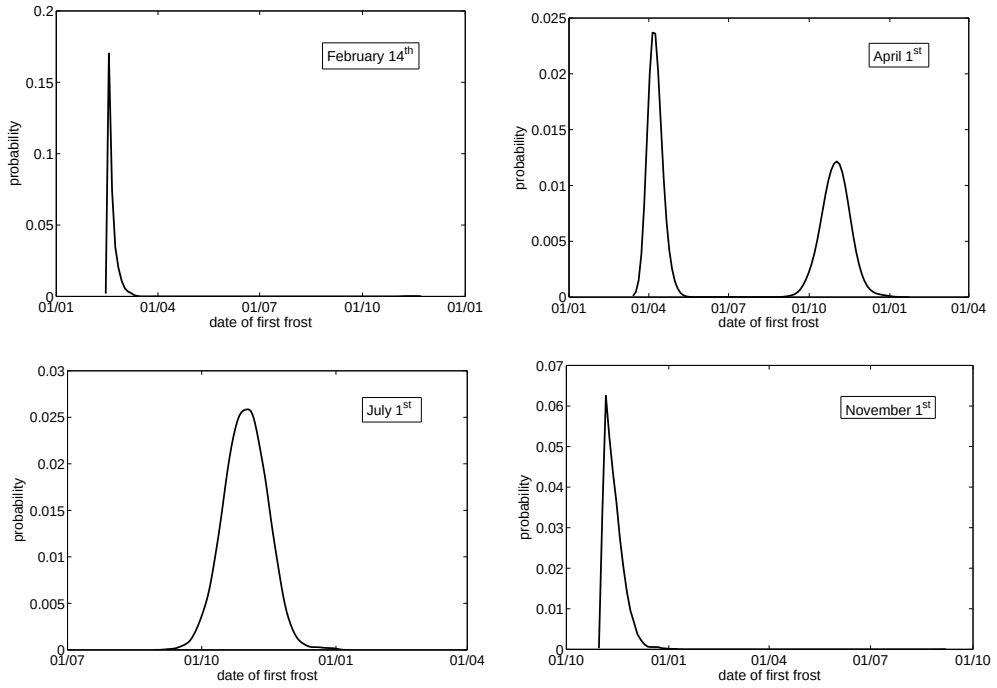


Figure 3.23: Kernel density estimation of $P(t_{\text{frost}}|t_0)$ for four representative values of t_0.

before and after summer[18].

There are therefore three distinct regimes for $P(t_{\text{frost}}|t_0)$: For a date t_0 when frost is very likely to occur, the distribution of the first passage time is close to an exponential decay[19] with a peak in the very near future. For t_0 in spring, there is a clear bimodality with frost on some occasions not observed until the next winter, while on other occasions it still occurs rather soon after the initial date. Finally, for t_0 during a time when no frost is to be expected, we observe an almost normal distribution centered around a fixed date and with a constant spread independent of the exact value of t_0.

3.5.2 Comparison to an AR(1) model process

One problem of the previous analysis, as already stated, is the lack of available data if only a precise initial date is considered or if most of the data do not satisfy the condition of $T > 0$ °C. We therefore estimated parameters for an autoregressive model process in Sec. 3.4 to be able to address this problem. Using this AR(1) model process to generate a time series of short-range correlated anomalies of an equal number of data points as the temperature measurements, we repeat our previous analysis[20].

[18] The more Gaussian appearance of the earlier peak when compared to the distributions for initial dates in winter is due to the choice of a slightly larger kernel width in the estimate to smooth out statistical fluctuations in the second peak. That the estimated distribution begins before t_0 is due to neglecting to consider the bounded support of first passage times.

[19] It does, however, not truly match such a distribution.

[20] Using this model process, the future objective is of course to generate large ensembles of such anomaly time series for every possible initial date. However, since we have only a single long realisation of the temperature data, treating the AR(1) process exactly the same makes for a better comparison at this stage, since it then also contains all smoothing and windowing artefacts, as well as statistical fluctuations present in the measured data. In this way, any significant differences between model and data results are therefore due solely to the inadequacy of the model process.

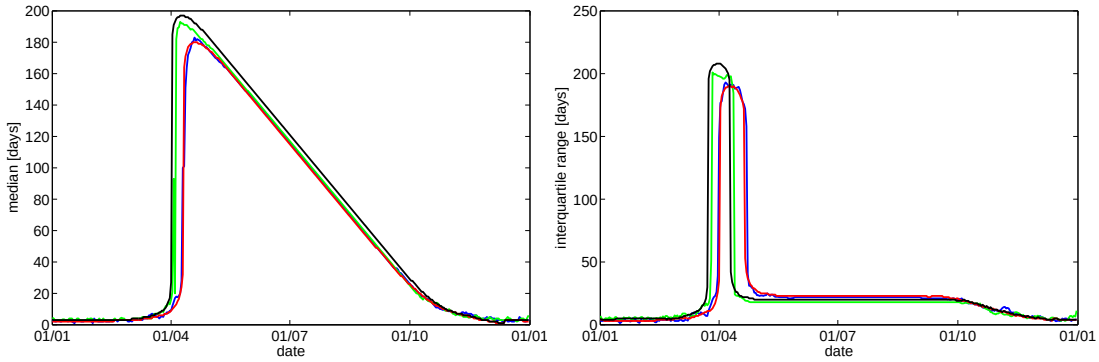


Figure 3.24: Median (left panel) and interquartile range (right panel) of $P(t_{\text{frost}}|t_0)$ generated from the temperature anomalies without windowing (green) and with a window width of 91 days (black), as well as from the AR(1) model process without windowing (blue) and with a window width of 91 days (red).

The most visible difference between model and data, as expected, is the evidence of stationarity of the AR(1) anomalies: As can be seen in Fig. 3.24, the median and IQR of $P_{\text{AR(1)}}(t_{\text{frost}}|t_0)$ do not change when using all anomalies irrespective of the date on which they were recorded, instead of only those in a window of width $w_t = 91$days around t_0.

However, when directly comparing $P_{\text{data}}(t_{\text{frost}}|t_0)$ and $P_{\text{AR(1)}}(t_{\text{frost}}|t_0)$, more differences emerge. It is clearly visible in Fig. 3.24 that while $P_{\text{AR(1)}}(t_{\text{frost}}|t_0)$ shows the same seasonal changes of the location and spread as $P_{\text{data}}(t_{\text{frost}}|t_0)$, the switching to the bimodal regime and the associated jump in location occur at a slightly later time.[21]

This can be understood by looking at the probability distribution of both anomalies: The anomalies obtained from the temperature measurements have a longer negative tail than the normally distributed anomalies generated by the AR(1)-process. This tail does not contribute to $P(t_{\text{frost}}|t_0)$ for t_0 in March since the very low anomalies then do not satisfy the condition of positive initial temperatures. However, the shorter positive anomaly tail in the data has a correspondingly enlarged weight compared to a normal distribution. This leads to a slight prevalence of higher temperatures in the data and thus to an earlier onset of first passage times to dates beyond the summer. This earlier onset is then directly responsible for the slightly longer passage times in the peak and the correspondingly broader distribution in the data.

Looking again directly at kernel density estimates of the full distributions for characteristic initial dates, these results are substantiated by Fig. 3.25. While the AR(1) model seems to reproduce the first passage time distributions almost perfectly for t_0 in winter and comes quite close for t_0 in summer where it leads to only slightly earlier dates than the temperature data, there is quite a large difference in the bimodal case. Indeed, the relative weight of the two peaks is not reproduced well at all by the model, which underestimates the number of cases in which frost will not occur until after summer. This is again due to the enhanced weight of large positive anomalies from the temperature data when compared to a normal distribution.

The AR(1) model is therefore most appropriate for the estimation of $P(t_{\text{frost}}|t_0)$ for t_0 in winter or summer, but much less so when considering the rather more interesting case of t_0 in spring.

[21] The same conclusions apply if the model process is changed to the higher order AR(8)-process (see Sec. 3.4). The curves from the AR(1) and the AR(8) process are almost indistinguishable.

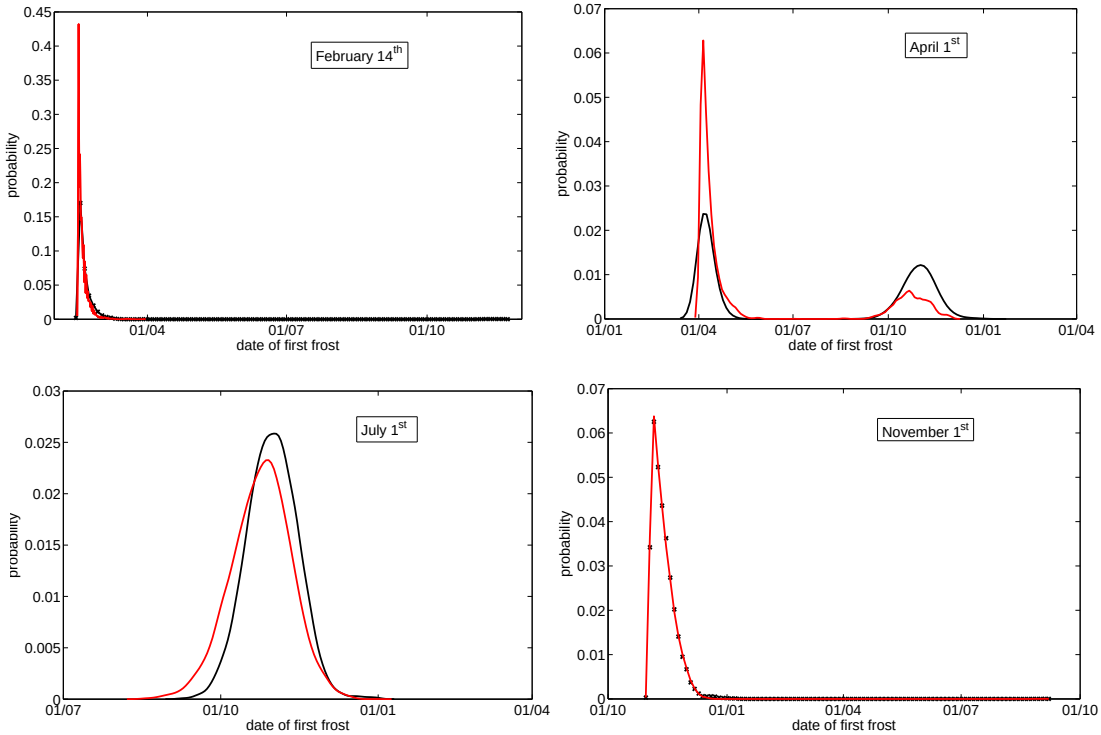


Figure 3.25: Kernel density estimation of $P_{\text{data}}(t_{\text{frost}}|t_0)$ (black) and $P_{\text{AR(1)}}(t_{\text{frost}}|t_0)$ (red) for four representative initial dates.

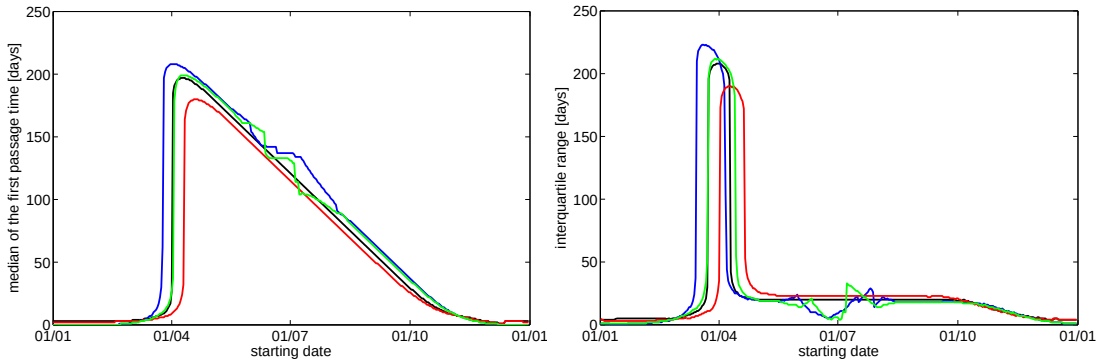


Figure 3.26: Median (left panel) and interquartile range (right panel) of $P_{\text{data}}(t_{\text{frost}}|t_0)$ (black) and $P_{\text{AR(1)}}(t_{\text{frost}}|t_0)$ (red) for a windowing width of 91 days, as well as the corresponding statistics for $P_{\text{data}}(t_{T>13.8^o\text{C}} + 188|t_0)$ (blue) and $P_{\text{AR(1)}}(t_{T>13.8^o\text{C}} + 188|t_0)$ (green).

3.5.3 First passage time to 13.8 °C

Until now only the first passage times to frost were considered, because of their special relevance with regards to applications, for instance in agriculture and transportation. However, it remains to be seen whether this is indeed also a special threshold with regards to the time series themselves. To analyse this, we will now look at the first passage times to crossing the threshold $T = 13.8$ °C from below. This threshold was chosen because it represents the mirror quantile of $T = 0$ °C with regards to the median of the temperature distribution.

Fig. 3.26 shows the median and interquartile range of $P_{\text{data}}(t_{\text{frost}}|t_0)$ and $P_{\text{AR(1)}}(t_{\text{frost}}|t_0)$, as well as those of $P_{\text{data}}(t_{T>13.8^o\text{C}}|t_0)$ and $P_{\text{AR(1)}}(t_{T>13.8^o\text{C}}|t_0)$, shifted by exactly half a year to look for symmetries. It can be seen that while the general shape of the curves is indeed very similar, there are some differences especially for the bimodal cases, that occur earlier for the upper threshold and lead to larger median and IQR. The shift also shows that there are large fluctuations present for the upper threshold for t_0 in summer that do not occur for the lower threshold. These are in fact a windowing artefact that disappears when only taking initial anomalies recorded on t_0 exactly into account (not shown here).

Therefore, while frost is a very significant threshold value in applications, its influence on the first passage time distributions to a threshold crossing from above is not significantly different from the mirror problem of a threshold crossing of $T = 13.8$ °C from below. It is therefore not a special threshold value in the context of our analysis.

3.6 Conclusion

In order to show the predictive skill of statistical forecasts on seasonal time scales in the extratropics, we chose to analyse forecasts of the first passage time to frost in this thesis, as more time-resolved forecasts of surface air temperatures are needed for many applications and temperatures have longer intrinsic correlations and less measurement errors than for instance precipitation.

Before starting on the predictions themselves in the next chapters, we first analysed the measurement data. We used the daily morning temperatures from the Potsdam station, a homogeneous time series containing 118 years of measurements without missing data points and of good data quality with an accuracy of ±0.2K.

As any prediction effort will already score quite well by just predicting the seasonal cycle of

the mean temperatures for each calendar day t_0, we proceeded to model it using a global mean and a sinusoidal model of the two main frequencies. After subtracting this climatology, we only retain the temperature anomalies ΔT, i.e. the fluctuations around the mean seasonal cycle.

Estimating any conditional first passage time distribution $P(t_{\text{frost}}|\Delta t, t_0)$ from the data which only contains 118 recorded values for each t_0 poses the problem of very poor statistics. Taking advantage of the assumed pseudo-stationarity of the anomaly time series, we wanted to relax the conditioning on t_0 and use all anomalies irrespective of the day they were recorded on. Then we could simply add the correct value of the mean seasonal cycle for the desired dates to recover a temperature time series.

However, some violations of the stationarity assumption were found in the anomaly time series. The mean anomaly increases with time due to a temperature trend that was not removed with the seasonal cycle. This directly influences the relative weight of different years within the time series when conditioning on very large absolute values of ΔT. The trend is also reflected in the autocorrelation function of the anomalies where it enters as the most obvious long-range correlation. Nevertheless, as can be seen both in the actual trend magnitudes as given by Table 3.3 and in the good climatological model fit for both the first and the last years in the time series as shown in Fig. A.1, the trend magnitude is very small. Any trend estimation also contains large errors, so that for the moment, we will not detrend the data.

Both the variance and the skewness of the anomaly distribution also show a dependence on the calendar month in which the anomalies were recorded, influencing the relative weight of different months in the analysis due to the use of time windowing for ΔT.

The violations of the stationarity assumption pose a bias-variance tradeoff problem between obtaining adequate statistics for the estimation of conditional first passage time distributions and avoiding the introduction of artefacts. As a compromise, we therefore chose to use a window of three months' width centered around the initial calendar day for our choice of initial anomalies.

Another method to enhance the statistics for the conditional first passage time distributions would be to model the anomalies using an autoregressive process. The AR(1)-process captures the main characteristics of the temperature anomalies quite well. However, it evidently does not capture the nonstationarities or the weakly red power spectrum, nor does it reproduce the deviations from a normal distribution observed in the temperature anomalies.

Before starting the predictability analysis by looking at $P(t_{\text{frost}}|\Delta T, t_0)$, we first studied the general shape of $P(t_{\text{frost}}|t_0)$. We found three distinct cases: An exponentially decaying distribution if the initial day falls within the coldest months of the year, a bimodal distribution for initial days in spring and a normal distribution with a mean first time of frost at the end of October whose shape also does not depend on the exact initial date for a large period of time in summer.

We were able to confirm that our choice of a 3 month window around the initial day for the anomalies adequately reproduced the estimated distribution $P(t_{\text{frost}}|t_0)$ obtained without windowing.

Studying $P_{\text{AR(1)}}(t_{\text{frost}}|t_0)$, we found that the AR(1)-process was very adequate for initial days in summer and winter but much less so for spring.

Finally also considering the mirror problem of a threshold crossing of $T = 13.8$ °C from below, we found very similar unconditioned first passage time distributions to the case of first frost, shifted by half a year. First frost is therefore not a special threshold with regard to the time series, making our findings on time series predictability more general than if it were.

After thus concluding the preliminary analysis of the properties of the temperature data as well as the possibilities of enhancing the statistics in the following, we can now start analysing the predictability of first passage times to frost.

4 First passage time prediction in temperature data

4.1 Introduction

In the previous chapter, we have completed the preliminary analysis of the temperature data. By subtracting the significant seasonal cycle, we obtained the more interesting fluctuations around the long-time mean. This led to a temperature anomaly time series. Now we want to turn to the goal of our thesis, namely the prediction of the first passage time to frost.

In a first step, however, we will analyse the potential of predictability contained in the data. If the first passage time probability distribution does not depend on the initial conditions observed on the day we issue a forecast, there is no possibility of a meaningful prediction. In mathematical terms, this means that the conditional probability distributions $P(t_{\text{frost}}|t_0, \Delta T)$ need to show a non-trivial dependence not only on the initial date t_0 but also on the initial anomaly ΔT. As seen in the previous chapter, $P(t_{\text{frost}}|t_0)$ depends strongly on t_0. In order to predict more than the seasonal cycle, however, we will now need to see also the influence on the initial anomalies.

As stated before, our time series only contains 118 recorded values for each calendar day. Even using a temporal window around t_0, the number of data points in winter that fulfill the condition $T > 0$ ℃ for any given calendar day is not large enough to further partition it into too many initial anomaly categories. We will therefore need to coarse-grain also this second initial condition significantly.

This leads to many parameters for estimating the probability distributions $P(t_{\text{frost}}|t_0, \Delta T)$: Not only the two different initial conditions can change, but also the extent of their coarse-graining. Additionally, the use of an AR(1) model process to simulate temperature anomalies as discussed before in Sec. 3.4 offers another prediction approach.

From this wealth of possibilities, we want to find the parameters for which the dependence of $P(t_{\text{frost}}|t_0, \Delta T)$ on ΔT is largest as this represents the true predictability potential. In order to do this, we will first pick some exemplary parameter combinations and look at the resulting distribution estimates to get a feel for the different dependences. In a next step, we will conduct a more systematic analysis of the parameter dependences before concluding the predictability evaluation with an analysis of two specific examples derived from possible applications of such first passage time predictions.

In the second part of this chapter, we will then proceed to issue systematic first passage time predictions for different application needs. Using appropriate forecast scoring schemes, we will verify the quality of the predictions. The analysis of its change with different prediction parameters provides the means to identify the best forecast schemes and the situations in which such a data-based prediction provides significantly more information than simply long-term averages.

4.2 Potential for predictability

Before proceeding to these full predictions for the first passage time problem, we will evaluate in this section the potential for predictability contained in our time series. We will focus especially on its dependence on the precise initial conditions. In the previous chapter, we have seen the

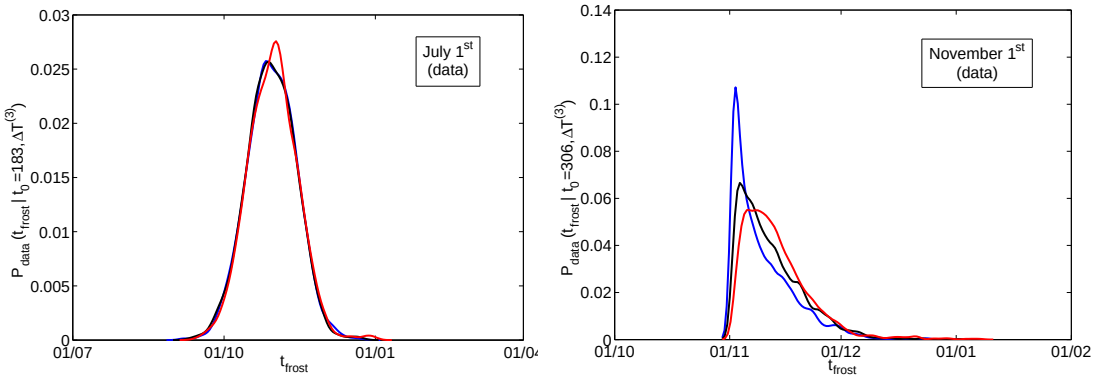


Figure 4.1: First passage time distribution estimates for initial dates within a 3 month window around July 1st (left panel) and November 1st (right panel) conditioned on initial anomaly bands, with colder than average anomalies depicted in blue, average in black and warmer than average anomalies in red. Normal kernels of width 3 days (July) and 1 day (November) were used for the density estimation.

seasonality of the autocorrelation (as shown in Sec. 3.3.2). The AR(1) model process might be used to enhance the statistics of the data. Its suitability, however, also varies across the calendar year. Therefore, especially the initial date might have a large influence on the predictive power.

4.2.1 Change of first passage time distributions under conditioning

The first indication for potential predictability would be a distinct change in the first passage time distribution if one conditions it on different initial temperature anomalies. If the first time of frost depends on the precise conditions observed on the initial date, i.e. if $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T)$ shows a nontrivial dependence on ΔT, then some predictive power is contained in the time series.

Estimating the full conditional first passage time distribution $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T)$ from the data presents the challenge of retaining an adequate number of data points for the statistics. We will therefore coarse-grain the condition on the initial date by taking t_0 to mean $t \in [t_0 - 45, t_0 + 45]$, i.e. by considering a 3 month window around it. We will also loosen the condition on the initial temperature anomalies ΔT by considering only a separation into terciles, i.e. we will condition the distribution on below average, average and above average initial anomalies. This will be denoted in the following by $\Delta T^{(3)}$.[1]

Fig. 4.1 shows the kernel density estimates of the resulting conditional probability distributions for initial dates of July 1st and November 1st, i.e. $t_0 = 183$ and $t_0 = 306$, as examples[2]. It can clearly be seen that starting in summer, the distributions resulting from different initial anomaly terciles are indistinguishable, i.e. $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3)})$ does not depend on $\Delta T^{(3)}$ for $t_0 = 183$. This leads to the conclusion that there is no information about the date of first frost contained in the time series this much in advance.

An initial date much closer to freezing temperatures on the other hand, namely November 1st which has a climatological temperature $\widetilde{T} = 4.3$ °C, results in larger differences between the corresponding conditional first passage time distributions, as can be seen in the right panel of Fig. 4.1.

[1] This specific choice of coarse-graining for the initial anomalies is made here for low statistical fluctuations. Other coarse-graining schemes as well as their influence on the predictability potential will be discussed later.

[2] Conditional probability distributions for other values of t_0 and also for more extreme initial anomaly bands, as well as corresponding distributions gathered from the AR(1) process, are shown in Appendix B.2.

Looking at the probability distributions for other initial dates t_0, one can see that they do indeed depend on the initial anomaly tercile $\Delta T^{(3)}$ for most of the initial dates, except when starting in the middle of summer. However, these differences are generally rather slight and further analysis is needed to see whether they are even statistically significant.

Separating the initial conditions only into three categories is a very coarse-grained approach. Weigel et al. suggested that looking at more extreme initial conditions instead of exact terciles might significantly improve predictability[26]. Indeed, considering more extreme anomaly bands of only 500 points each instead of the approximately 3500 data points in each tercile (depending on t_0) does seem to enhance the differences. However, as shown in Appendix B.2, they are still slight enough that they would not necessarily be statistically significant considering the reduced number of points to base the estimation on.

For the corresponding first passage time distribution estimates gathered from the AR(1) model process, the results look very similar[3]. There are some notable differences though. For initial dates very early in the year, the measured data contains some instances when first frost does not occur until after summer. These do not exist in the model process. There, on the other hand, starting in late spring, there are still some instances of frost in the immediate future that are not observed in the data. In both cases, the distributions obtained from one time series are slightly bimodal, while those from the other show only one peak.

Moreover, the figures in Appendix B.2 show that while for an initial date in July the data-generated curves are nearly indistinguishable (as stated before), a slight shift in the location of the peak might be contained in the AR(1)-generated data with the more extreme anomaly bands. If this could be confirmed to be significant, it would be a very interesting finding: The AR(1) process involved has a decorrelation time of around 5 to 6 days (see Sec. 3.4). That some influence of the initial condition could be observed after around 120 days would be rather unexpected. A Kolmogorov-Smirnov test of the null hypothesis that the 500 data points with lowest initial anomaly and the 500 data points with highest initial anomaly were drawn from the same underlying distribution returns a p value of $p = 0.19$. This probability is not sufficiently small to reject the null hypothesis.

Even in the other cases where the differences between the data-derived first passage time distributions and those generated from an AR(1) model appear to be very small, this is difficult to evaluate without a direct comparison. Indeed, Fig. 4.2 shows two initial dates for which there is some discrepancy between the two distributions that was not immediately apparent before. Note that while only the middle anomaly tercile is plotted here, the results for the other terciles are very similar. As can be seen, the relative weight of the two peaks observed in the bimodal case with an initial date in April is indeed not the same for the data and the model. One might even conjecture that the second peak occurs slightly earlier in the model case. A similar effect is visible for an initial date in July, when the model shows first frost occurring with significant probability a few days earlier than for the data.

Since the AR(1) model process seems to lead to similar results as the data for many initial dates, it might still be used to improve the statistics of any prediction made. However, there are some indications that the first passage times might then be underestimated such as a larger weight of the earlier peak in the bimodal case or a slightly shorter mean first passage time for autumn. The occurrence of some very large values in the data is also not supported by the model. A more thorough analysis of model suitability is therefore necessary.

The preliminary analysis of the data has shown that initial dates in spring and autumn seem to induce larger differences in the data-generated distributions than those in summer and winter[4].

[3] See the more extensive representation in Appendix B.2.

[4] While the differences in winter appear quite large, part of this is due to the narrower support of the distributions which results in a higher resolution in the figures.

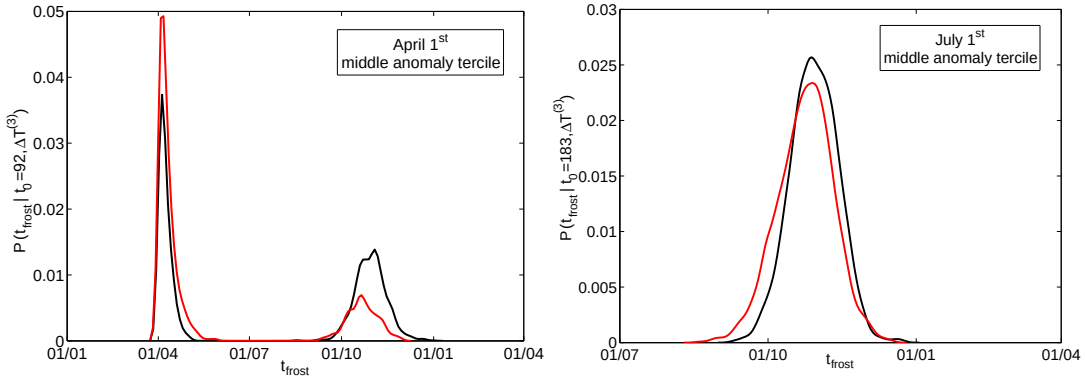


Figure 4.2: Conditional first passage time distribution estimates for the middle anomaly tercile on two different initial dates. The black curve shows the temperature data-generated case, the red curve depicts the case of anomalies generated by an AR(1) model process.

This might albeit be a discouraging sign: In spring and autumn, the next frost day will very probably occur within a short time frame. If the differences in distribution occur exclusively for short first passage times within the usual prediction lead time of conventional weather forecasts, then our scheme might not be very helpful. We will therefore in the following not only conduct a more systematic analysis of the predictability effects for other parameters but also see which effects are remaining on longer time scales.

4.2.2 Change of distribution summary measures under conditioning

In the next part we want to better assess the influence of the initial date on the predictability of the time to first frost. We will therefore consider the effect of t_0 on relevant summary measures of the full conditional probability distributions $P(t_{\text{frost}}|t_0, \Delta T)$, namely the median and the most probable date of first frost, i.e. the location of the peak of the distribution. We also consider the standard deviation, since the information content and therefore also the predictability increase with narrowing distribution.

Figure 4.3 shows the dependence of these summary measures on t_0 both for the measured time series in the left panels and for the AR(1) model process on the right. As can immediately be seen, the changes in these measures brought about by conditioning the first passage times on $\Delta T^{(3)}$ are very slight. Indeed, the median seems to diverge from the unconditioned case $P(t_{\text{frost}}|t_0)$ only in autumn, the standard deviation mostly in winter, while there are no discernible differences in the most probable date of first frost. Note that the median and standard deviation were not considered for initial dates in spring as these measures do not represent a bimodal distribution well.

Looking more closely at the four initial dates considered as representative examples in Sec. 3.5, one can see that both February 14$^{\text{th}}$ and April 1$^{\text{st}}$ were omitted here as they lead to slightly bimodal distributions. This can also be seen in the two distinct peaks in the first passage time distribution evidenced in the lowest panels of Fig. 4.3. For July 1$^{\text{st}}$, no effects of the conditioning are visible here, and the effect for November 1$^{\text{st}}$ is quite small and mostly consists of a slight difference in the median. This confirms our findings of Sec. 4.2.1 rather well.

The differences between the data and the AR(1) model process are more startling. Indeed, not only is the median for initial dates from May to October earlier by more than a week for the model process, also the standard deviation for initial dates from June to November is larger. Moreover, while the large standard deviation resulting from a residual bimodality of the

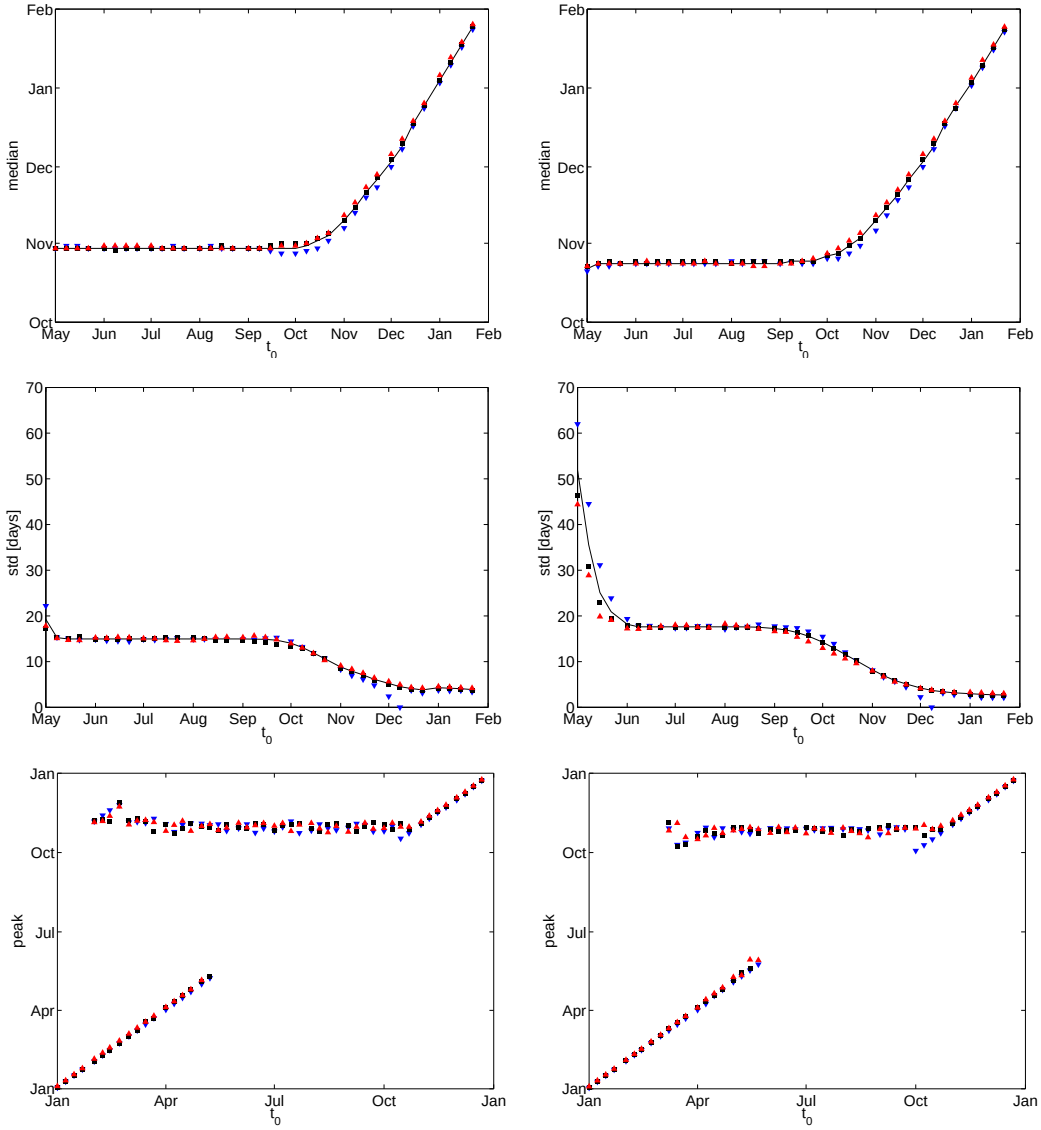


Figure 4.3: Median (upper panels) and standard deviation (middle panels) of $P(t_{\text{frost}}|t_0, \Delta T^{(3)})$, as well as the location of the peaks of maximum first passage time probability (lower panels) for data-generated (left) and AR(1)-generated (right) temperature anomalies. The continuous lines represent $P(t_{\text{frost}}|t_0)$, the lower tercile of ΔT is represented by blue down-facing triangles, the middle tercile by black squares and the upper anomaly tercile by red up-facing triangles.

distribution is absent in the temperature data between May and February, its effects are still visible in May in the model process. Since these effects are highly influenced by the conditioning on initial anomalies, this leads to a large potential predictability in the model that is absent in the measured data for these values of t_0.

The dates of maximal first frost probability, i.e. the location of the peaks of the first passage time distributions, also show significant differences between the model process and the data. Indeed, the peak of first frost after summer appears earlier in the measured data, where the peak of next frost still before summer also disappears earlier, as already seen in the respective standard deviations.

While the predictability effects visible for our previous four example dates were rather slight and therefore not too encouraging, this analysis discovered significant effects for other times of the year. We will therefore choose four more initial dates that are representative of these and should be looked at more closely: May 15$^{\text{th}}$ is a good example of the large difference in standard deviation between model and data, where the bimodality vanishes earlier; September 15$^{\text{th}}$ represents a slight predictability effect in the standard deviation of the data but not the median, contrary to the previous example of November 1$^{\text{st}}$; October 1$^{\text{st}}$ shows the largest effect on the median in the data, which however is almost absent in the model process; and December 1$^{\text{st}}$ shows an effect in both median and standard deviation for data and model.

These eight chosen dates encompass the range of predictability effects contained in the time series. In order to better evaluate these effects, we now need to look in more detail at the influence of the initial anomaly ΔT on $P(t_{\text{frost}}|t_0, \Delta T)$ for these eight dates t_0.

We will therefore look at the summary measures of the conditional first passage time distribution again, analysing how they change with increasing initial temperature anomaly ΔT. In order to enable a more detailed evaluation, we change the coarse-graining of the initial anomalies for each date t_0 to deciles, i.e. ten separate groups of increasing magnitude, denoted in the following by $\Delta T^{(10)}$. Since the initial anomalies are subject to the additional condition of an initial temperature $T > 0$ °C, both the number and the actual anomaly values within a decile differ from one initial date to the next.

Between November 1$^{\text{st}}$ and April 1$^{\text{st}}$, the median of $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(10)})$ clearly depends on the initial anomaly decile, as Fig. 4.4 shows exemplarily in the top left panel[5]. In order to make sure that these effects are not due to statistical fluctuations arising because of the limited number of data points for which the initial temperature still lies above 0 °C, we used the standard error from a bootstrap calculation with 1000 samples (see e.g. [26]) to obtain an estimate of the statistical error[6]. This indicated that the dependence is indeed statistically significant when considering a 95%-confidence interval. Note that this time our analysis does include bimodal distributions for initial dates in spring. For this case, the mean directly reflects the shifting relative weight of the two peaks and is thus an interesting quantity for predictability even it is meaningless as an indication of the location of the distribution.

For an initial date of May 15$^{\text{th}}$, no influence of the initial anomaly is visible. We would also have expected this result for an initial date of July 1$^{\text{st}}$, since the probability distributions for different anomaly terciles were identical in Fig. 4.1. However, the top right panel of Fig. 4.4 shows that the mean seems to change slightly for the warmest initial anomalies. In this case, however, the effect is rather small.

For initial dates in autumn, we observe the same as in summer: For initial anomalies in very few extreme deciles (the cold ones in autumn, the warm ones in summer), the mean or median of the distribution is different than in all other cases.

The spread of the conditional first passage time distributions on the other hand only shows a

[5] As before, a full display of all figures can be found in Appendix B.3, with only some illustrative examples shown here.

[6] For more details on the bootstrap and the choice of the specific number of samples used here, see Sec. 2.2.6.

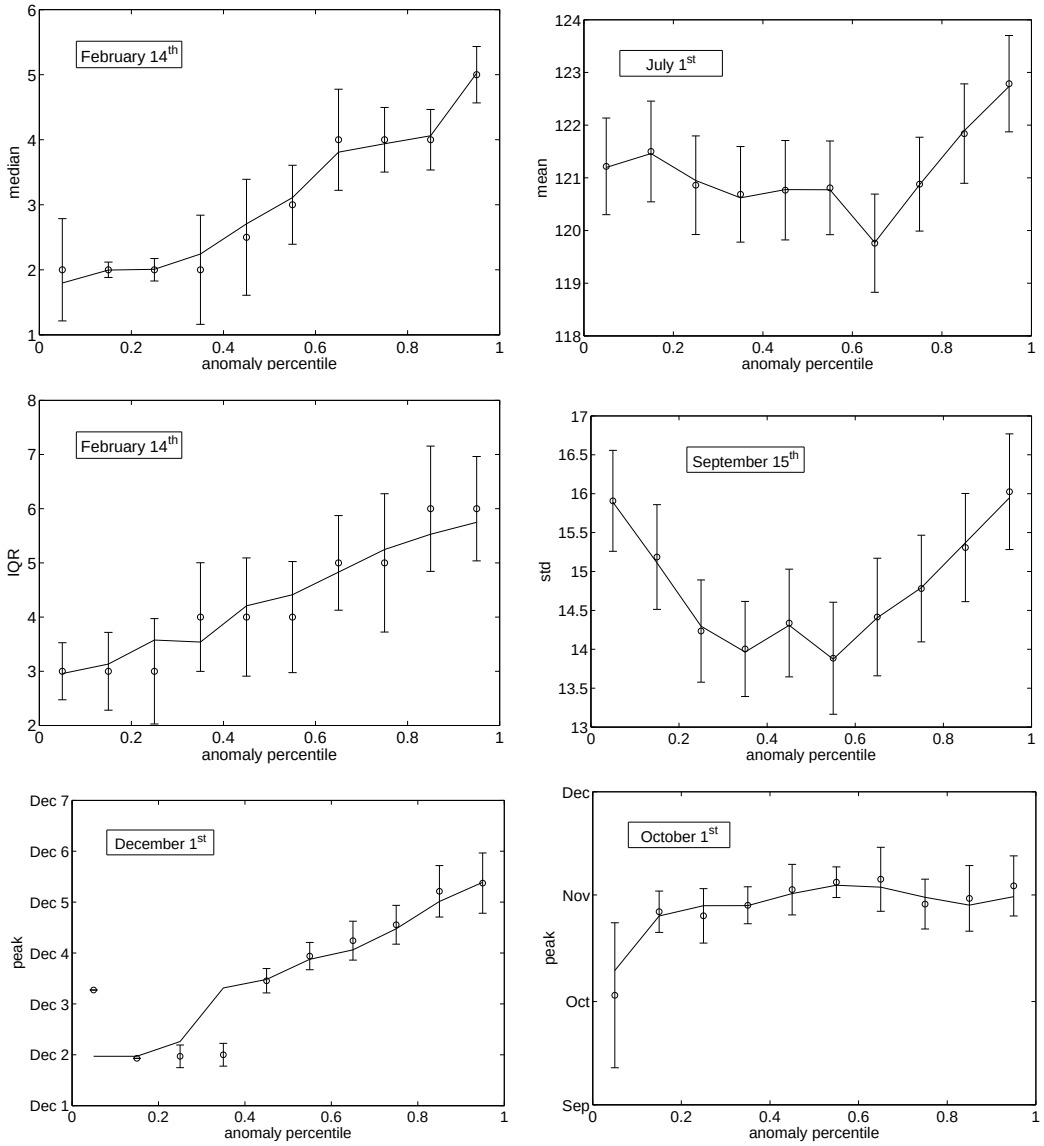


Figure 4.4: Location and spread of $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(10)})$ measured in days, as well as the date of the most probable time of first frost for different initial dates. The continuous lines represent the mean of 1000 bootstrap samples from the data, the error bars the corresponding 2σ confidence intervals.

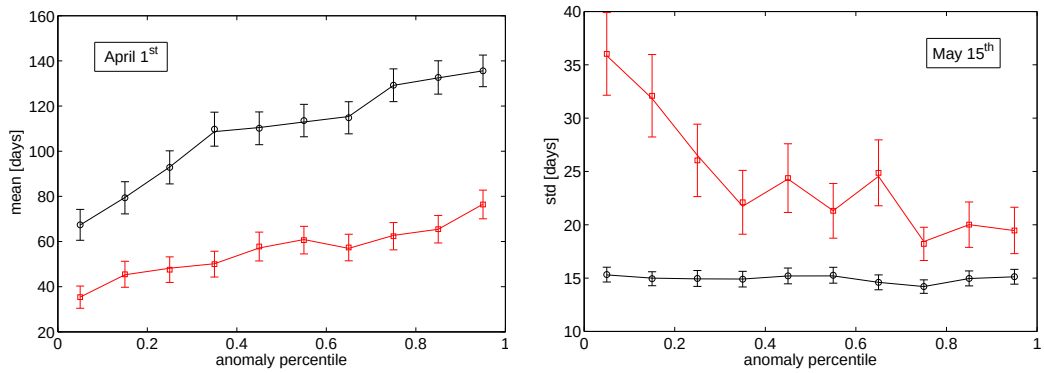


Figure 4.5: Location (left panel) and spread (right panel) of $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(10)})$ (black) and $P_{\text{AR(1)}}(t_{\text{frost}}|t_0, \Delta T^{(10)})$ (red) for two specific values of t_0. The continuous lines represent the mean of 1000 bootstrap samples from the data, the error bars the corresponding 2σ confidence intervals.

similar dependence on $\Delta T^{(10)}$ for the initial dates of December 1$^{\text{st}}$ and February 14$^{\text{th}}$ (as shown in the middle left panel of Fig. 4.4), where it increases with the initial anomaly.

Interestingly, as can be seen in the middle right panel of Fig. 4.4, there are initial dates for which the spread shows a clear but non-monotonic dependence on the initial anomaly. Thus the first passage time distribution to frost contains more information for extreme initial anomalies if one considers initial dates in late spring, while the spread is narrower for average $\Delta T^{(10)}$ for an initial date of September 15$^{\text{th}}$.

Looking again at the location of the peaks of maximum probability in the distribution as illustrated in the lowermost two panels of Fig. 4.4, it can be seen that there only seems to be a dependence on $\Delta T^{(10)}$ for initial dates from November 1$^{\text{st}}$ to mid February, with some effect visible at least for the coldest anomalies already at the beginning of October. However, the bimodality in the distributions with initial dates in spring becomes visible as early as February 14$^{\text{th}}$ at least for larger anomaly values[7].

The figures in Appendix B.3 not only show the summary measures of $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(10)})$, but also those for $P_{\text{AR(1)}}(t_{\text{frost}}|t_0, \Delta T^{(10)})$. Comparing their respective dependencies on $\Delta T^{(10)}$ allows us to get additional insights into the validity of the model for our purposes.

The left panel of Fig. 4.5 shows the largest mismatch between the data and the model in terms of the mean (or median) of the first passage time distribution. Indeed, for initial dates between April 1$^{\text{st}}$ and October 1$^{\text{st}}$, the model severely underestimates the average time to first frost by up to two months. During the winter months, however, the median of the measured first passage time is well represented by the model process.

When considering the spread instead, a similar picture emerges. For initial dates between April 1$^{\text{st}}$ and September 15$^{\text{th}}$, there is a mismatch between the model and the data: Not only the absolute value of the spread is different between the two time series, but also the functional dependence on $\Delta T^{(10)}$. This can be seen exemplarily in the right panel of Fig. 4.5. Only for an initial date of November 1$^{\text{st}}$ the spread is well-represented by the model, with the other cases leading to a similar picture at least within the error bars gathered from a bootstrap calculation.

The most directly obvious problem posed by the model process occurs in the peaks, i.e. the dates of maximum first frost probability, as already seen before in Fig. 4.3. Indeed, the bimodality in the first passage time distribution occurs later in the model than in the data, leading to

[7] The bimodality in this case is very slight with between zero and five data points per decile indicating first frost only after the summer.

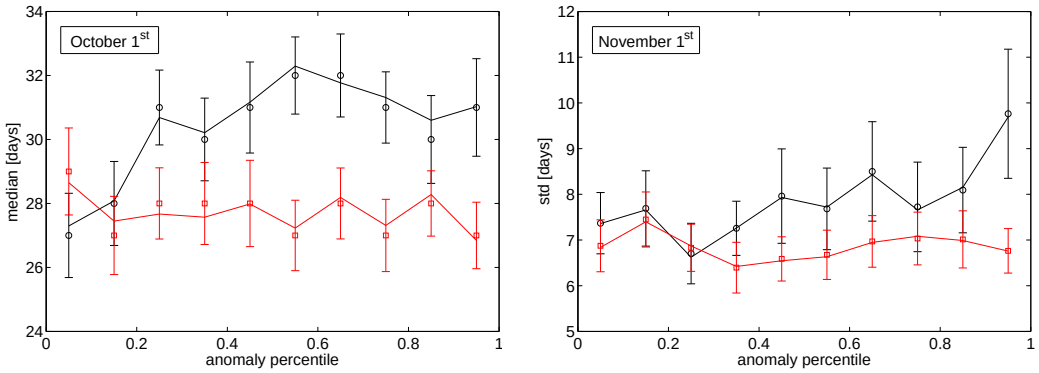


Figure 4.6: Location (left panel) and spread (right panel) of $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(10)})$ (black) and $P_{\text{AR(1)}}(t_{\text{frost}}|t_0, \Delta T^{(10)})$ (red) for $t_{\text{frost}} > t_0 + 8$. The continuous lines represent the mean of 1000 bootstrap samples from the data, the error bars the corresponding 2σ confidence intervals.

initial dates for which the model shows cases of first frost still before summer that do not occur in the data and vice versa.

Therefore, the model only represents the data well when considering initial dates in winter, making it rather unsuitable for enhancing the statistics at any other time. Any results gathered from longer simulated time series or an ensemble simulation using the model process should therefore be viewed with caution. While this is at first glance somewhat disappointing, it also shows that there are potentially still long-range correlations inherent in the data, an encouraging result for the predictability of first frost for longer lead times.

Apart from showing the model inadequacies, we have also seen that predictability effects are visible for initial dates roughly between September 15$^{\text{th}}$ and April 1$^{\text{st}}$, while $\Delta T^{(10)}$ does not influence the first passage time distribution to frost for initial dates in summer. This is hardly surprising as it suggests an upper bound on the lead time of any prediction efforts, meaning that the initial date needs to be sufficiently close to the date of first frost for useful predictions.

However, since the largest effects of the initial condition can be observed when the first passage time to frost is rather short, it gives rise to a different question. Since the current operational weather forecasts are already useful for lead times of up to 8 days[8], any prediction scheme gleaned directly from the data should retain indications of predictability over a longer time scale than this.

In the following, we shall therefore explore the persistence of the predictability effects when conditioning additionally on first passage times larger than 8 days, i.e. $P(t_{\text{frost}}|t_0, \Delta T^{(10)})$ for $t_{\text{frost}} > t_0 + 8$. This excludes those cases for which the current operational weather forecasts prove already adequate and thus focuses only on the addtional information obtained through our approach. This analysis is superfluous for initial dates in summer such as July 1$^{\text{st}}$, or indeed already May 15$^{\text{th}}$, since then all first passage times at least in the measured data are much larger than two weeks and therefore fulfill the extra condition.

Since this additional conditioning further reduces the number of data points available, any scheme to enhance the statistics would prove especially helpful here. We shall therefore again include a comparison with results gathered through a single realisation of an AR(1) model

[8]The European Centre for Medium-Range Weather Forecasts defines the useful range of a deterministic forecast as the time for which the anomaly correlation coefficient ACC, which can be used to determine potential forecasting skill (see also Sec. 2.3.3) remains above 60%. This is currently the case for up to 8 days' lead time[137].

process. From the overview of all relevant figures in Appendix B.4 and the two examples shown in Fig. 4.6, it is immediately evident that the discrepancies between $P_{\mathrm{AR}(1)}(t_{\mathrm{frost}}|t_0, \Delta T^{(10)})$ and $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(10)})$ for $t_{\mathrm{frost}} > t_0 + 8$ are still significant even though they are not quite as large as for the full distributions. Therefore the model process definitely should not be used to enhance the statistics of the measurements.

However, leaving the model concerns aside, Fig. 4.6 also shows evidence of some predictability effects even after 8 days, although this does not hold for every t_0. Indeed for autumn, i.e. October 1$^{\mathrm{st}}$, both the mean first passage time to frost beyond 8 days into the future and its standard deviation still seem to be increasing with $\Delta T^{(10)}$ (this is also valid for September 15$^{\mathrm{th}}$ as shown in App. B.4). For an initial date of November 1$^{\mathrm{st}}$ on the other hand, only the standard deviation still depends on the initial anomaly decile.

We therefore indeed find some effects beyond the lead time which is covered by operational weather forecasts, but they are much less pronounced and occur for fewer initial dates than when considering all first passage times.

For initial dates in the winter months where the additional condition of an initial temperature above freezing drastically reduces the available number of data points, a scheme to compensate the lack of statistics from the measured data would be highly desirable. However, the previous parts have shown that using a simple AR(1) model process to this effect is not promising due to the significant differences in the respective conditional first passage time distributions. Since the predictability effects observed until now are very slight, a reduction of the errors involved in any prediction scheme is desirable to obtain significant results but it does not seem achievable by enhancing the statistics.

Sec. 4.2.1 has shown that changing the coarse-grained conditioning on ΔT from terciles to much smaller bands of just 500 points each enhances the resulting differences in the distributions. In the first part of Sec. 4.2.2, it could also be seen that there are some predictability effects when conditioning on anomaly deciles that were not visible for terciles. Since the coarsening into anomaly bands is necessary to retain a reasonable number of data points for the estimation of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T)$, an optimisation of the width of the initial anomaly bands might further improve predictability.

To analyse this, we will separate all initial anomalies still corresponding to an initial temperature above freezing into two outer bands of equal number of data points containing the warmest and coldest p% of anomalies. The third and middle band is then made up of all other initial anomalies, i.e. $(100 - 2p)$% of data points. This coarsening scheme of weighted initial anomaly terciles will be designated in the following by $\Delta T^{(3w,p)}$. We will now look at the change in location and spread of the resulting conditional first passage time distributions with changing outer anomaly bandwidth p in order to identify an optimum p_{opt}.

Fig. 4.7 shows three exemplary initial dates from the period between September 15$^{\mathrm{th}}$ and April 1$^{\mathrm{st}}$ where predictability effects were found previously.[9] As can be seen in the first row for initial dates in autumn, the mean first passage time to frost is significantly different for the lowest anomalies with differences in the mean of up to 10 days. The results for average and high initial anomaly bands, however, cannot be distinguished. While this effect is significant over almost the whole range of p, it increases for very small outer bands that only contain around 5% of all data points. Looking at the spread of the conditional first passage time distributions, the influence of the anomaly bandwidth is much less clear.

Starting the predictability analysis later in the year on December 1$^{\mathrm{st}}$, the mean values for the three anomaly bands are significantly different with very small bootstrapping errors. These differences again extend over the whole range of anomaly bandwidths with an optimum around

[9] In fact, we confirmed that there are no anomaly bandwidths leading to the appearance of statistically significant predictability effects for the previously used initial dates of May 15$^{\mathrm{th}}$ and July 1$^{\mathrm{st}}$.

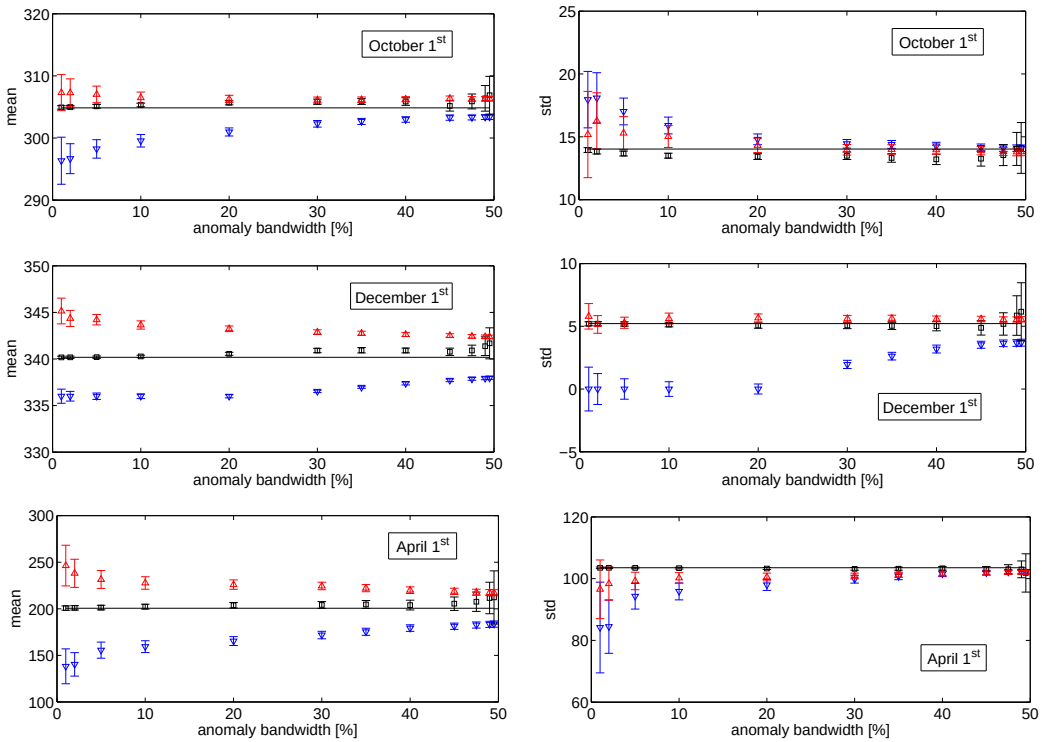


Figure 4.7: Location and spread of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(3w,p)})$ and their evolution with p for different values of t_0. The results from the highest anomalies are depicted using red upper triangles, from the middle anomalies with black squares, from the lowest anomalies with blue lower triangles. The continuous black line shows the result for $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0)$. The error bars indicate the 2σ confidence interval obtained using a bootstrap calculation with 1000 samples. The standard deviation is measured in days, but contrary to before, the mean value is given here by the calendar day, where October 1$^{\mathrm{st}}$ corresponds to $t = 275$, December 1$^{\mathrm{st}}$ to $t = 335$ and April 1$^{\mathrm{st}}$ to $t = 92$.

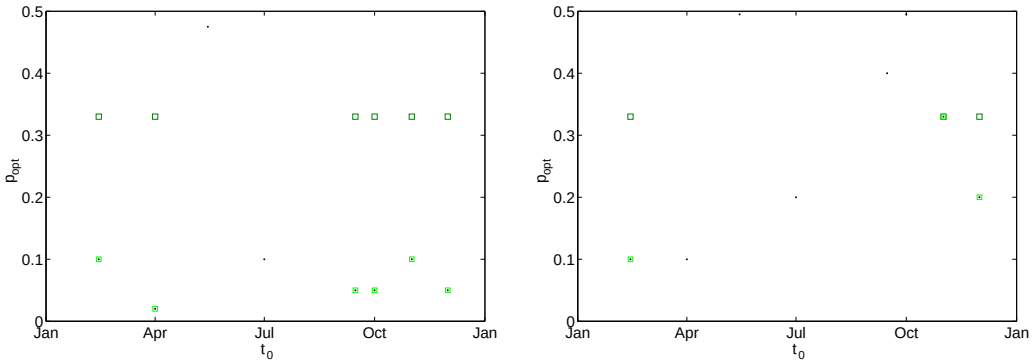


Figure 4.8: Optimal percentile of data points in the outer anomaly band when considering the mean first passage time (left panel) or its standard deviation (right panel) depicted by black dots. The light green squares show where p_{opt} leads to a statistically significant distinguishability between the lower and upper anomaly bands, the dark green squares show for which t_0 this is also valid for $\Delta T^{(3)}$.

$p = 5\%$ and a corresponding maximal difference in mean of seven days. However, the magnitude of this effect is almost the same for the optimum bandwidth and exact terciles. The standard deviation of the first passage time distribution conditioned on the lowest anomalies is significantly lower than in the unconditioned case, going so far as to drop to 0 days for the smallest outer anomaly bands. This means that frost will occur with certainty on the next day in these cases, a fact confirmed by the mean of day 337 of the calendar year in these cases, which corresponds to December 2$^{\mathrm{nd}}$.

The bottom row in Fig. 4.7 shows the latest initial date within the period when some predictability can be observed, namely April 1$^{\mathrm{st}}$. In this case, the three anomaly bands lead to mean first passage times to frost as far apart as 70 days between the outer bands. This large value is of course due to the bimodality in the distribution for which the mean is an ill-suited location measure. The difference here represents the changing weight of the two peaks rather than a true shift in peak location. The observed differences in the mean again increase with smaller outer bands, leading to an optimum of 2% as for even smaller bands the errors in the distribution estimation become too large. However, as before the improvement of the optimum over terciles is not large. The standard deviation in this case is much bigger than before due to the bimodality. We can see here that the conditional distributions have smaller standard deviation than the unconditioned one, something that ought to have a positive impact on predictability.

As Fig. 4.7 only highlighted the results for three different initial dates, we look at the optimal anomaly bandwidth p_{opt} for all eight dates considered up to now in the following. Fig. 4.8 shows that if one is considering the mean value of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(3w,p)})$, both using p_{opt} and a separation into terciles lead to statistically significant differences between the upper and lower anomaly bands[10] except for those initial dates where no predictability effects were found previously. The relative improvement in the separation of the mean first passage times for the upper and lower anomaly band when changing from terciles to p_{opt} is smaller than 2%, except in April when it is closer to 10%.

When looking at the standard deviation of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(3w,p)})$ instead, the statistical significance of the differences between the upper and lower initial anomaly bands is not nearly as widespread as for the mean values. In fact, only in winter these differences are statistically significant, but then this is again valid for both the optimum and terciles. However, in this

[10] A statistically significant difference in this case is taken to mean that the 2σ error bars obtained through a bootstrap calculation with 1000 sample points do not overlap.

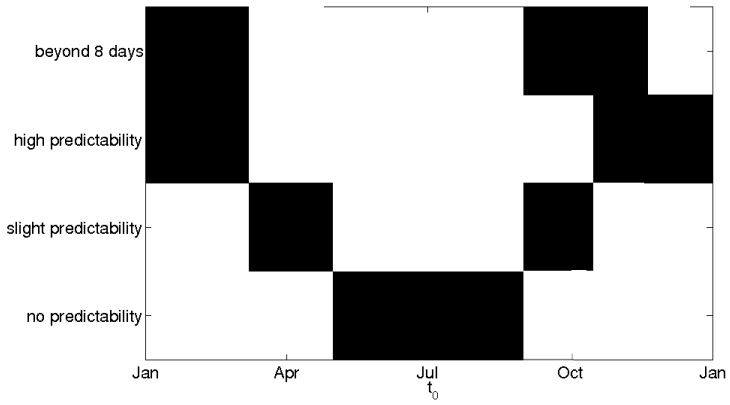


Figure 4.9: Schematic showing in black the initial dates t_0 for which predictability effects were found in this and the previous subchapter.

case, the relative improvement of an optimal bandwidth over simple terciles in the statistically significant cases of February 14$^{\text{th}}$ and December 1$^{\text{st}}$ is of the order of 50%.

The analysis of p_{opt} shows that even though optimal bandwidths do not introduce predictability effects where there were none before, predictions might indeed be improved by choosing $\Delta T^{(3w,p)}$ instead of $\Delta T^{(3)}$. On the other hand, the improvement is overall rather small.

Summarising, predictability effects are almost non-existent for initial dates in the summer months and very strong for initial dates during the winter, i.e. close to the threshold of frost. If one excludes all first passage times that fall within the short lead times for useful current operational weather forecasts, then slight predictability effects remain for initial dates between January and March and between September and the end of November[11].

To provide a clearer overview of the results from this section, Fig. 4.9 shows a schematic summary of the initial dates for which predictability effects were observed.

4.2.3 Statistical tests of significance in distribution differences

Even though the error bars in the previous analysis help to assess the significance of the differences in first passage time distributions conditioned on very low or high initial temperature anomalies, the effects observed are not extremely large. Moreover, while bootstrapping is the best method to assess errors caused by statistical fluctuations, it cannot produce data points in the undersampled tails of the distributions and there are some issues when applying it to correlated data[77]. Another limitation of the previous approach is the difficulty of summarising results for different initial dates across the calendar year and gaining an insight into the more precise seasonal evolution of these predictability effects. One way to evaluate this and to provide additional evidence that the observed predictability effects are indeed significant are statistical hypothesis tests (see Sec. 2.2). They can assess the difference in probability distribution or in specific distribution summary measures such as mean or variance.

An important caveat is provided by Fig. 4.10: The null hypothesis that $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3)})$ is a normal distribution for all t_0 and $\Delta T^{(3)}$ is firmly rejected by a Jarque-Bera test (see also Sec. 2.2.4). Since normality of the distribution is an underlying assumption of most other hypothesis tests, this significantly reduces the available choices of statistical tests to confirm the predictability effects observed in the previous section.

[11] The month of December was excluded from this analysis because the first passage times then almost exclusively fall within the first week after the initial date and thus within the range of current model-based weather

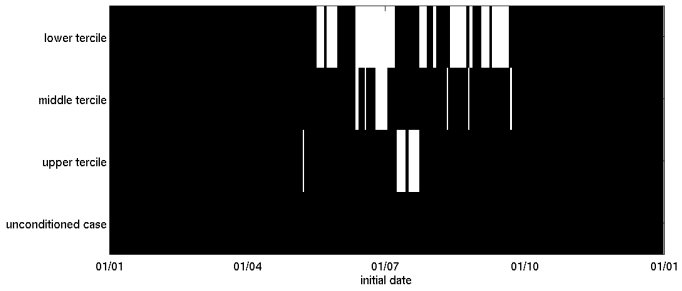


Figure 4.10: Results of a Jarque-Bera hypothesis test for $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(3)})$, as well as $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0)$. Black denotes the cases in which the null hypothesis of a normal distribution was rejected with a significance level of $\alpha = 5\%$.

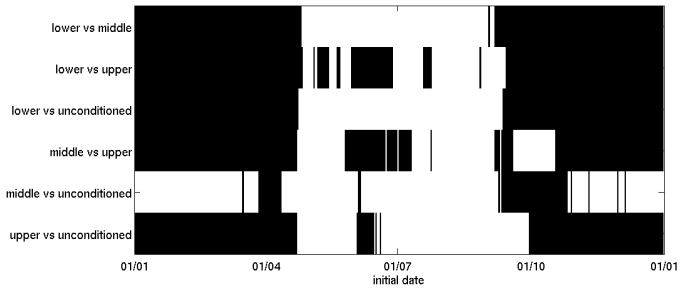


Figure 4.11: Results of a Kolmogorov-Smirnov hypothesis test comparing the first passage time distributions conditioned on the initial date and different initial anomaly terciles. Black denotes the cases in which the null hypothesis of equal underlying probability distribution was rejected with a confidence level $\alpha = 5\%$.

The most robust remaining method is a Kolmogorov-Smirnov test that compares two cumulative probability distributions directly, as introduced in Sec. 2.2.5. Fig. 4.11 shows a summary of the resulting hypothesis test results with a significance level $\alpha = 5\%$ for pairs of differently conditioned first passage time distributions. As can be seen, conditioning on the middle initial anomaly tercile mostly has no effect as the resulting distribution cannot be distinguished from $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0)$, except around the beginnings of April and October. However, in the winter half of the year, the test firmly rejects the possibility of equal underlying probability distributions for all other pairs. The first passage times are therefore clearly influenced by the initial conditions and contain potential for predictability, confirming our previous results.

Moreover, interestingly, the Kolmogorov-Smirnov test also rejected the null hypothesis of equal underlying distribution for initial dates in June, whenever first passage times conditioned on the highest initial anomaly terciles were part of the comparison - a result that was not readily visible in the previous section.

This can be seen more clearly in Fig. 4.12 which plots the p value of the test, i.e. the estimated probability that the first passage time distributions conditioned on the initial date and the lower or upper initial anomaly tercile were indeed drawn from the same underlying distribution. There, two distinct stretches of initial dates for which the null hypothesis cannot be rejected are visible, with a clear exception for the month of June in between.

forecasts.

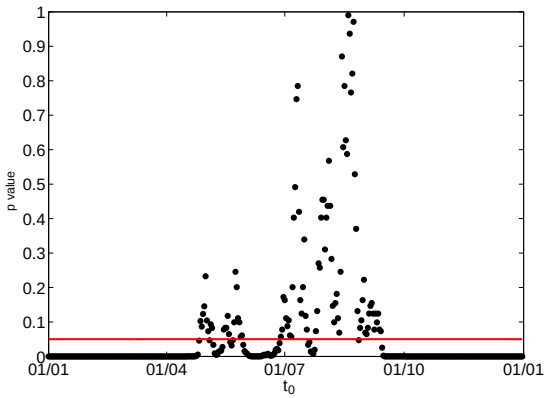


Figure 4.12: p value of the Kolmogorov-Smirnov hypothesis test comparing $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3)})$ with ΔT in the lower and upper tercile, i.e. the probability that the underlying distribution is one and the same. The horizontal red line represents the 5% confidence level.

Since the earlier analysis showed even stronger predictability effects for smaller outer anomaly bands, we repeated the Kolmogorov-Smirnov test using the lowest and highest decile as outer bands. In this case, however, the p value of the test is larger than using terciles for 80% of all initial dates. The distributions conditioned on the outer bands are thus less distinguishable by the test overall. Also, the test fails to reject the null hypothesis of equal underlying probability distribution for more initial dates if smaller outer bands are used. The scarcity of data points when compared to full terciles therefore has a larger influence than the increase in differences observed before.

Long periods of initial dates for which there are indeed statistically significant predictability effects have become evident using the Kolmogorov-Smirnov test. This is very encouraging for our efforts at issuing full predictions of first frost, especially since the hypothesis test used here is rather conservative in cases of discrete distributions such as first passage times[70].

4.2.4 Predictability: Summary and examples

This section has revealed that the differences in $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T)$ are significant over a wide range of initial dates, namely from mid September to the end of April. This was shown not only through an analysis of the dependence of several distribution summary measures on the initial anomaly decile but also through the use of a Kolmogorov-Smirnov test on the cumulative conditional first passage time distributions.

This influence of the initial conditions on the first passage times leads us to conclude that there is significant potential for predictability in the temperature data. In order to illustrate the impact of these findings, we will in the following consider two specific consequences of these distribution differences as examples of potential prediction tasks.

First, let us consider an initial date in spring, namely $t_0 = 92$, i.e. April 1$^{\text{st}}$. As seen before and as shown again in the left panel of Fig. 4.13, $P_{\text{data}}(t_{\text{frost}}|t_0 = 92, \Delta T^{(10)})$ is bimodal with a clear separation of the two peaks. The previous analysis has shown that the specific shape of this distribution significantly depends on ΔT, i.e. on whether the observed temperature on April 1$^{\text{st}}$ is warmer or colder than normal for this date.

In agriculture at this point in time, if it is already warm enough, one might start debating whether to begin certain tasks such as for example sowing the fields or shearing the sheep. Both

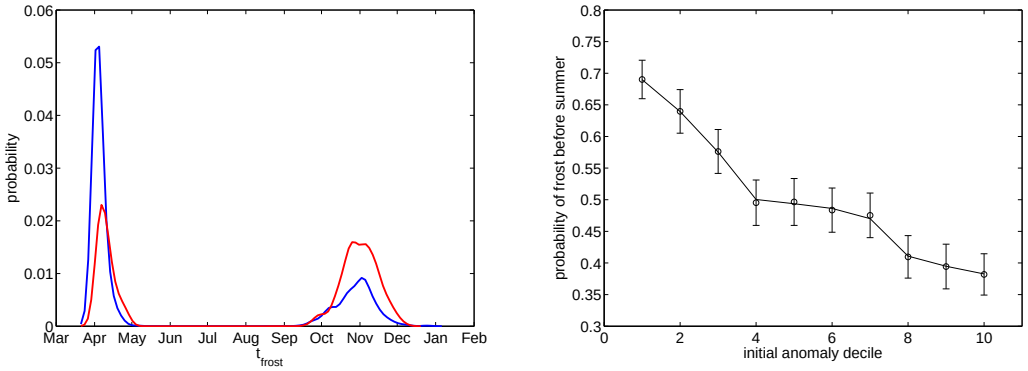


Figure 4.13: Conditional probability distribution of first frost for $t_0 = 92$, i.e. April 1$^{\text{st}}$, and the lowest (blue) and highest (red) initial anomaly decile (left panel). The resulting probability of frost occurring before summer, depending on the initial anomaly decile, is shown in the right panel. The continuous line represents the mean value obtained through 1000 bootstrap calculations, the error bars show double the resulting standard deviation.

are decisions rather sensitively depending on the temperature since one does not want to lose seeds or sheep to freezing. It is therefore important to know whether the temperatures will drop below 0 ℃ once again or not.

The right panel of Fig. 4.13 shows for each initial anomaly decile the probability of frost occurring again before summer. As can be seen, this changes from around 70% if the current temperature conditions on April 1$^{\text{st}}$ are still very cold to less than 40% if it is on the contrary already very warm. There is therefore good potential of providing a measure of weighing the gain of early action against the risk of loss if frost does indeed occur again.

The second example is an initial date in autumn, when $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T)$ is not bimodal, as reminded in the left panel of Fig. 4.14 for an initial date of October 15$^{\text{th}}$. In this case, questions of deadlines arising from the approaching first occurrence of frost are of more interest. A car owner might want to know the time he has left before he needs to mount the winter tires, a winemaker needs to know when his latest opportunity to harvest the grapes is so they can fully mature but do not freeze to death. The right panel of Fig. 4.14 shows that this date of first frost shifts by almost a week depending on initial anomaly conditions.[12]

The analysis in this section thus indeed shows good potential for the prediction of such first passage time related quantities. We will therefore move on to assess actually issued predictions for several such more specific questions in the next section.

4.3 First passage time prediction

The last section has shown that $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T)$ depends on ΔT and therefore on the precise initial conditions in a nontrivial way. These differences in probability distributions hold up to statistical significance tests for most initial conditions. Looking at more specific examples such as the probability of frost still before summer if t_0 is in spring, there is a clear dependence on the initial anomaly decile. However, actual predictions are needed in order to truly assess the quality of purely data-based prediction schemes.

[12] Fig. 4.14 does not actually show the date of first frost, but the date for which in 10% of all cases frost has already occurred. This was chosen to reduce the statistical fluctuations.

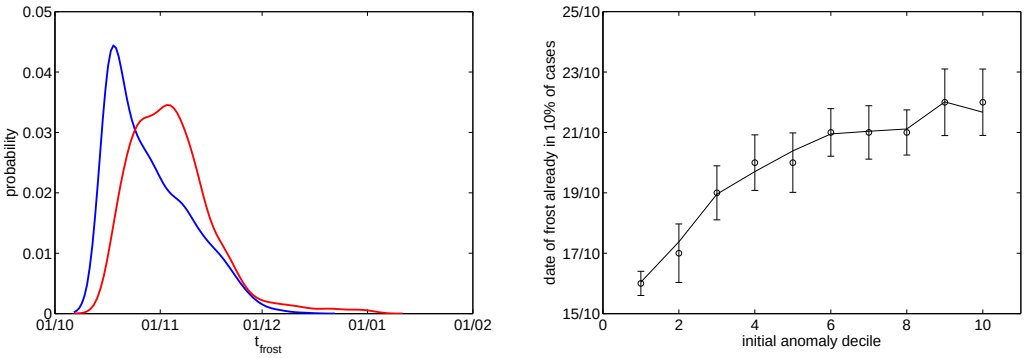


Figure 4.14: Conditional probability distribution of first frost for an initial date of October 15$^{\text{th}}$ and the lowest (blue) and highest (red) initial anomaly decile (left panel). The right panel shows the resulting date for which in 10% of cases first frost has already occurred, depending on the initial anomaly decile. Here, the continuous line again represents the mean value obtained through 1000 bootstrap calculations, the error bars show double the resulting standard deviation.

Up to now, we have taken all available data within our time series to perform the analyses. However, forecast verification needs to be conducted out-of-sample, i.e. the data used to build a prediction scheme should not be used to also verify it. In this way, the pitfalls of overfitting, i.e. finding a prediction scheme that fits the time series specifically but does not generalise to other data sets, can be avoided[13].

In order to achieve out-of-sample predictions, we will start by calculating the climatology, resulting anomalies and first passage times from the first 30 years of the time series, i.e. from 1893 to 1922. We then verify the predictions on the next year, namely 1923. This process can be iterated so that we will use the whole time series up to the year 1923 to predict the data measured in 1924, and so on. We therefore operate with a yearly updated forecast scheme. Often, cross validation is used as an alternative, where only the verification year is left out from the forecast process which therefore uses data both from the future and the past (see for instance [49]). However, we chose the yearly update as it tends to underestimate the actual future forecast skill rather than being overconfident, as the forecasts will mostly be based on much shorter time series than those that are actual forecasts issued at the present time[74].

The actual forecasts and also the scoring used to assess their quality depend significantly on the precise question one wants to answer and on the costs and benefits associated with wrong or correct forecasts. In the case of the first passage time to frost, there are several distinct possibilities.

4.3.1 Deterministic predictions

We will start with the most immediate prediction goal: A deterministic forecast of the specific day on which we anticipate first frost to happen. But before proceeding to the predictions, we still need a benchmark to compare the forecast scores against. This is usually the most simple and uninformed scheme available that we intend to improve upon. In the case of the deterministic forecast, the simplest prediction would be based on the climatology, since this already describes how the temperatures change throughout the seasons. Depending on the initial date t_0, there are two possibilities: Either the climatological temperature $\widetilde{T}$ is already

[13]For more information on forecast verification see Sec. 2.3.

forecast scheme	benchmark	no ΔT	$\Delta T^{(3)}$	$\Delta T^{(3w,10)}$	$\Delta T^{(10)}$
RMSE [days]	73.5	31.1	30.8	31.0	30.8

Table 4.1: Root mean square error for deterministic forecasts of the first day of frost, using the climatology for the benchmark, an estimate of $P_{\text{data}}(t_{\text{frost}}|t_0)$ and an estimate of $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T)$ with differently coarse-grained initial anomalies ΔT.

below the threshold of 0 ℃. Then we expect frost to happen at any moment and issue the forecast of 1 day for the first passage time to frost. If, on the other hand, $\widetilde{T} > 0$ ℃, then we issue the time left until $\widetilde{T} \leq 0$ ℃ as prediction. This is obviously independent both of the initial temperature anomalies and of any knowledge of first passage time distributions.

To issue more informed predictions from the data, we use the mean of $P_{\text{data}}(t_{\text{frost}}|t_0)$ to compare the benchmark against another scheme that also does not use the initial anomalies. We then also calculate the mean of $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T)$ with the condition on ΔT again coarse-grained into terciles $\Delta T^{(3)}$, weighted terciles $\Delta T^{(3w,p)}$, and deciles $\Delta T^{(10)}$.

The most accessible score for such a deterministic forecasting scheme is the root mean square error (RMSE) as described in Sec. 2.3.3 that evaluates by how many days the forecast is off on average. Table 4.1 lists the scoring results for the different forecasting schemes employed here.

As expected[80], the benchmark performs much worse than all predictions that rely on a first passage time distribution estimate. This is mostly due to the correlations in the temperature data that reduce the probability of temperatures falling below the threshold on the next day if they currently still lie above it. In winter, the benchmark thus issues far too many forecasts of $t_{\text{frost}} = 1$ day, corresponding to a large negative bias. Moreover in spring, when frost occurs in the immediate future but the climatological temperatures are already positive, the benchmark will be off by more than half a year. If the more informed forecasts based upon the mean conditional probability are wrong, the error will only be about half as large. These differences between forecasts and benchmark lead to a skill score, i.e. an improvement over the benchmark, of around 58%. However, the different coarse-graining schemes of the initial anomalies do not seem to have much influence on the prediction quality. The error of the RMSE value as given by Eq.(4.1), where $\hat{\sigma}$ denotes the standard deviation of the score values across all verification days and n the total number of verification days, amounts to roughly 0.4 days[14]. They are therefore statistically indistinguishable, and even the prediction based only on $P_{\text{data}}(t_{\text{frost}}|t_0)$ does not perform significantly worse than the others, when evaluated with this score.

$$\Delta(\text{RMSE}) = \frac{1}{2 \cdot \text{RMSE}} \frac{\hat{\sigma}}{\sqrt{n}} \tag{4.1}$$

In the previous section, we have looked at the evolution of predictability with the initial date. Using the deterministic predictions and evaluating each initial date t_0 separatedly, we can assess this seasonality in the prediction quality directly. Fig. 4.15 shows the resulting RMS error for the prediction using $\Delta T^{(3)}$. As expected, this reproduces the seasonality of the standard deviation of $P_{\text{data}}(t_{\text{frost}}|t_0)$ as shown previously in Fig. 3.21 quite nicely and confirms that predictions in summer are less accurate than those in winter. However, the spread of the distribution is very large in spring and the mean value does not represent a bimodal distribution at all well. The deterministic predictions of a specific day of first frost therefore perform rather poorly for initial dates in spring even though there were significant predictability effects then.

Fig. 4.15 also illustrates that the RMSE amounts to at least two weeks for most initial dates across the calendar year. This shows that such a forecast directly from the data is in general

[14] Using a normal approximation to the distribution of MSE verification score values, we assume that they have standard error given by $\hat{\sigma}/\sqrt{n}$. Eq.(4.1) then gives the resulting error for the RMSE.

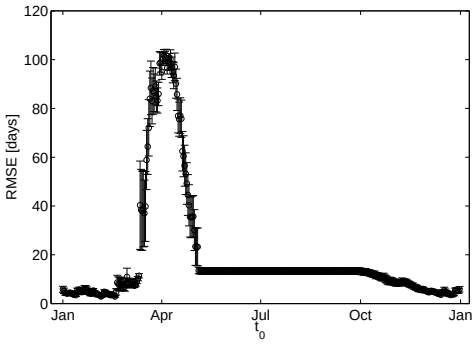


Figure 4.15: Dependence of the root mean square error for the predictions using $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3)})$ on t_0. The error bars were obtained using Eq.(4.1).

not very good and therefore also not very useful, even though it is still significantly better than the benchmark forecast for most initial dates t_0. Only in winter, when both forecast schemes predict a first passage time to frost of 1 day, they are of comparable quality.

In Sec. 4.2.4, we looked at the date in autumn on which in 10% of all cases first frost after summer had already occurred. This is a possible alternative for the deterministic forecast that would score very badly with the root mean square error, since the RMSE is sensitive to larger outliers. While the score used up to now is therefore good for conservative forecasts, it strongly penalises schemes such as the benchmark or using the first observed day of frost instead of the mean, where the errors will occur mostly in one direction and contain large outliers.

If one wants to know the first possible occurrence of frost and needs the exact day, i.e. if it does not matter by how many days the forecast was off, but only if it was exactly correct or not, then other scores are therefore much more appropriate. The most accessible of these is the proportion correct (PC)[15], the percentage of forecasts that fell on the correct day. This has the added advantage of using discrete data such as first passage times without any continuity assumption contrary to the RMSE.

The left panel of Fig. 4.16 compares the proportion correct of different forecasting schemes that are all based on $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3)})$. As can be seen, the mean first passage time to frost actually performs worst when one wants the greatest possible number of correct forecasts, even the benchmark is better for this requirement. This is in no small part due to the contribution of initial dates with a bimodal first passage time distribution. Then, the mean value of the distribution falls on a date in the middle of summer which has a vanishing probability of first frost, while the other versions insure that the forecast date always has some probability of first frost[16]. Among the other possibilities that all outperform the benchmark, choosing the date for which in 10% of all cases first frost had already occurred in the training set proves to be the forecasting scheme that most often predicts the correct date of first frost. However, it predicts correctly only in roughly 6.5% of cases and is therefore not as accurate as might be desired either.

The right panel of Fig. 4.16 shows that for the best prediction scheme in terms of the proportion correct of the exact date of the prediction, one can indeed achieve an improvement over the benchmark from spring to autumn, even though the absolute level is dismal for both schemes. In winter, however, the benchmark can outperform the prediction scheme. This is in part due

[15] For more detail see Sec. 2.3.3.

[16] The case of the median is intermediate. If the two peaks have the same weight, then the median might actually also fall in summer, but this occurs only very rarely.

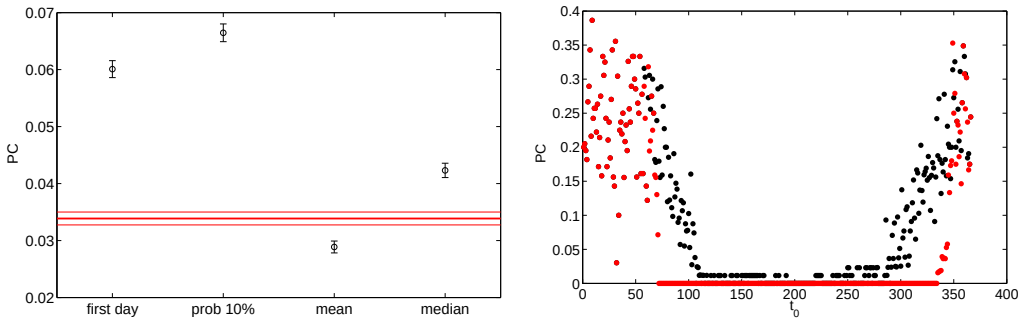


Figure 4.16: Comparison of different forecasting schemes based on the estimate of $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3)})$, namely using the first day after t_0 with non-vanishing probability of frost, the date on which in 10% of all cases in the training set first frost after t_0 had already occurred, as well as the mean and the median of the distribution (left panel). The horizontal line shows the proportion correct (PC) of the corresponding benchmark forecast. All error bars show the estimated standard error of the mean proportion correct. The right panel shows the seasonal variation of the PC of a forecast of first frost using t_{forecast} such that $P_{\text{data}}(t_{\text{frost}} \leq t_{\text{forecast}}|t_0, \Delta T^{(3)}) = 10\%$ (black) as well as the PC of the benchmark forecast (red).

to the fact that the proportion correct is heavily influenced by the most probable outcome[49]. If the most probable date of first frost is the next day (i.e. $t_{\text{frost}} = t_0 + 1$), the benchmark is always correct if the verification falls into this category and thus achieves better proportion correct. However, in this case the error bars are very large (they are not shown in the figure for clarity), so that this is not statistically significant.

The two scores employed in this section have shown that the best scheme to forecast the next day of frost highly depends on the aim. If one wants to be closest to the true date on average, then the RMSE shows that using the mean conditional first passage time to frost is the best predictor. If on the other hand one needs to predict the exact correct date as often as possible, using the date for which in 10% of all cases in the training set frost had already occurred proves to be much better.

4.3.2 Binary predictions

As seen before, the deterministic forecasts are rather bad for initial dates in spring when the first passage time distributions to frost are bimodal and the variation in observed first passage times are therefore very large. However, in that case there is a more coarse-grained prediction task that would be useful. It corresponds to the other predictability example addressed in the previous section: Will there be frost again before the summer or not?

This is in fact a binary prediction of the event "next frost still before summer", which aims at a yes/no answer. In contrast to the deterministic predictions looked at until now, it will be a prediction of the probability of the event occurring. While it does not answer the question of how soon frost will occur, thus giving much less precise information, the answer is still very important also in agriculture.

For these predictions, it does not make sense to look at initial dates across the calendar year, since the duality of possible answers only occurs for t_0 in spring. For all other t_0, the rate of occurrence of the event that is to be predicted is either 1 or 0, the outcome is certain and any

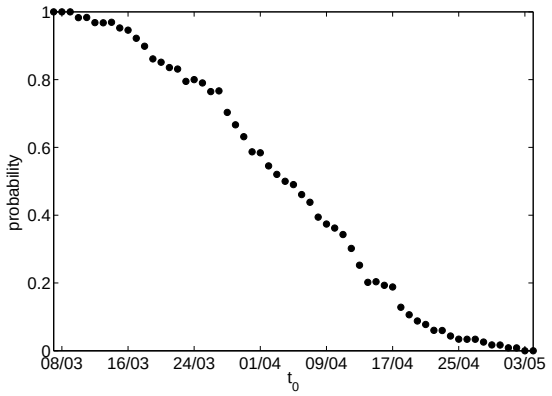


Figure 4.17: Probability of frost still occurring before summer, i.e. the event rate for the binary prediction, depending on the initial date t_0.

prediction will be trivial. We are therefore only going to look at $t_0 \in [70, 123]$, i.e. at the period between March 10^{th} and May 2^{nd} (see Fig. 4.17).

As before, we need a benchmark to compare the forecasts against and a score to rate our forecast skill. We will again base our benchmark on the climatology, always issuing the climatological rate r of frost still occurring before summer for the initial date t_0 in question. The rate is obtained with the same yearly update as the other forecasts following Eq.(4.2).[17] This is, of course, a much more refined benchmark than the one used for the deterministic predictions: While it is independent of the initial temperature anomaly, it still makes use of first passage time properties in addition to the climatology.

$$r = P_{\text{data}}(t_{\text{frost}} < 183 | t_0) \tag{4.2}$$

The most widespread score for this type of predictions is the Brier score as introduced in Sec. 2.3.3. For an easier comparison with the benchmark, we will also look at the Brier skill score (BSS) defined by Eq.(4.3). It directly measures the improvement of the forecasting scheme over the benchmark forecast.

$$\text{BSS} = 1 - \frac{\langle \text{BS}_{\text{forecast}} \rangle}{\langle \text{BS}_{\text{benchmark}} \rangle} \tag{4.3}$$

To get a first impression of the performance of our forecasting scheme, we will consider the average of the Brier scores over the whole time period as previously determined, namely for $t_0 \in [70, 123]$. Table 4.2 shows the results. As can be seen, the scores are very close together. It is therefore imperative to also assess the statistical significance of the observed differences. However, the Brier score values for different verifications are not distributed according to a normal distribution but they rather form several δ-peaks. The standard error of the mean is therefore not an appropriate measure of the variation in the mean score that can be expected from the procedure. We instead used the bootstrap method to get an idea of the statistical significance of the improvement the forecast makes over the benchmark (for a more detailed explanation see Sec. 2.2.6).

Table 4.2 shows that while all forecast schemes seem to improve on the benchmark, the overall improvement is very slight with less than 4%. Moreover, taking into account the statistical errors obtained through the bootstrap method, this improvement is not statistically significant for any of the forecast schemes.

[17] Before summer here means before July 1^{st} ($t = 183$). Note that the rate is calculated without coarse-graining the initial condition on t_0 in contrast to the other forecasts.

	benchmark	$\Delta T^{(3)}$	$\Delta T^{(3w,10)}$	$\Delta T^{(10)}$
Brier score	0.139 ± 0.003	0.134 ± 0.003	0.136 ± 0.004	0.134 ± 0.003
Brier skill score	0%	3.3%	2.1%	3.4%

Table 4.2: Brier score and Brier skill score for different forecasting schemes and the benchmark applied to initial dates $t_0 \in [70, 123]$. The errors were determined using a bootstrap procedure with 1000 samples.

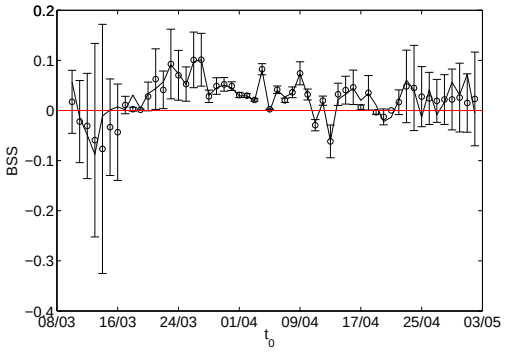


Figure 4.18: Brier skill score comparing the binary forecast of frost still occurring before summer using $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3)})$ to the benchmark. The error bars were obtained using twice the standard deviation of 1000 bootstrap samples, the horizontal red line denotes no forecast improvement over the benchmark.

However, Brier skill scores using the climatology as a benchmark are often very low and are thus considered a very harsh standard to measure a forecast against[90]. Looking at the Brier scores themselves, a value of BS ≈ 0.14 lies at the lower end of the usually obtained score values[85] and does therefore not appear to be uncompetitive overall.

However, the Brier score is also very rate-dependent, with good scores the easier to obtain the rarer the event is. The forecast quality is therefore likely to change across the time window. A summary score across the whole time window might therefore be dominated by regions with large uncertainty that influence the average more strongly[138]. One should therefore also look at the scores for each of the initial calendar days in order to consider quasi-homogeneous subsamples.

Fig. 4.18 shows the change of the Brier skill score with the initial date. As t_0 was chosen in such a way that $0 < r < 1$, there is no automatic certainty in the benchmark forecast in this setup. As can be seen, while the rate is very close to $r = 1$, the errors for the Brier skill score are very large due to the existence of very few non-events that act like outliers. Since the score is very sensitive to the climatological rate, a single good forecast of a rare case happening, i.e. of next frost only occurring after summer, results in a much better score than all other verifications, therefore heavily influencing the mean score value. The same is true for the reverse case of $r \approx 0$. No statistically significant conclusions as to the forecast quality can therefore be gathered before March 17^{th} or after April 22^{th}, leaving a very small useful forecasting window.

Within this window, there are a few calendar days when the forecast performs worse than the benchmark. Most of these are at the beginning or end of this forecasting period, leaving only two dates towards the middle of the window where the forecast quality is dismal, namely April 11^{th} and April 13^{th}. For all other initial dates, the forecast outperforms the benchmark, although the improvement is mostly between 0 and 5% and does not exceed 20% even considering the error bars.

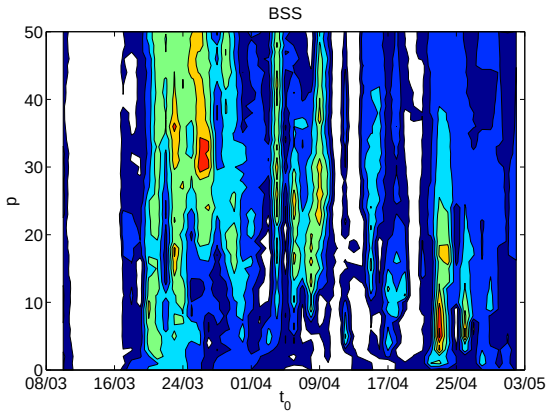


Figure 4.19: Contour plot of the Brier skill score comparing the binary forecast of frost still occurring before summer using $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3w,p)})$ to the benchmark. White denotes the region of negative skill scores, the other colours denote increasing skill scores in increments of 2% from dark blue for BSS $\in [0\%, 2\%]$ to dark red for BSS $\in [12\%, 14\%]$.

Up to now, we have only looked at prediction schemes with $\Delta T^{(3)}$, $\Delta T^{(10)}$ and $\Delta T^{(3w,10)}$, never assessing whether other methods of conditioning on the initial anomalies might improve the prediction quality. Specifically using weighted terciles but for other widths of the outer anomaly bands might be an interesting choice.

Fig. 4.19 shows the dependence of the Brier skill score both on the proportion p of the outer initial anomaly bands and on the initial date t_0. As can be seen, forecasts that perform worse than the benchmark occur almost independently of p and mostly at the beginning of the analysed time window and around mid April, where the two isolated dates were previously observed in Fig. 4.18. Forecasts for t_0 later than April 24$^{\text{th}}$, which were previously rejected for poor quality, now appear reasonable, as the large errors were not considered in this case.

Interestingly, the previously considered exact terciles lead to the best forecasts for initial dates in late March. The later the initial date, the smaller the outer initial anomaly bands should be to obtain the best possible forecast quality. In late April, there is even an additional peak of the BSS visible for outer anomaly bands that contain less than 10% of all initial anomalies each.

Therefore, while the overall forecast quality across the whole time window for which there is no certainty of the event is dismal, there are initial dates t_0 for which a significant improvement over the benchmark can be obtained.

4.3.3 Full probability prediction

Up to now, we have only looked at coarse-grained predictions, and in the binary case we only issued forecasts for initial dates in spring and not across the whole calendar year. Now, we will move on to predicting the probability of the next frost being in t days, but with an upper maximal first passage time of $t_{\max}$ days. This means that we will use $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T)$ directly, binned into daily probabilities with a single larger bin for $t_{\text{frost}} \geq t_{\max}$.

As a benchmark model, we first construct the histogram $p_t = P_{\text{data}}(t_{\text{frost}} = 1|t-1)$ from the data, using only the cases in which the initial temperature on day t is still above freezing. We then issue the probability distribution given by Eq.(4.4) as a forecast for the true distribution

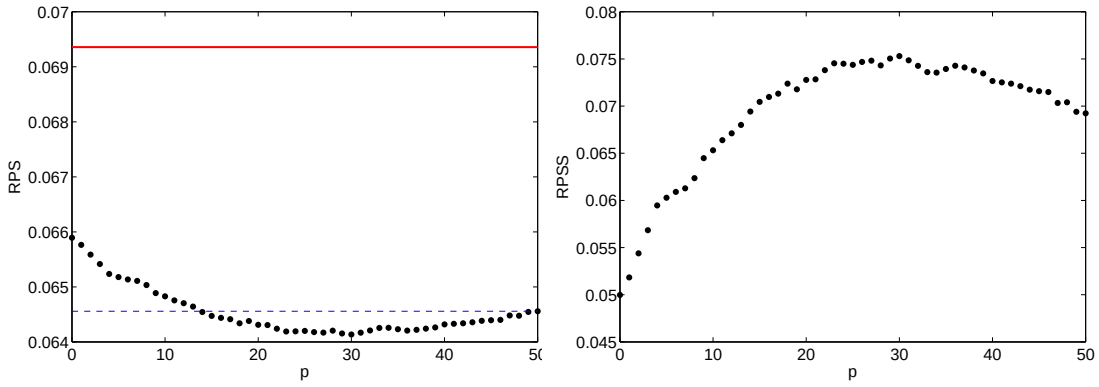


Figure 4.20: Ranked probability score (left panel) and ranked probability skill score (right panel) of the full probability distribution prediction for a lead time of up to 30 days using $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3w,p)})$. The red horizontal line shows the score of the benchmark forecast, the blue dashed line the score of the forecast using $\Delta T^{(10)}$.

of first passage times to frost, for fixed t_0 and $i = \{1, ..., t_{\max}\}$ days into the future:

$$\hat{p}(t_{\text{frost}} = i|t_0) = (1 - \hat{p}(t_{\text{frost}} = i - 1|t_0)) \cdot p_{t_0+i}, \tag{4.4}$$

with $p(t_{\text{frost}} = 0|t_0) = 0$ as we only consider initial temperatures above 0 °C, and $p_{t_0+t_{\max}} := 1$ so that the last bin contains the probability of $t_{\text{frost}} \geq t_{\max}$. This means that we only use information about the probability of frost on the next day to construct the benchmark forecast of the probability distribution of next frost, corresponding to a Markov-like approximation that neglects longer correlations but respects the seasonal change of the frost base rate.

As this can be considered a categorical prediction with ordered categories, it is best verified using the ranked probability score (RPS), as explained in more detail in Sec. 2.3.3. As before, we start by estimating both the benchmark histogram p_{t+1} and the conditional first passage time probabilities $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T)$ from the first 30 years of the data. We then forecast t_{frost} for the next year before using next year's data to first verify the forecasts and then reestimate the histograms, thus implementing a yearly update of the forecasting algorithm, leading to conservative estimates of forecast skill due to the reduced sample size[74].

As before, we coarse-grain the initial condition on the anomalies by using $\Delta T^{(3w,p)}$ and $\Delta T^{(10)}$, i.e. a separation into three bands, where the outer bands contain p% of all initial anomalies each, and a separation into deciles. The scores of the resulting forecasts for $t_{\max} = 30$ days can be seen in Fig. 4.20, where the left panel shows that all forecasts perform better than the benchmark (red line), but three weighted bands of 30%, 40% and 30% of all initial anomalies are best. The skill score is plotted in the right panel and shows that using our method, an improvement of up to 7.5% over the benchmark forecast is possible.[18]

Of course the forecast quality also depends on the maximal first passage time $t_{\max}$. Using the three initial anomaly bands that were optimal for $t_{\max} = 30$ days before, we analyse this dependence in Fig. 4.21. As can be seen, both the benchmark and the forecast show a very similar non-trivial dependence on the maximal first passage time with the forecast always beating the benchmark. As expected, the largest improvement over the benchmark with almost 12% is obtained for very short maximal first passgae times. There, persistence effects that are

[18] Note that the benchmark used for this forecast is much more refined than the one used for the deterministic forecasts, as it also uses first passage time properties albeit only one day in advance. Moreover, the RPS skill score with reference to the climatology has been considered a harsh standard similarly to the Brier skill score[90].

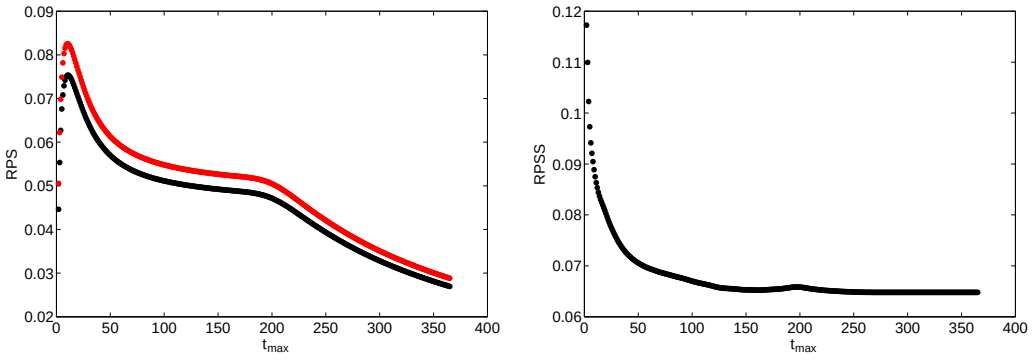


Figure 4.21: Ranked probability score (left panel) and ranked probability skill score (right panel) for the forecast using $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(3w,30)})$ for different maximal first passage times.

prevalent in weather conditions cause the initial temperature anomaly to have a large influence on the resulting first passage time to frost. For larger maximal first passage times, the skill score rapidly drops until it attains a constant value of $\approx 6.5\%$. This occurs when all forecasting skill is contained within the resolved time scale so any added resolution for the longest first passage times does not change anything.

Since the forecast quality for deterministic and binary forecasts heavily changed with the initial date and pooling forecasts might result in an overestimation of forecast skill[138, 76], we will also analyse this dependence for the full probabilistic forecast. As can be seen at first glance in Fig. 4.22, if the skill score is only coarsely colour-coded, the dependence on the size of the outer initial anomaly bands is negligible, resulting in the striped appearance of the contour plot. The initial date t_0 on the other hand has again a large influence on the forecast quality. Indeed, from May to the beginning of September both the forecast and the benchmark obtain perfect scores, as all probability for frost lies in the last bin, namely $t_{\mathrm{frost}} \geq t_{\mathrm{max}}$. In the first half of September and for isolated days in spring and winter, the forecast actually performs worse than the benchmark. This is due to the forecast already predicting frost with non-vanishing probability for days that are closer to t_0 than the maximal first passage time considered in the forecast, when it seldom occurs, while the benchmark still does not put any probability there. For the rest of the year, the forecast performs better than the benchmark, the improvement even exceeding 10% (areas marked in red) for a sizable number of initial dates.

4.4 Conclusion

After preparing and analysing the temperature data itself, this chapter investigated the possibilities of predicting the first passage time to frost exclusively from the time series.

In a first step we found that the conditional probability distribution of the next day of frost $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T)$ depends in a non-trivial way on the initial temperature anomalies ΔT. This means that the initial conditions influence the first passage time to frost and can be used for predictions. This influence, however, holds true only if the initial date t_0 lies in autumn or winter, i.e. close enough to the next occurrence of frost. Indeed, for the median of the distribution this dependence on ΔT was in evidence for t_0 between November 1$^{\mathrm{st}}$ and April 1$^{\mathrm{st}}$.

As we have a limited amount of data available from the measured temperature time series, we needed to coarse-grain the condition on the initial anomalies. By changing from exact to weighted terciles, where the two outer ones contain $p\%$ of all anomalies each while the middle

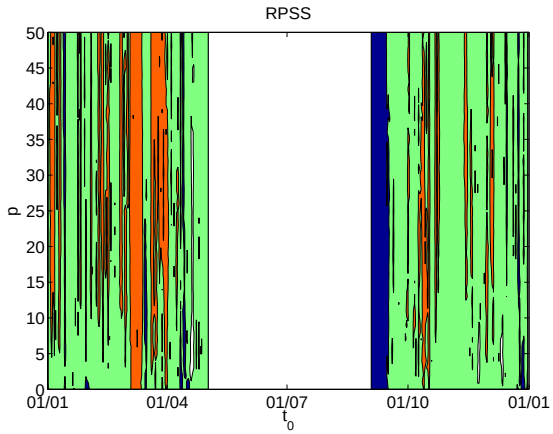


Figure 4.22: Contour plot of the ranked probability skill score using $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(3w,p)})$ with $t_{\max} = 30$ days. White denotes perfect forecast and benchmark, dark blue denotes forecasts worse than the benchmark, green are forecasts better than the benchmark by less than 10% and red forecasts that are better than the benchmark by more than 10%.

one is made up of $(100 - 2p)\%$, the differences in $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T)$ were enhanced. The improvement was, however, rather small.

Using an autoregressive model to simulate additional initial anomalies introduced systematic errors in the estimation of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T)$. This was therefore not a successful strategy to improve the statistics and thus allow for a finer conditioning on the initial anomalies. However, the discrepancies between model and measured time series shows that there are still long-term correlations in the anomaly time series. This further indicates that there is indeed potential for predictability.

In order to more conclusively analyse the significance of these predictability effects, we also used a Kolmogorov-Smirnov test. It confirmed that the differences in $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T)$ between the highest and lowest initial anomaly tercile are indeed statistically significant for t_0 in the winter half of the calendar year or t_0 in June.

Since the current operational weather models are useful only for up to 8 days into the future, we analysed our findings also for the case $t_{\mathrm{frost}} > t_0 + 8$. We found that some of the effects for initial dates t_0 in autumn are even significant in this case - an encouraging finding for the use of purely data-based prediction schemes.

To issue actual predictions of the next date of frost, we first needed to specify the forecast task needed. The most immediate task that we started with was a deterministic forecast of the exact date. A preliminary analysis showed that this date changes significantly with ΔT for $t_0 = 289$, i.e. October 15^{th}. Moving on to true forecasts and their out-of-sample verification, we found results that differed strongly depending on the verification aim. If one wants to be as close as possible on average to the next frost date, then the root mean square error (RMSE) showed that a forecast based on the mean of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T)$ always beats the benchmark, regardless of the precise coarse-graining scheme used for ΔT. However, the absolute quality of the forecast is not very good, as evidenced by a large RMSE. One might, however, rather want to hit the correct date as often as possible or one might need the forecast to be earlier than the actual date of first frost. For these forecast requirements, a forecast based upon the date on which in 10% of all cases in the training set frost had already occurred led to a nice proportion correct that again beat the benchmark.

Actual predictions of the date are most relevant for initial dates in autumn and perform worst for initial dates in spring. Then, when the probability distribution of first passage times is bimodal, it is already of interest to predict occurrence within the first or rather the second peak, i.e. before or after summer. A preliminary analysis showed that the probability of frost still occurring before summer changes significantly with the initial anomalies. The actual forecast, however, did not perform well across the whole time period when the first passage times are bimodal and was, overall, actually not significantly better than the benchmark. Looking at the initial dates t_0 individually, however, showed that between March 20^{th} and April 18^{th}, forecasts that significantly outperform the benchmark are obtained. To obtain the best forecasts, a coarse-graining of the initial anomalies as $\Delta T^{(3w,p)}$ with p diminishing with t_0 should be chosen.

As this prediction is only relevant for specific initial dates, we then moved on to a full forecast of the probability distribution of the next date of frost. Predicting this for fully resolved first passage times up to $t_{\max} = 30$ days using again estimates of $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T)$ beat the benchmark by around 7.5%. Analysing the seasonal changes in forecast quality, we saw that for initial dates in summer, both forecast and benchmark performed perfectly, while for a short time in early autumn the benchmark was actually better than the forecast. However, most of the time, the forecast outperformed the benchmark, in some cases even by more than 10%.

We have therefore seen that there are significant indications for predictability in the temperature anomaly time series. However, when looking at actual predictions, there were situations in which the forecasts did not perform well and were, in some cases, even beaten by the chosen benchmarks. Therefore, while the forecast quality was in general acceptable, there is still quite some room for improvement.

5 Improvement of the anomaly stationarity

5.1 Introduction

Chapter 4 has shown that while statistically significant predictability effects were found in the temperature anomaly time series, the quality of the predictions themselves still leaves room for improvement. Some of the problems might be due to the efforts needed to obtain enough data to adequately estimate the full conditional probability distribution $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T)$. Indeed, we did not only coarse-grain the condition on the initial anomalies ΔT by grouping them into three or at most ten separate categories, we also coarse-grained the condition on the initial date t_0. By using all anomalies recorded in a time window of $[t_0 - 45, t_0 + 45]$ and treating every initial year equally, we effectively assumed stationarity of the time series not only over a period of three months but also from one year to the next for the 118 years in total. In view of the climate change debate and also the seasonal change of the circulation patterns, this assumption is, as stated before in Sec. 3.3.1, rather daring.

Indeed, there is a positive trend contained in the Potsdam morning temperature time series. It was carried over into the anomalies (see Sec. 3.3.1), since the climatology only took out the mean and the yearly and twice-yearly frequency components but no polynomial contributions. The trend magnitude of around 0.01 K/year as determined in Sec. 3.3.1 is small enough to be considered negligible on the time scales of less than a year considered in the first passage time distributions in Chapter 4. However, it does affect which initial anomalies are selected in the conditioning on different anomaly bands and it therefore influences the whole prediction procedure (see also [24]). Correcting the temperature time series for the trend would remove these effects and collapse the anomaly distributions for different years that were previously slightly shifted in mean. The resulting distribution is therefore less broad and could lead to more precise results.

A trend also causes measurements at different times to be weakly related to each other, leading to long-range correlation effects as found in Sec. 3.3.2. These might not only overshadow any other correlations in the underlying dynamics, but also add to the model inadequacy of the AR(1)-process (see Sec. 3.4). Correcting the time series for the trend might therefore also better permit enhancing the statistics by modeling these new temperature anomalies with an AR(1) model process. A trend correction could therefore improve the prediction task outcome for several distinct reasons.

Apart from a trend, there was another stationarity problem revealed in Sec. 3.3.1: the seasonality of the variance and of the shape of the anomaly distribution. This also affects the selection of anomalies into bands for the prediction procedure - in this case across different calendar months and not different years - thus also contributing potential artificial effects.

We will therefore in the following attempt to use a more involved procedure to construct a new anomaly time series where, in addition to the seasonal cycle of the mean temperature as previously, both the trend and the seasonality of the variance are removed.

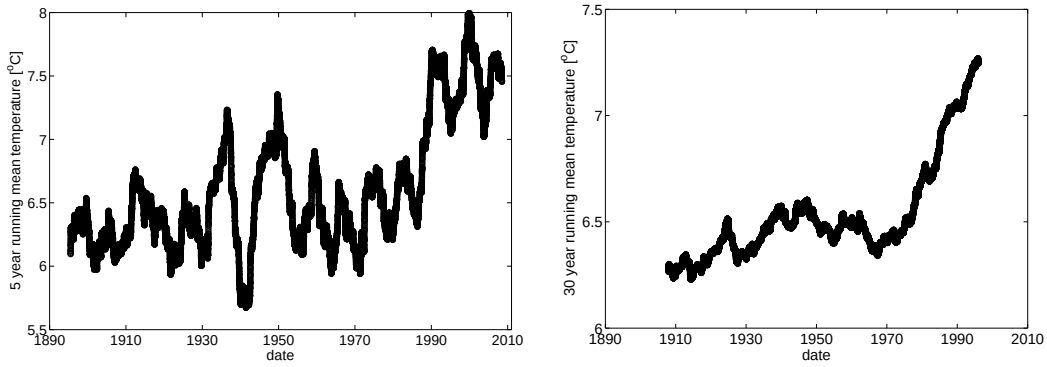


Figure 5.1: Mean Potsdam temperature over a running window of 5 years' width (left panel), as well as 30 years' width (right panel).

5.2 Properties of the corrected anomaly time series

5.2.1 Construction of the new anomaly time series

We start by removing the temperature trend. To do this, we first simply calculate the running average temperature over a time window of 5 years. This mean temperature is shown in Fig. 5.1. There exist, of course, much more accurate and involved methods for determining actual trends in the data (see also Sec. 3.3.1), especially in the presence of long-term correlations. However, our goal is not to separate out any effects rather arising from oscillations on very large timescales such as the longer cycles of the North Atlantic Oscillation (NAO) or even solar cycles. We are not interested in the purest trend value possible but simply want to insure stationarity of our resulting anomaly time series. Therefore the running average should be sufficient for our purposes, but we will analyse the resulting time series in terms of stationarity in Sec. 5.2.2.

An interesting feature of this trend can be seen in the more coarse-grained calculation in the right panel of Fig. 5.1: While the mean temperatures overall increase, there exists a period between 1940 and 1970 where the trend appears to vanish and might even be slightly negative. This general time dependence of the mean temperatures confirms previous findings for global mean temperature anomalies gleaned from many data sets[139].

Having thus obtained an estimate of the temperature trend, we subtract it from the data to obtain the detrended time series T^{detr}. We then proceed with determining the climatology $\widetilde{T^{\mathrm{detr}}}$: As described in Sec. 3.2.2, we first calculate the mean detrended temperature value for each calendar day. Then we compute a smooth sinusoidal model using the yearly and twice-yearly frequencies as given previously in Eq.(3.1). Subtracting this climatology from the detrended temperatures, we obtain a preliminary anomaly time series.

Finally, we still need to divide out the seasonal variance as described in Eq. (5.1), where $\hat{\sigma}_i$ denotes the standard deviation of the preliminary anomalies recorded on calendar day i. This hopefully makes the resulting final anomaly time series not only stationary across different years, but also across different calendar months.

$$\Delta T_i = \frac{T_i^{\mathrm{detr}} - \widetilde{T_i^{\mathrm{detr}}}}{\hat{\sigma}_i} \tag{5.1}$$

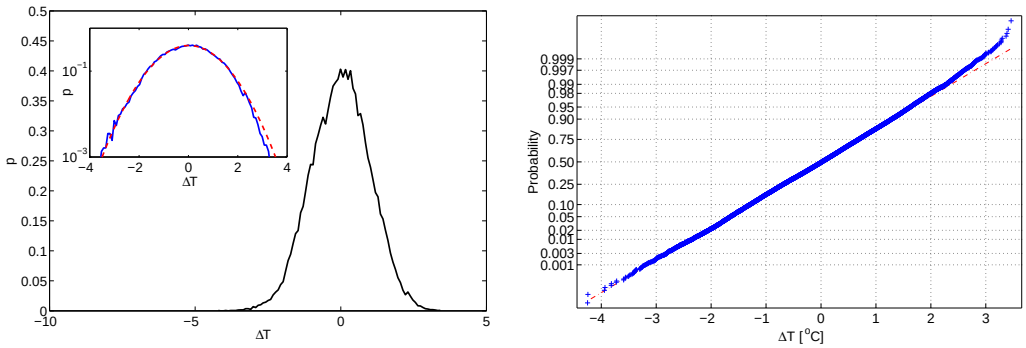


Figure 5.2: Probability distribution of the anomalies generated with additional detrending and deseasoning of the variance, plotted both directly and on a semilogarithmic scale (left panel), and in a normal probability plot (right panel). The red dashed lines in both panels represent the corresponding figures for an exact normal distribution.

5.2.2 Stationarity issues and correlations

As there is no guarantee of true stationarity even after using this more involved way of determining the temperature anomalies, we will need to analyse the resulting time series carefully again.

Fig. 5.2 shows that the probability distribution of the new anomalies can be considered normal[1]. There are only very slight deviations from the expected shape for a normal distribution as marked by the red dashed lines, and they occur for very large anomaly values. The undersampling of these positive extremes might simply be due to the finiteness of the recorded time series. Comparing this with Fig. A.12 shows that the heavier tail for the negative extremes has vanished with this redefinition of the anomaly time series.

Another issue revealed in Sec. 3.3 was the change in anomaly distribution when conditioning on different calendar months. The left panel of Fig. 5.3 confirms the findings of Fig. A.5, namely that the deseasoning of the variance mostly remedies this seasonality. However, it also shows that the distribution still shifts in skewness over the calendar months so not all seasonal changes have been eliminated. The right panel also shows that while there are still some differences in the mean anomaly across the calendar months, they have diminished considerably when compared with the earlier definition of the anomalies as used in Chapters 3 and 4. So while some issues with the stationarity of the probability distribution remain, they are not as glaring as before.

However, the difference in skewness of the distribution still shifts the anomaly range in the course of a calendar year. While this only affects the more extreme anomalies, it still gives more weight to some months than to others when selecting initial anomaly bands. Therefore, even this more involved way of constructing the anomaly time series did not manage to render the selection procedure free from seasonal influences. However, the changes in anomaly range are rather slow and smooth so this effect should be less important for the new anomaly time series than for the original one.

Another concern with the anomaly selection procedure for the conditioning was the trend in the data. Fig. 5.4 shows in the right panel that the fraction of anomalies in each decile of the distribution that is coming from the first half of the time series systematically depended on the percentile of the distribution. After detrending, it now fluctuates around the expected value of 50%, as can be seen in the left panel. The trend in the temperature data is therefore not a concern anymore.

[1]For an explanation of the normal probability plot see Sec. 2.2.4

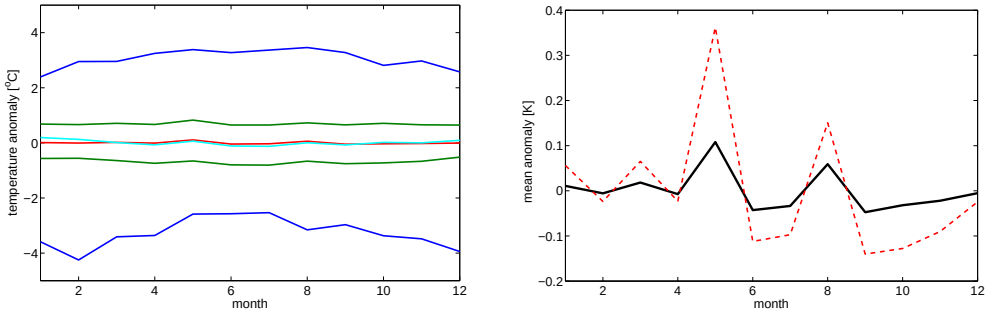


Figure 5.3: Seasonality of the anomaly probability distribution across the calendar months. The left panel shows its mean (magenta), median (cyan), quartiles (green) and extrema (blue) for each calendar month. The right panel shows only the monthly mean (black line), as well as the corresponding values for the previous anomaly time series (red dashed line).

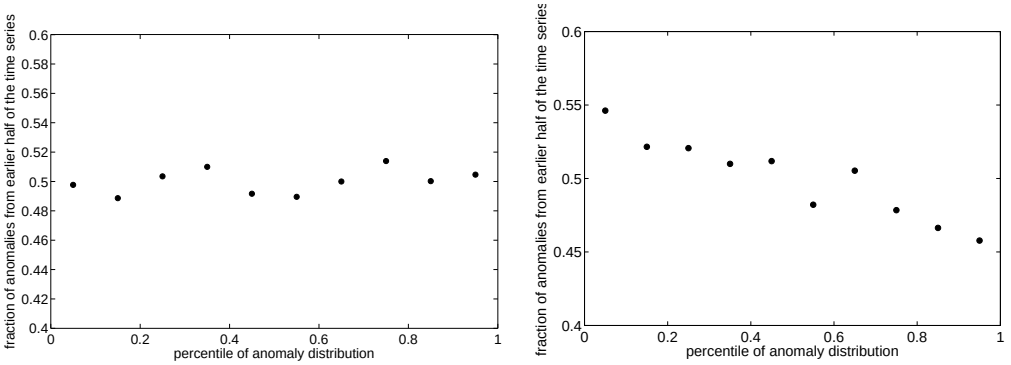


Figure 5.4: Fraction of anomalies in each decile that was recorded in the earlier half of the time series, i.e. between 1893 and 1951. The left panel shows the result for the time series with improved stationarity, the right panel for the original version.

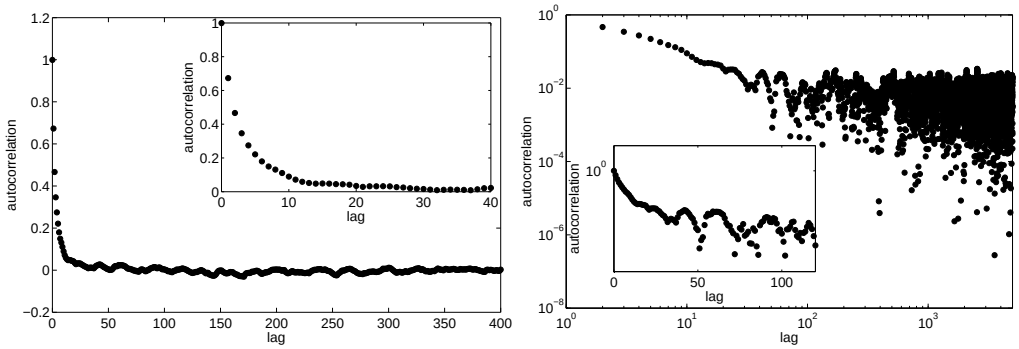


Figure 5.5: Autocorrelation function of the anomaly time series with improved stationarity for two different zoom depths (left panel), as well as on a double-logarithmic scale (right panel) and a semilogarithmic scale (right panel, inset).

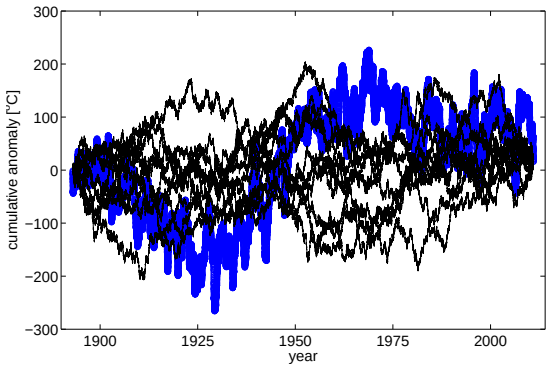


Figure 5.6: Cumulative anomalies as defined in Eq. (3.3) for the anomaly time series with improved stationarity (blue line) and ten different surrogate realisations (black lines).

With the detrending, one aspect of the temperature time series that was leading to the appearance of long-range correlations in the data has been removed. This will have changed the autocorrelation structure of the anomaly time series. If only short-range correlations remained, we would expect an exponential decay of the autocorrelation function. Fig. 5.5 shows, however, that the autocorrelation function looks remarkably similar to that obtained using the previous anomaly definition (shown in Fig. 3.13). Indeed, the right panel confirms that the small lags can be approximated by an exponential decay, but that there is a crossover to a much slower decay. Therefore, removing the temperature trend did not remove all long-term correlations in the data.

To better visualise these correlations, Fig. 5.6 shows again the fluctuation landscape of these new anomalies, as defined before in Eq. (3.3). Contrary to the previous anomaly time series, whose fluctuation landscape was shown in Fig. 3.14, the deviations from the zero line are much smaller in this case and their amplitude is not distinguishable from that of the surrogate realisations. It has, however, a systematic-looking shape, leading to the conclusion that while the true long-range correlations were removed with the detrending, there are remaining correlations on an intermediate time scale.

It has been found previously that the autocorrelation structure in temperatures is seasonal, with a greater persistence in summer than in winter[134]. Now that this effect in the correlations cannot be overshadowed by the temperature trend and the seasonality of the variance any longer, we analysed whether this seasonality of correlations was carried over into the anomaly time series.

Fig. 5.7 shows that the autocorrelation structure of the winter anomalies indeed decays less rapidly than that for all the other seasons, confirming longer persistence in winter. The other three seasons, however, have a rather similar autocorrelation structure. While the stationarity of the anomalies has been improved across the calendar months, this result shows that the time series is not stationary enough to further coarse-grain the conditioning on t_0 in the estimation of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0)$.

5.2.3 Modeling with an AR process

The removal of the long-range correlations means that any low-order autoregressive model should be more adequate for this new anomaly time series than for the previous version. This is further supported by the truly normal distribution of the anomalies as seen in Fig. 5.2. Testing the optimal model order as described in Sec. 2.1.4 leads to the conclusion that most criteria again agree on an optimal order of $p = 8$ and also still show the large improvement already

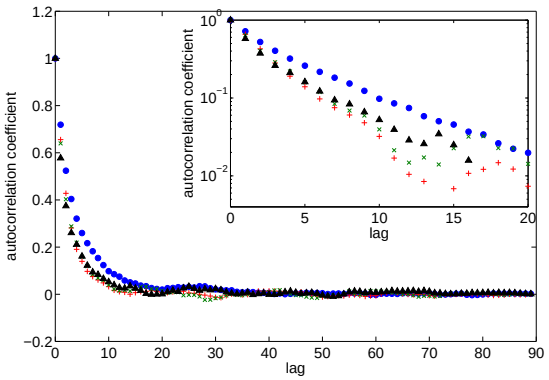


Figure 5.7: Autocorrelation coefficients for winter (blue dots), spring (red plus signs), summer (green crosses), and autumn anomalies (black triangles), averaged over the 118 years in the anomaly time series. The inset shows a zoom into small lags on a semilogarithmic scale.

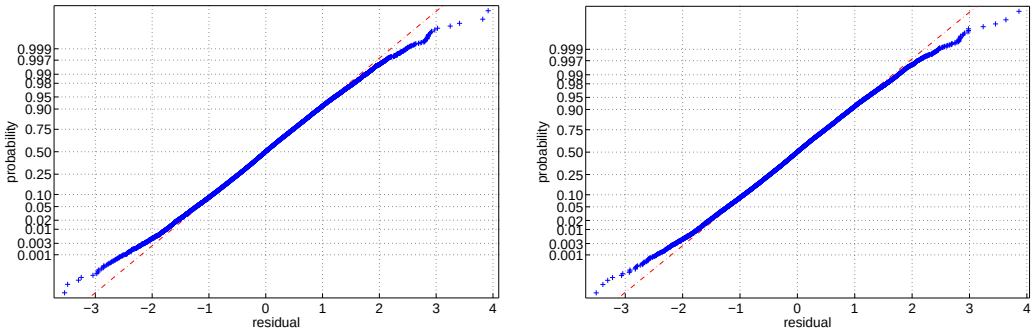


Figure 5.8: Normal probability plots of the residuals after fitting the anomaly time series with an AR(1)-process (left panel) and an AR(8)-process (right panel).

for $p = 1$ with much smaller corrections for the higher model orders. However, the improved criterion as defined by Schwarz yields an optimal order of $p = 5$ and the reflection coefficients of the partial autocorrelation function have equal optimal results for $p = 10$ and $p = 20$.

Staying with the previous results of $p = 1$ and $p = 8$, we calculate the residual probability distributions after subtracting the autoregressive models from the anomalies. As shown through their normal probability plots in Fig. 5.8 (see also Sec. 2.2.4), both the AR(1) and the AR(8) process lead to a normal residual distribution with slight deviations for the more extreme anomalies, as was to be expected. There is no difference in model fit quality detectable directly from the residual distribution.

However, trying to determine the capacity of the two models to fit the correlation structure of the anomaly time series, significant differences are visible. Fig. 5.9 shows that while the AR(1)-process is rejected consistently, the AR(8)-process can capture the correlation structure of the anomalies to some extent: While the p value of the observed Ljung-Box test statistic again declines rapidly beyond lag 20 (compare to Fig. 3.18), it stays above any probability usually taken as low enough to reject the null hypothesis of equal correlation structure. The AR(8)-process is therefore a better model for the temperature anomalies than the AR(1)-process.

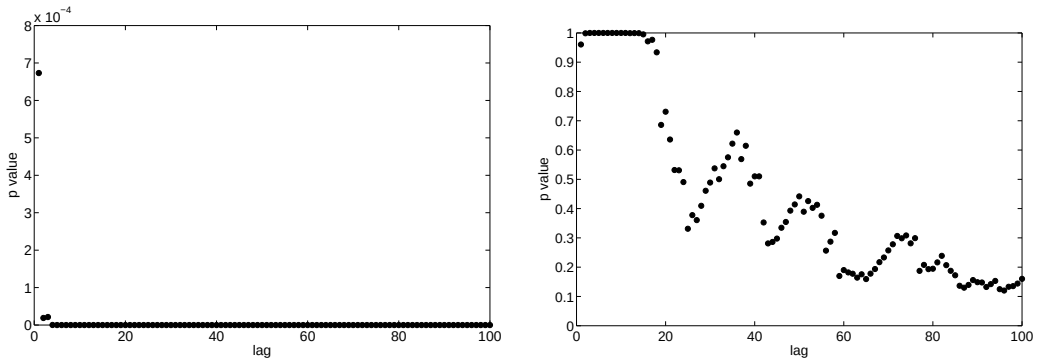


Figure 5.9: P value from the Ljung-Box test statistic for the anomaly time series with improved stationarity fitted with an AR(1) process (left panel), and an AR(8)-process (right panel).

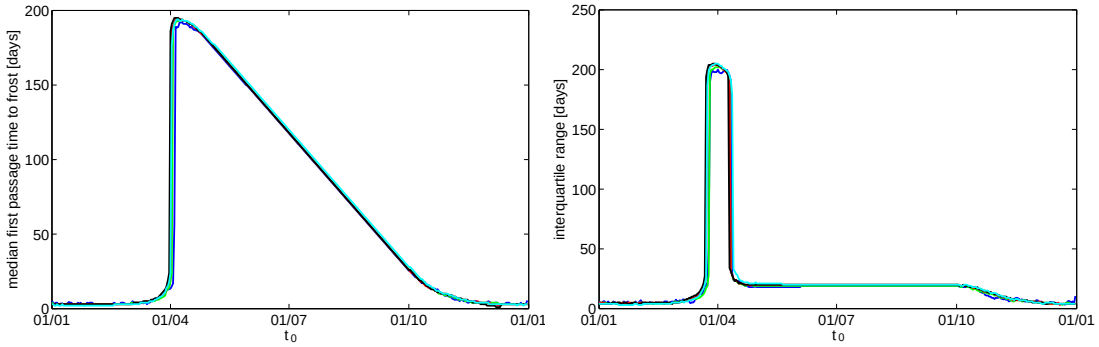


Figure 5.10: Median and interquartile range of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, y = 1965)$ for different window widths w_t of 31 days (green), 61 days (red) and 91 days (black) around t_0, as well as using all available anomalies (cyan) and only those recorded on t_0 itself (blue).

5.2.4 First passage time properties

We will now turn to the properties of the first passage time distributions to frost $P(t_{\mathrm{frost}}|t_0, y)$ as in Sec. 3.5. In this case, we again consider all anomalies ΔT_j with $|t_0 - j| \leq \lfloor w_t/2 \rfloor$ as possible starting values. The additional detrending and deseasoning means that now, for each such j which also satisfies $\hat{\sigma}_{t_0} \cdot \Delta T_j + T^{\mathrm{trend}}_{t_0} + \widetilde{T^{\mathrm{detr}}_{t_0}} > 0\,°\mathrm{C}$, the first passage time to frost t_{frost} is defined as the smallest number of days k, $k \in \mathbb{N}$, for which $\hat{\sigma}_{t_0+k} \cdot \Delta T_{j+k} + T^{\mathrm{trend}}_{y,t_0+k} + \widetilde{T^{\mathrm{detr}}_{t_0+k}} \leq 0\,°\mathrm{C}$. Since the trend T^{trend} differs between calendar years, the distribution does not only depend on the specific calendar day t_0, but also on the calendar year y. For the start, we will do our analysis with $y = 1965$ fixed, i.e. with a mean value corresponding to the global mean temperature across the whole time series, before analysing the influence of different initial years.

Fig. 5.10 shows that both the median and the interquartile range of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, y = 1965)$ are independent of the anomaly window width w_t. Contrary to before, even choosing the anomalies entirely independently from the date they were recorded on does not change the location or spread of the resulting first passage time distribution from those obtained from anomalies only recorded on t_0 itself. The stationarity of the anomaly time series has therefore been improved drastically, aiding in recovering better statistics than before.

As seen in Sec. 5.2.3, this increased stationarity has also led to an increased goodness of

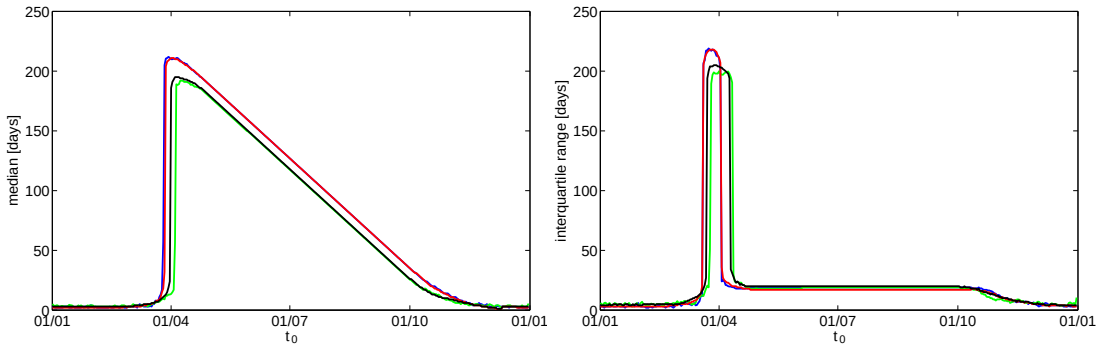


Figure 5.11: Median and interquartile range of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, y = 1965)$ and $P_{\mathrm{AR(8)}}(t_{\mathrm{frost}}|t_0, y = 1965)$ for two different window widths w_t (1 day and 91 days).

fit of the AR(8) process to the anomalies. However, Fig. 5.11 shows that while the location and spread of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, y = 1965)$ and $P_{\mathrm{AR(8)}}(t_{\mathrm{frost}}|t_0, y = 1965)$ are very similar, some essential differences remain. Indeed, the model process seems to show a slightly earlier jump to first passage times after the summer than the data and then also a slightly later date with maximum probability of first frost in autumn. Even though the goodness of fit was certainly encouraging, these differences show that the difficulties using the very simple model to generate more anomalies for better statistics remain.

Fig. 5.12 shows the full first passage time distributions for both the AR(8) model and the data using four representative dates. It can be seen that while the date of maximum probability can be reproduced by the model process except for initial dates in summer, the relative weight of the two peaks in the bimodal case is represented wrongly. This directly explains the difference in the precise date on which the shift to a greater weight at later dates occurs. Therefore, the model introduces systematic errors that still make it inherently unsuitable for the prediction task even with this new anomaly time series.

Interestingly, one cannot help but notice that the distribution estimates from the data and the model process shown in Fig. 5.12 seem to match even worse than those obtained with the less elaborate anomaly definition match those obtained from the statistically less suited and simpler AR(1) model process (see Fig. 3.25).

Up to now, we have always considered $y = 1965$, when the mean temperature was closest to the global average. However, this would not be very representative for other years, when the mean temperature differs from the global average by more than 1 °C. In the following, we shall therefore analyse the influence of the initial year on the first passage time probability to frost.

As can be seen from Fig. 5.13, there is, as expected, a marked influence of different mean values on $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, y)$. The spread of the first passage time distribution is only significantly affected in spring when the mean temperature determines the exact date at which the later peak for frost after summer acquires the larger weight. The location, however, depends on y almost across the whole calendar year, with a first passage time difference outside of spring of as much as three weeks.

This can also be seen in the full first passage time distributions shown in Fig. 5.14. There the differences are slight for t_0 in late winter but appear quite remarkable when looking at initial dates in spring, going so far as to determine whether frost before or after summer is more likely. In summer, the mean first passage time is also shifted by about two weeks depending on y. The precise choice of the mean temperature value therefore greatly influences the first passage time distributions.

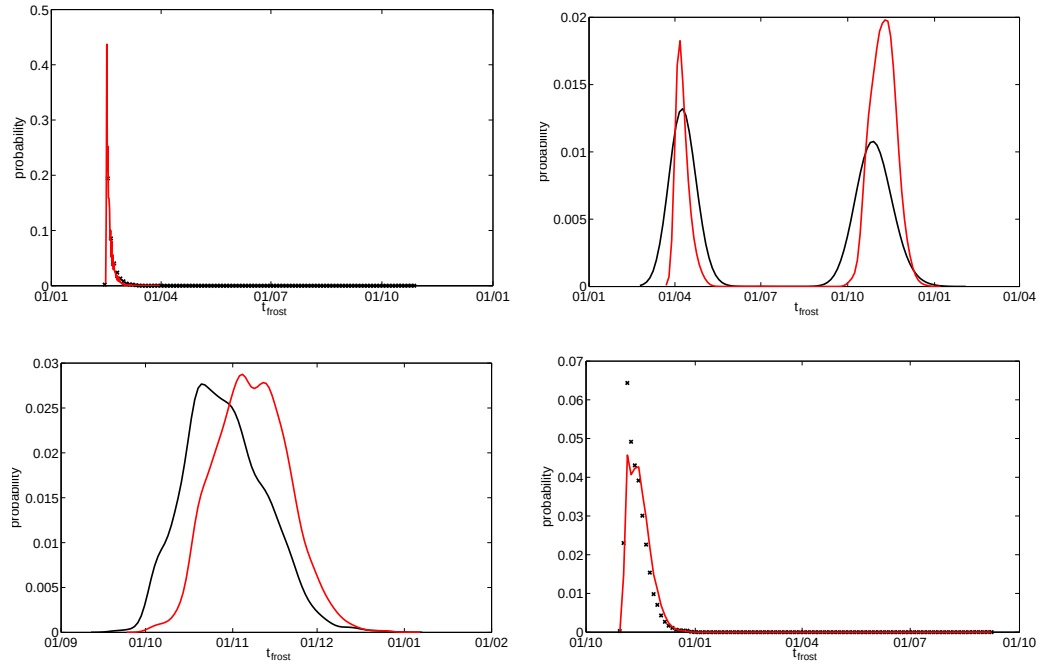


Figure 5.12: Kernel density estimates of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, y = 1965)$ in black and $P_{\mathrm{AR(8)}}(t_{\mathrm{frost}}|t_0, y = 1965)$ in red for initial dates t_0 of February 14$^{\mathrm{th}}$ (top left panel), April 1$^{\mathrm{st}}$ (top right), July 1$^{\mathrm{st}}$ (bottom left) and November 1$^{\mathrm{st}}$ (bottom right).

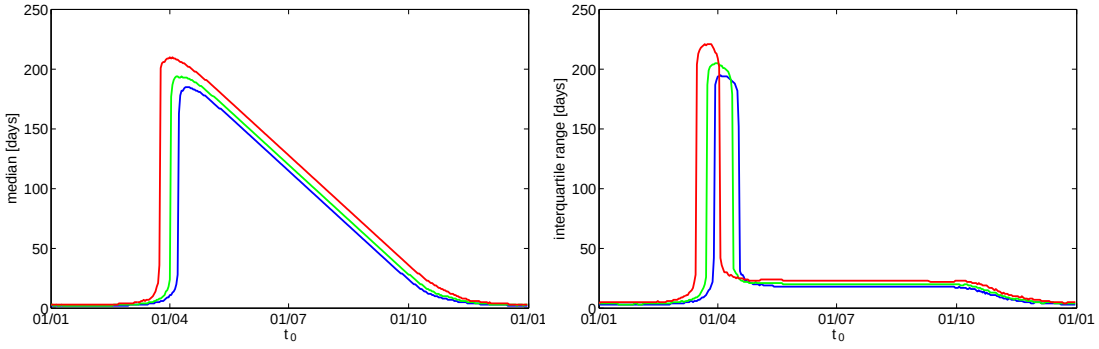


Figure 5.13: Median and interquartile range of an estimate of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, y)$ where all anomalies were considered for each value of t_0. The initial year was chosen as that with the minimal value ($y = 1941$, in black), the global average value ($y = 1965$, in red) and the maximal value ($y = 1999$, in green) of the 5 year running mean temperature.

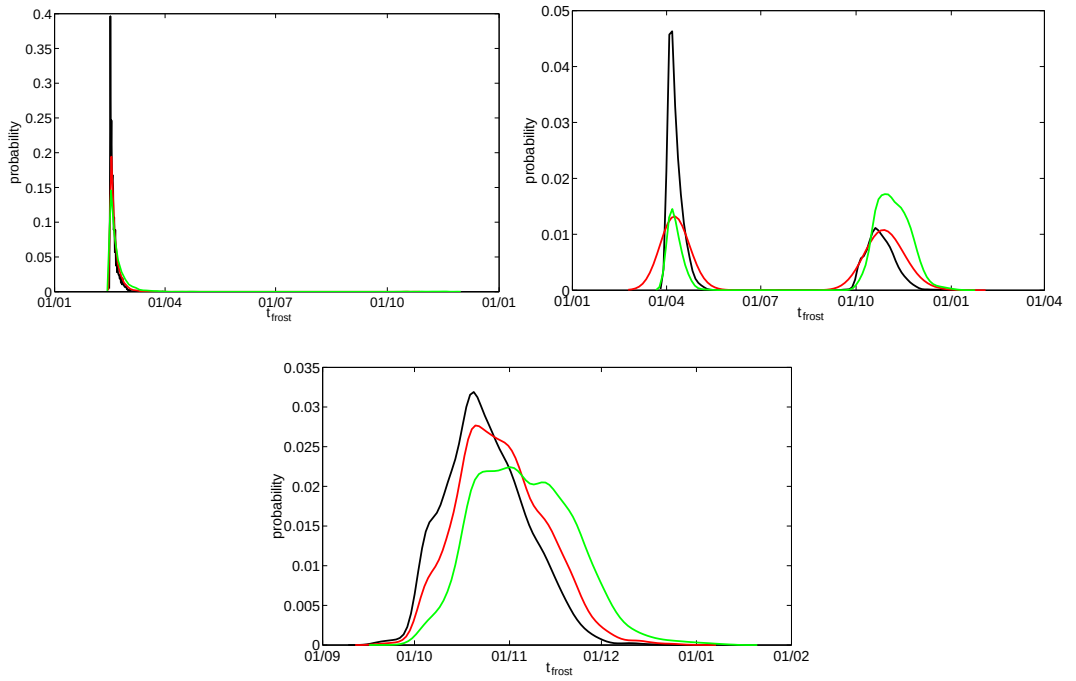


Figure 5.14: Kernel density estimates of the first passage time distributions starting on February 14$^{\text{th}}$ (upper left panel), April 1$^{\text{st}}$ (upper right panel) and July 1$^{\text{st}}$ (bottom panel) using an anomaly window w_t of 91 days around t_0. The initial year was chosen for minimal (black), average (red) and maximal (green) mean temperature.

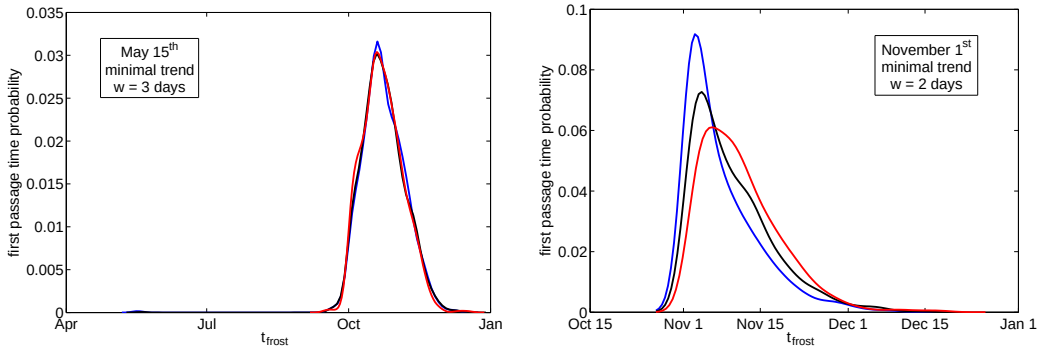


Figure 5.15: Kernel density estimates of $P_{\text{data}}(t_{\text{frost}}|t_0, y = 1941, \Delta T^{(3)})$ for two different values of t_0, showing the lowest anomaly tercile in blue, average anomalies in black and the highest anomaly tercile in red. The distribution estimates were obtained by using normal kernels of width w.

5.3 Predictability analysis

5.3.1 Change of first passage time distributions under conditioning

Moving on to the first passage time distribution to frost conditioned also on the initial anomaly terciles[2], we can see that there are again some predictability effects visible in the differences between the respective probability distributions. While the full spread of figures is displayed in Appendix B.5 for the three different initial years representing the minimum, average and maximum mean temperature, Fig. 5.15 contains two examples.

The left panel shows the case of an initial date of 15 May 1941, when different initial anomalies do not seem to have any influence on the resulting first passage time distribution. In the right panel, the opposite can be seen for an initial date of 1 November 1941, when the date of maximal first frost probability shifts by a week between the cases of low and high initial anomalies.

When comparing the effects to the results of the previous analysis for the time series without improved stationarity shown in Appendix B.2, it becomes evident that they do not seem to have changed markedly. The initial date of May 15$^{\text{th}}$ shows the largest difference, as there are still some very small first passage times present in the distribution, forming a long left tail that was not there in the original time series.

The results from the AR(8) model process are also depicted fully in Appendix B.5. Comparing them to those generated directly from the measured temperature time series, one can immediately see that for an initial date of February 14$^{\text{th}}$, the long and vanishingly small tails that extend until after the summer are missing in the model. This means that first frost does not occur nearly as late anymore in the model as in some instances in the data. The figures also hint at some other differences in spread and location between the corresponding first passage time distributions, but this is less immediately apparent.

We therefore again looked at a direct comparison between the first passage time distribution estimates from the data and the model for equal conditions. Fig. 5.16 shows two representative examples, illustrating in the left panel that the model can still lead to a reversal in the relative importance of the two peaks in the case of a bimodal distribution when compared to the data. For initial dates in the summer, as seen in the right panel, the AR(8) process even consistently overestimates the mean first passage time to frost.

Table 5.1 shows an overview of representative examples where the model-based estimate of the conditional first passage time probability differs significantly from the measured data. As can be seen, simulating temperature anomalies using the AR(8) model introduces systematic errors in the first passage time that would harm any prediction effort. This reflects the seasonality still present for instance in the skewness of the anomalies which acts as a correlation on longer time scales. Also the seasonality in the autocorrelations introduces a seasonality in the goodness of fit of the AR parameters. Even though the model fit was good in the more general tests in Sec. 5.2.3, using it to enhance the statistics even for this more stationary version of the anomaly time series remains problematic.

5.3.2 Change of distribution summary measures under conditioning

Continuing the model fit analysis, Fig. 5.17 shows a more systematic comparison for initial dates across the calendar year, again using an initial year of average temperatures. As can be seen, the median of the first passage time distribution lies about ten days later in the model than in the

[2] Due to the seasonality of the correlations and skewness of the anomalies, we have again chosen to use 3 month windows around the initial dates, even though Fig. 5.10 has shown that all anomalies might be used for this time series. However, we found that using the whole anomaly time series does not improve or in any way change the results of the predictability analysis.

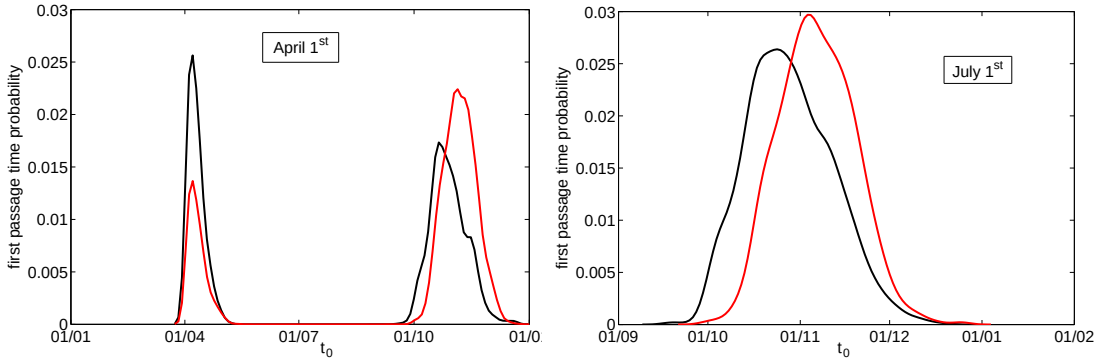


Figure 5.16: Conditional first passage time probability distribution for 1 April 1965 with the upper anomaly tercile and 1 July 1965 with the lower anomaly tercile estimated from the measured data (black) and the AR(8) model process (red).

t_0	y	Problem with the model fit
February 14$^{\text{th}}$	1941 and 1965	slight probability of frost after summer only in the data
May 15$^{\text{th}}$	1999	slight probability of frost still before summer only in the data
April 1$^{\text{st}}$	1941 and 1965	difference in the relative peak weight up to a reversal
summer	1941 and 1965	mean first passage time consistently later in the model

Table 5.1: Illustrative examples of parameters for which $P_{\text{data}}(t_{\text{frost}}|t_0, y, \Delta T^{(3)})$ and $P_{\text{AR(8)}}(t_{\text{frost}}|t_0, y, \Delta T^{(3)})$ differ significantly.

data for initial dates between the beginning of May and mid November.[3] Also, the difference in median between the lower initial anomaly tercile and the other initial conditions that is visible in autumn starts around two weeks later in the model than in the data. These results are also observed when using an initial year with minimal mean temperature; for an initial year with maximal mean temperature, the difference in median between the lower anomaly tercile and the others even starts almost six weeks later in the model than in the data (not shown here).

The model also underestimates the standard deviation of the first passage time distribution for initial dates between May and mid October and from mid December to mid January as shown in Fig. 5.17. A difference in the standard deviation for the upper initial anomaly tercile when compared to the other initial conditions is visible in the data in late spring to summer (at least for years with average and maximum mean temperatures), but not in evidence in the model. However, the model shows a difference in standard deviation of the lower initial anomaly band in autumn (for an initial year with minimum and average mean temperature) that is not in evidence in the data.

Lastly, the bimodality in the first passage time lasts around two weeks longer in the data than in the model and the peak of first frost after summer takes also place later in the model than in the data. This is in nice agreement with the later median observed in the model before.

After this analysis, we have seen that the model not only misrepresents the precise timing of first frost events, but also the existence of predictability effects for specific initial conditions. Therefore, the model truly is not representing the data well even when the stationarity of the anomaly time series is significantly improved. This is a very interesting finding indeed: In the original anomaly time series, much of the model failure could be attributed more to the lack of stationarity in the anomaly data than to any long-range correlations in the data that would

[3]The bimodal distributions were again excluded when calculating the median and standard deviation.

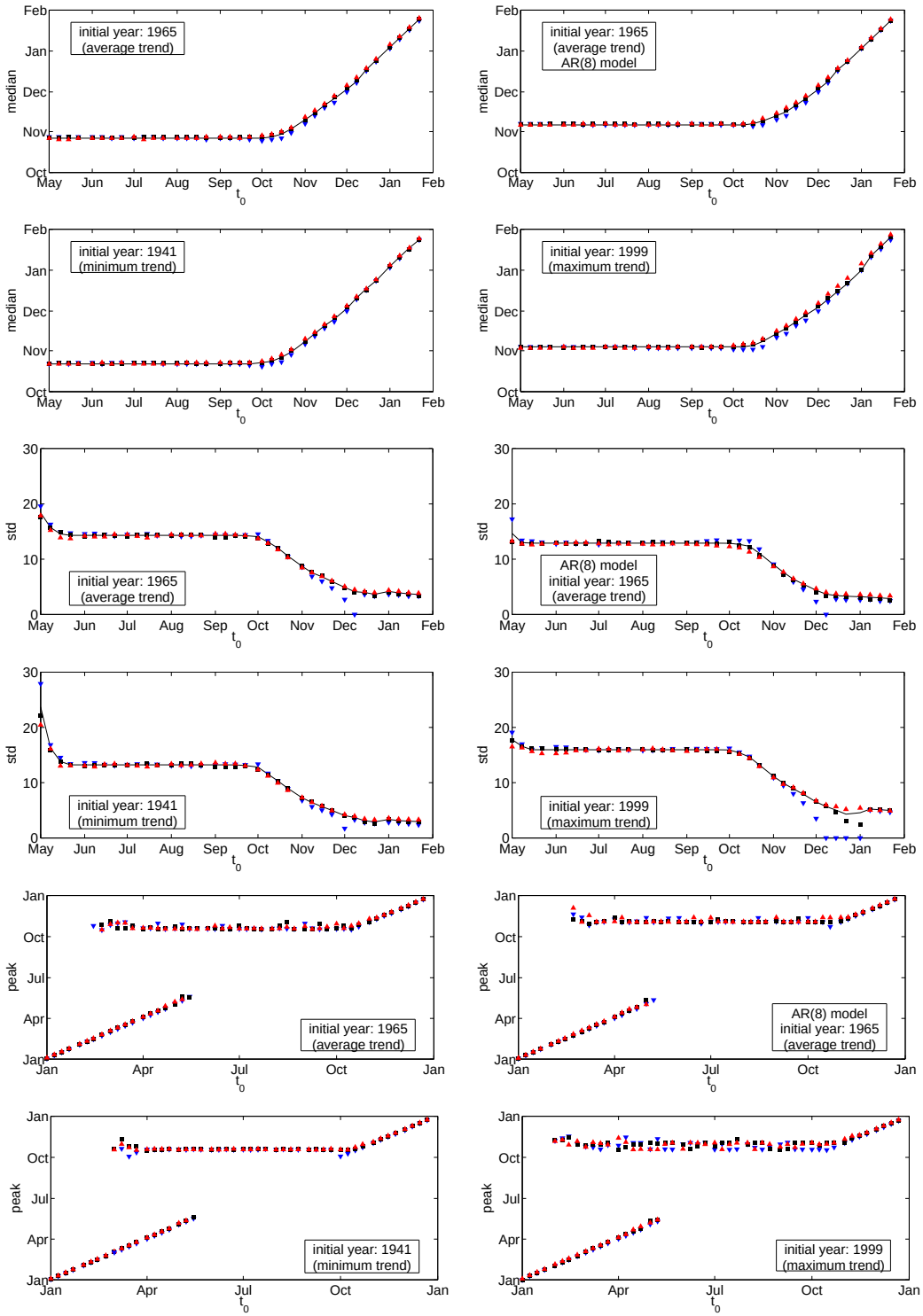


Figure 5.17: Median, standard deviation (in days) and location of the peaks of maximum probability of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, y, \Delta T^{(3)})$. The lower initial anomaly tercile is depicted using blue downward triangles, the middle with black squares and the upper with red upward triangles. The continuous lines represent $P(t_{\mathrm{frost}}|t_0, y)$. In three cases, $P_{\mathrm{AR}(8)}(t_{\mathrm{frost}}|t_0, y, \Delta T^{(3)})$ is shown instead (as noted directly on the figures).

allow for true predictability. However, the lack of model suitability actually persists here, even though the stationarity is much improved and the AR(8) model is actually not rejected over a large range of lags anymore (see Sec. 5.2). This suggests that there are indeed long-range correlations in the data.

Having rejected the model, we will again turn to a more thorough analysis of the predictability effects visible across the calendar year for the three different initial years considered up to now. Fig. 5.17 shows that when considering the median of the first passage time distribution, only for initial dates from October to mid December any differences between the initial anomaly terciles are in evidence. These effects are slightly more pronounced for the initial years with average or maximum mean temperature. However, these effects appear about a week later than for the previous time series analysed in Fig. 4.3. Except for some initial dates in mid December of a year with maximum mean temperature, they are also smaller than before.

For the standard deviation, some predictability effects are visible for initial dates in May and June, especially for an initial year with maximum mean temperature. This was not the case for the previous version of the anomaly time series. The predictability effects occurring in December and January do also seem slightly larger than before for an initial year with maximal mean temperature, but not changed in the other cases.

In the peaks of maximum probability, as before, no predictability effects are visible.

Moving from a separation into terciles to the dependence of the first passage time distribution on initial anomaly deciles, we again look for predictability effects for the previously considered eight different initial dates and three different initial years. An overview of all relevant figures can be found in Appendix B.6.

The median of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, y, \Delta T^{(10)})$ shows a strong systematic dependence on $\Delta T^{(10)}$ between October and April that slowly diminishes with t_0. For April 1$^{\mathrm{st}}$, the mean[4] shifts by as much as 50 days depending on the initial anomaly decile. This reflects the changing weight of the two peaks of the distribution. These effects are most pronounced for $y = 1999$ and least pronounced for $y = 1941$.

For the spread, the effects are less clear. Only for t_0 in winter a significant dependence on $\Delta T^{(10)}$ is visible with a notable exception for December 1$^{\mathrm{st}}$ when the spread drops to zero, i.e. certain frost on the next day. The magnitude of these effects does not appear to change consistently with the initial year. For the initial date of April 1$^{\mathrm{st}}$, the influence of the different initial years is, however, very large: Here, it determines whether the standard deviation increases or decreases with the initial anomaly decile or is indeed independent from it. This can be understood when thinking about the origin of the large spread: If the initial conditions contain both a high initial mean temperature due to $y = 1999$ and high initial anomalies, then the peak of early frost in the bimodal case has very low weight so that its contribution to the standard deviation decreases, resulting in a lower overall standard deviation. The opposite case of low initial mean temperature due to $y = 1941$ and low initial anomalies results in a corresponding decrease of the standard deviation due to a highly suppressed later peak of next frost only after summer. For $y = 1999$, the standard deviation therefore decreases with ΔT while for $y = 1941$, it increases.

When looking at the peak of maximum first frost probability, a dependence on the initial anomaly is only visible for initial dates during winter as before and then slightly more so for an initial year with maximal mean temperature than for the other cases.

Comparing these results more directly to those gathered from the original time series as detailed in Sec. 4.2.2 and Appendix B.3, we can better assess whether the more involved way of constructing anomalies from the temperature measurements constitutes an improvement.

[4]The mean was again chosen as the more robust measure for disconnected bimodal distributions.

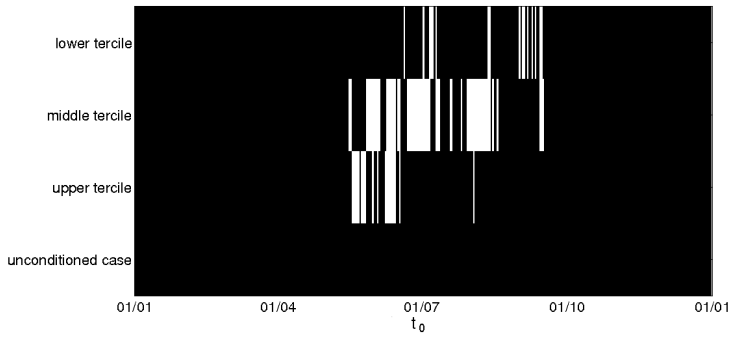


Figure 5.18: Results of a Jarque-Bera hypothesis test for $P_{\text{data}}(t_{\text{frost}}|t_0, y = 1941, \Delta T^{(3)})$, as well as $P_{\text{data}}(t_{\text{frost}}|t_0, y = 1941)$. Black denotes the cases in which the null hypothesis of a normal distribution was rejected with a significance level of $\alpha = 5\%$.

For initial dates in summer, i.e. July 1$^{\text{st}}$ and September 15$^{\text{th}}$, this new anomaly time series actually mostly yields less indications of predictability than the earlier version, especially when considering the mean of the conditional first passage time distributions.

For initial dates in winter and spring, i.e. December 1$^{\text{st}}$, February 14$^{\text{th}}$ and April 1$^{\text{st}}$, the comparison yields no clear conclusion: For some initial years, there is a definite improvement, for others this new anomaly time series actually leads to worse results, while in some cases the results are very similar.

For the other three cases of May 15$^{\text{th}}$, October 1$^{\text{st}}$ and November 1$^{\text{st}}$, however, there is a definite improvement in the predictability effects, meaning that the dependence of key summary measures of the conditional first passage time distributions on the initial anomalies is generally more pronounced in this case than before. This means that the more short-term prediction is improved, impacting both initial dates in autumn when the date of first frost is close, and the last dates in spring for which frost might still be observed within the next days.

5.3.3 Statistical tests of significance in distribution differences

To verify that the predictability effects are statistically significant also for this version of the anomaly time series, we again use statistical hypothesis tests. In keeping with the results for the anomaly time series without detrending, we start by checking whether the underlying conditional first passage time distributions are still not normal for this case. A Jarque-Bera hypothesis test as introduced in Sec. 2.2.4 confirms that $P_{\text{data}}(t_{\text{frost}}|t_0, y, \Delta T)$ is almost never a normal distribution. Fig. 5.18 shows the test results for $y = 1941$, i.e. a minimal average temperature, which has the most instances for which the test cannot reject the null hypothesis. Starting instead with $y = 1999$, there are no such exceptions any longer, the null hypothesis of normality is then rejected for all other parameter values.

The Kolmogorov-Smirnov test that was introduced in Sec. 2.2.5 is therefore still the most suitable one for the conditional first passage time distributions. For the detrended anomaly time series, we again looked at test results for three different initial years y. Fig. 5.19 shows the case with the least indications of predictability, namely $y = 1965$. It contains the largest number of parameter values for which the equality of the underlying probability distribution could not be rejected with a confidence of 5%. As can be seen, the null hypothesis can still mostly be rejected during the winter half of the year, while the distributions are very similar in the summer half. The first passage time to frost is therefore, as expected, still not predictable for t_0 in summer.

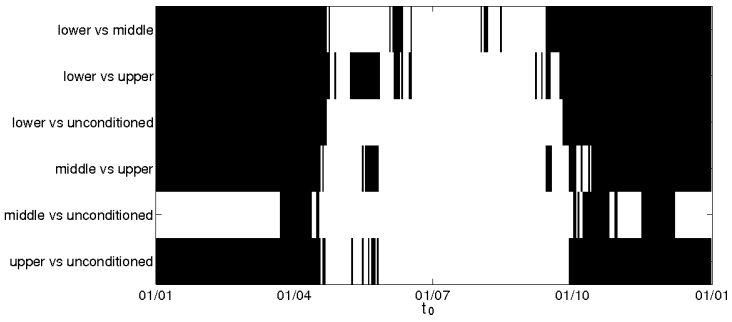


Figure 5.19: Results of a Kolmogorov-Smirnov hypothesis test with a confidence level $\alpha = 5\%$ comparing $P_{\text{data}}(t_{\text{frost}}|t_0, y = 1965, \Delta T^{(3)})$ for the different initial anomaly terciles, as well as $P_{\text{data}}(t_{\text{frost}}|t_0, y = 1941)$. Black denotes the cases in which the null hypothesis of equal underlying probability distribution was rejected.

time series ΔT	non-detrended	detrended y=1941	detrended y=1965	detrended y=1999
lower and middle tercile	234	247	237	228
lower and upper tercile	274	263	248	275
lower tercile and unconditioned	225	221	211	206
middle and upper tercile	245	221	210	211
middle tercile and unconditioned	70	56	68	109
upper tercile and unconditioned	220	213	212	213

Table 5.2: Number of initial calendar days t_0 for which the $P_{\text{data}}(t_{\text{frost}}|t_0, y, \Delta T^{(3)})$ for different combinations of ΔT are statistically distinguishable by a Kolmogorov-Smirnov test with confidence level $\alpha = 5\%$. Unconditioned refers to $P_{\text{data}}(t_{\text{frost}}|t_0, y)$.

However, the results of the test are very similar to those for the non-detrended anomaly time series as displayed before in Fig. 4.11. It is therefore not apparent whether the more involved method of generating anomalies from the temperature time series has led to an improvement in predictability indicators, such as an increase in initial days for which the conditional first passage time distributions to frost depend significantly on the initial conditions. In order to better evaluate this, Table 5.2 shows a summary for both anomaly time series, with the number of days for which the difference in distribution is significant for different combinations of initial anomaly conditions.

As expected, the first passage time distributions conditioned on the lower and upper initial anomaly tercile are the most distinguishable with a Kolmogorov-Smirnov test. More surprisingly, however, the non-detrended time series generally shows more distinguishable days than the detrended time series on average. In two thirds of the cases it even shows more than all of the initial years analysed here for the detrended time series. This directly implies that the detrending diminishes the predictability effects.

A closer look at the p value of the Kolmogorov-Smirnov test for $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3)})$ with ΔT in the lower and upper initial anomaly tercile for both the non-detrended and the detrended time series is taken in Fig. 5.20. For the latter, we chose the initial year $y = 1999$ when the predictability effects extend over the same number of initial days t_0 as for the non-detrended time series. The differences are not as slight as suggested by the almost equal time span for which the p value exceeds the confidence level $\alpha = 5\%$. In spring, the detrended time series performs better, with many cases close to the limit but not above it. In mid summer, however,

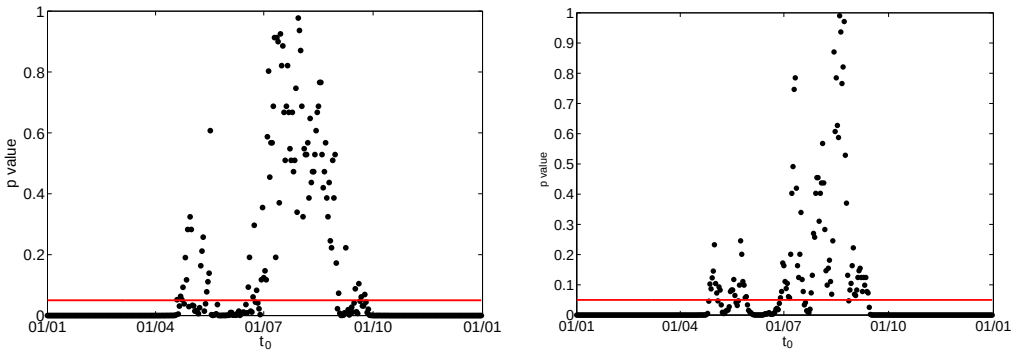


Figure 5.20: p value of the Kolmogorov-Smirnov test comparing $P_{\text{data}}(t_{\text{frost}}|t_0, y = 1999, \Delta T^{(3)})$ with ΔT in the lower and upper tercile, i.e. the probability that the underlying distribution is one and the same. The left panel shows the detrended temperature anomalies, the right panel the simpler approach (see Fig. 4.12). The horizontal lines represent the 5% confidence level.

there are no predictability effects in the detrended time series, when at the end of July there was some distinguishability in the non-detrended time series.

While the statistical hypothesis test hints that detrending the anomaly time series does not only fail to improve the predictability of first frost, it is not quite conclusive enough to state that it indeed makes it even worse. After all, even with less calendar days on which the underlying probability distribution are significantly different so that a forecast can make use of these effects, the differences on the distinguishable days might be much larger thus still leading to an improvement of the forecasts that the Kolmogorov-Smirnov test could not judge. We will therefore need a direct comparison of actual first frost forecasts to be able to make a definite statement.

5.3.4 Predictability: Summary and examples

Before looking at full predictions of the first passage time to frost in the next section, we first evaluate the two examples of predictability already shown in Sec. 4.2.4 for the non-detrended time series: The probability of frost occurring before summer when conditioning on an initial date on April 1$^{\text{st}}$ and the date of first frost when conditioning on an initial date on October 15$^{\text{th}}$.

Fig. 5.21 shows again a nice dependence of the probability of frost occurring again before the summer both for an initial year with minimal mean temperature and an initial year with maximal mean temperature. As can be seen from a direct comparison of the two cases, this probability is very dependent upon the initial year. Indeed the probabilities for $y = 1999$ are lower by around 30% than those for $y = 1941$.[5]

Comparing the detrended time series again to the earlier non-detrended version as shown in Fig. 4.13, we see a change in probability across the anomaly deciles of slightly less than 30% for $y = 1999$ and around 22% for $y = 1941$ in the detrended time series. For the non-detrended case it was around 30%, confirming the earlier conclusion that the detrending did not enhance the predictability.

Moving on to initial dates in October, we again looked at the date on which in 10% of all cases first frost had already occurred. This was chosen to minimise the influence of outliers and statistical fluctuations. Fig. 5.22 again shows a nice dependence of this date on the initial

[5] $y = 1965$ yields a result between the two shown here. This case was therefore left out in favour of displaying the limiting cases.

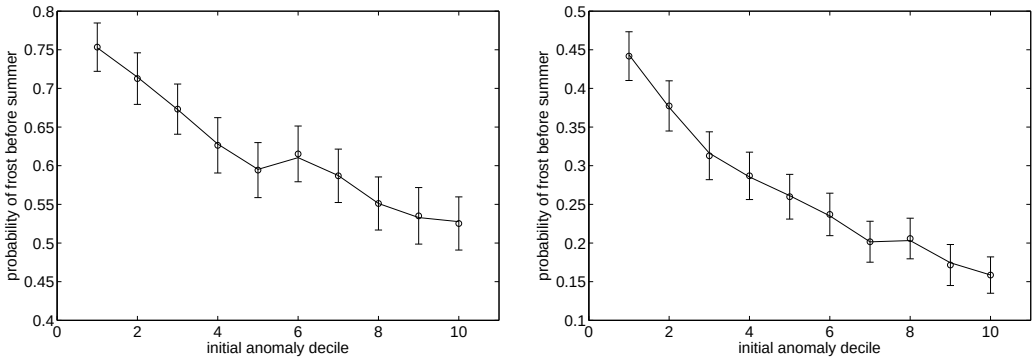


Figure 5.21: Probability of frost occurring before summer depending on the initial anomaly decile for an initial date of 1 April 1941 (left panel) and 1 April 1999 (right panel). The straight line represents the mean value obtained through 1000 bootstrap calculations, the error bars show double the resulting standard deviation.

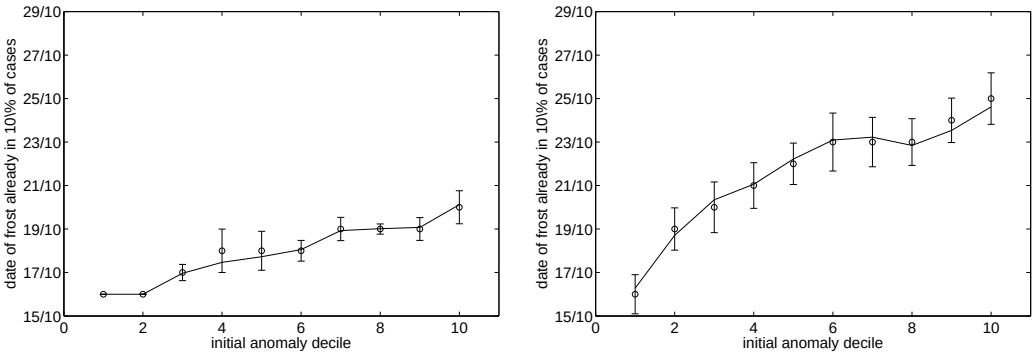


Figure 5.22: Date for which in 10% of cases first frost has already occurred, depending on the initial anomaly decile, for an initial date of 15 October 1941 (left panel) and 15 October 1999 (right panel). The straight line again represents the mean value obtained through 1000 bootstrap calculations, the error bars show double the resulting standard deviation.

anomaly decile both for $y = 1941$ and $y = 1999$. Again, this influence is stronger for an initial year with maximal mean temperature, where the initial date shifts by 9 days from the lowest to the highest initial anomaly decile. By comparison, for the initial year with minimal mean temperature the difference is only around 4 days.

Compared to the non-detrended time series as shown in Fig. 4.14, this is not conclusive: In that case, the date shifted by 6 days, which is the same result as for the initial year with average mean temperature. It therefore seems that in years with a high mean temperature, these predictability effects are stronger or at least equal to the non-detrended case, while for years with a low mean temperature, they are weaker than in the non-detrended case. In order to fully evaluate the usefulness of the detrending, we will however need to look at the actual predictions that also take into account the actual composition of the time series into years with low, average and high mean temperature.

forecast scheme	benchmark	no ΔT	$\Delta T^{(3)}$	$\Delta T^{(3w,10)}$	$\Delta T^{(10)}$
RMSE [days]	73.5	31.7	31.5	31.6	31.4
RMSE (original prediction) [days]	73.5	31.1	30.8	31.0	30.8

Table 5.3: Root mean square error of deterministic forecasts of the first day of frost for the time series with improved stationarity and for the simpler original version (see Table 4.1). The benchmark forecast uses the climatology, the others the mean values of an estimate of $P_{\text{data}}(t_{\text{frost}}|t_0)$ and an estimate of $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T)$ for differently coarse-grained initial anomalies ΔT.

5.4 Actual predictions

5.4.1 Deterministic prediction

As in Sec. 4.3.1, we will start by considering the deterministic prediction task of forecasting an actual day on which we expect first frost to happen. We construct our prediction by calculating the mean of the estimated $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T)$, again with different coarse-graining schemes of the initial anomalies ΔT. The benchmark is chosen as before, i.e. we forecast $t_{\text{frost}} = t_0 + 1$ if $\widetilde{T}(t_0) \leq 0$ °C. For $\widetilde{T}(t_0) > 0$ °C, we issue the remaining time to the date on which the climatological temperature drops below this threshold as a forecast.

Table 5.3 lists the resulting root mean square error (RMSE) of the prediction both for this new temperature anomaly time series with improved stationarity and for the original time series used in Chapter 4. The error of the RMSE score as obtained through the standard error of the mean is 0.4 days for all different cases considered here. As can be seen, the prediction using the more involved definition of temperature anomalies actually performs slightly worse than the original time series for all different prediction schemes used here. Considering the error of the prediction scores, this distinction, while systematic, does not appear to be statistically significant.

Moving on to the second score, namely the proportion correct (PC) which does not aim at predictions that lie as close to the verification on average as possible, but rather those that are as often correct as possible, we come to a different conclusion. As Table 5.4 shows, the PC is higher for the more involved anomaly definition. This even goes so far as to be statistically significant if one issues the date on which in 10% of all cases frost has already happened, or alternatively the mean of the distribution, as forecasts. Considering that calculating separate confidence intervals for each score is less powerful for determining significant differences than calculating a single confidence interval for the score difference to see whether it includes the zero or not, this is a very significant finding[75, 76]. However, the PC is still not very high overall.

The more involved definition of temperature anomalies therefore does not help with prediction quality as hoped for. The only improvement it has brought is a slightly higher proportion correct of the prediction, meaning that the correct date is forecast a little more often, even though the effect is small. The root mean square error, on the other hand, does not truly distinguish between the two possibilities explored up to now.

5.4.2 Binary forecasts

Changing to the binary forecasting of next frost still occurring before summer, we again look first at the whole prediction time window as determined from the benchmark in Sec. 4.3.2. Table 5.5 shows that not only the predictions are again not statistically distinguishable from the benchmark forecast, but they also score worse than those obtained from the original version of the anomaly time series as considered in Chapter 4.

forecast scheme	benchmark	first date	date of $p = 10\%$	mean	median
PC [%]	3.4 ± 0.1	6.0 ± 0.2	7.2 ± 0.2	3.8 ± 0.1	4.3 ± 0.1
PC (original prediction) [%]	3.4 ± 0.1	6.0 ± 0.2	6.7 ± 0.2	2.9 ± 0.1	4.2 ± 0.1

Table 5.4: Proportion correct of different forecasting schemes for the time series with improved stationarity and for the simpler original version, verifying how often the exact day of first frost was forecast. The benchmark again uses the climatology, the forecasts are based on different measures of the estimate of $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3)})$, namely the first day with non-vanishing probability of frost, the date on which in 10% of all cases in the training set first frost after winter had already occurred, as well as the mean and the median of the distribution. The errors were obtained using the standard error of the mean proportion correct.

	benchmark	$\Delta T^{(3)}$	$\Delta T^{(3w,10)}$	$\Delta T^{(10)}$
Brier score (BS)	0.139 ± 0.003	0.139 ± 0.004	0.140 ± 0.004	0.138 ± 0.004
Brier skill score (BSS)	0%	0.3%	-0.8%	0.6%
BS (original anomalies)	0.139 ± 0.003	0.134 ± 0.003	0.136 ± 0.004	0.134 ± 0.003
BSS (original anomalies)	0%	3.3%	2.1%	3.4%

Table 5.5: Brier score and Brier skill score applied to initial dates $t_0 \in [70, 123]$ for different forecasting schemes obtained from both the anomaly time series with improved stationarity and the original version, as well as the benchmark. The errors were determined using a bootstrap procedure with 1000 samples.

As the quality of the predictions for the original time series depended sigificantly on the initial date t_0, we again look at the variation of the Brier skill score across the time period considered for the binary predictions. Fig. 5.23 shows the new results and for better comparison also again the previous results for the original version of the anomaly time series. Note that the earliest dates in the prediction window were omitted for the new results as the error bars were too large to permit any conclusions to be drawn. As can be seen, the forecast quality is again rather low - compared to the original time series there are even more days for which the benchmark performs better than the forecast.[6]

To analyse the possibility of improving the forecasts by using weighted terciles with different anomaly band widths p, Fig. 5.24 shows the dependence of the Brier skill score both on p and on t_0. Only for initial dates around April 9^{th}, the forecasts perform well for a large range of p. For other values of t_0, the forecast performs mostly similar or worse than the one obtained from the original version of the anomaly time series. The regions in parameter space for which the benchmark actually outperforms the forecasts have also grown. The reduced amplitude of the error bars in the left panel of Fig. 5.23, however, makes the conclusions drawn from this contour plot more reliable.

Improving the stationarity did therefore not have a positive impact on the binary forecasts.

5.4.3 Probabilistic forecasts

Finally moving on to the probabilistic forecast of the full first passage time distribution up to a maximal first passage time $t_{\max}$, we will check whether improving the overall stationarity of the time series had a negative impact also in this case.

[6] Note that the benchmark for the binary forecasts is based on the estimate of $P_{\text{data}}(t_{\text{frost}}|t_0, y)$. It is therefore much more sophisticated than the benchmark used for the deterministic forecasts and generally a very harsh standard[90].

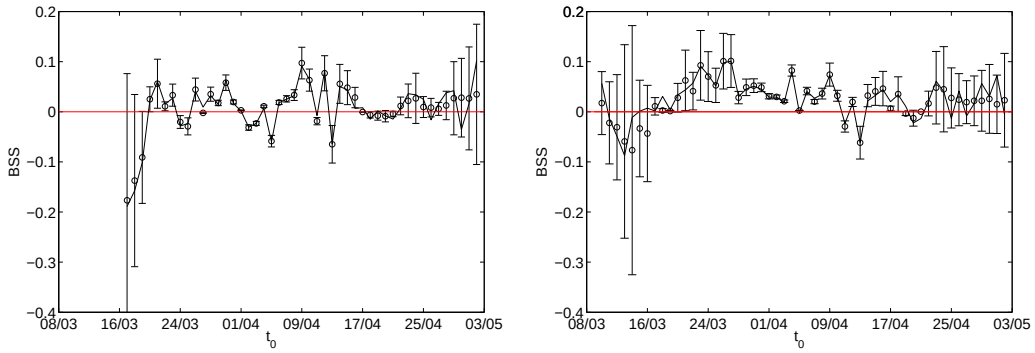


Figure 5.23: Brier skill score comparing the binary forecast of frost still occurring before summer using $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3)})$ to the benchmark both for the anomaly time series with improved stationarity (left panel) and for the original version (right panel - see Fig. 4.18). The error bars were obtained using twice the standard deviation of 1000 bootstrap samples, the horizontal red line denotes no forecast improvement over the benchmark.

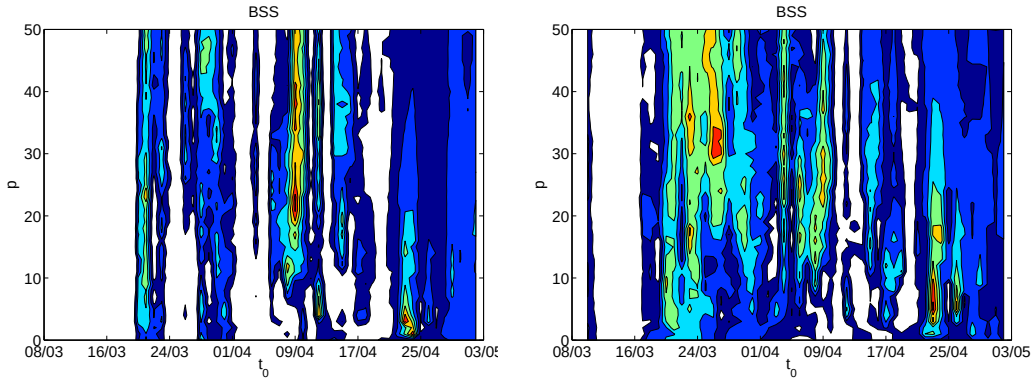


Figure 5.24: Contour plot of the Brier skill score comparing the binary forecast of frost still occurring before summer using $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3w,p)})$ to the benchmark both for the anomaly time series with improved stationarity (left panel) and for the original version (right panel - see fig. 4.19). White denotes the region of negative skill scores, the other colours show increasing skill scores in increments of 2% from dark blue for BSS $\in [0\%, 2\%]$ to dark red for BSS $\in [12\%, 14\%]$.

Fig. 5.25 shows that for both anomaly time series the dependence of the ranked probability score on the weight of the outer initial anomaly bands p is similar with a minimum at $p = 30\%$. Also the separation of initial anomalies into deciles obtains the same score as the separation into halves for both time series. However, the ranked probability score for the time series with improved stationarity is markedly higher than for the original version, meaning that here, too, the improved stationarity has worsened the forecast.

This can also be seen when looking at the seasonal dependence of the corresponding skill score: Fig. 5.26 shows less areas with more than 10% improvement over the benchmark and more areas with negative skill score than for the original anomaly time series. Especially in March, when forecasts were quite good for the original time series, they now score worse than the benchmark. Also for small values of p there is a definite lowering in forecast quality when compared to the first version.

When looking at the dependence of the forecast quality on the maximal fully resolved first passage time, the forecast and benchmark scores are also much closer together than before (see Fig. 5.27), leading to a reduced skill score.

5.5 Conclusion

In the previous chapter, even though the forecasts had significant quality, there was still room for improvement. Considering that we coarse-grained the initial condition on the date t_0 to retain an adequate number of data points for the estimation of $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T)$, the time series of anomalies was assumed to be stationary. This was clearly not the case as both the mean value of the anomalies increased over the years and their standard deviation changed depending on the calendar month. In this chapter, we therefore tried to improve the forecasts by using a more involved definition of the temperature anomalies that both detrended the time series and deseasoned the variance.

While the selection issues of initial anomalies across the calendar years have been eliminated with the trend in this new version of the anomaly time series, there are still some issues across the different calendar months. The skewness of the anomaly distribution and with it the anomaly range still change across the seasons, albeit with smaller magnitude and more smoothly than the variance did before. Also the autocorrelation structure shows some seasonality with a greater persistence in winter than in any other season. While the stationarity of the anomaly time series has indeed been improved with the new procedure, there are therefore still remaining issues.

Elimination of the trend also led to an improvement of the fit of the anomalies by an AR(8) model process that is not rejected by a Ljung-Box test. However, the AR(8) process is still inadequately reproducing the first passage time distributions due to the remaining seasonality. It also introduces systematic errors in the estimation of the conditional first passage times and can therefore still not be used to improve the available number of data points.

Comparing the conditional first passage time distributions to those obtained from the original version of the anomaly time series, there are small differences mostly for outliers around t_0 in spring. The change of their mean values with the initial anomalies is smaller than for the original time series, while their spread changes more than before. The predictability effects observed in summer are smaller than before, those in autumn and spring larger than for the original time series.

While the precise date on which the anomalies were originally recorded does not appear to have any impact on the conditional first passage time distributions, the initial year y has become another factor of great influence: It determines the mean temperature value to use when retransforming the anomalies back to temperatures to evaluate the threshold crossing criterion. The predictability effects are the most pronounced for an initial year with a high mean temperature.

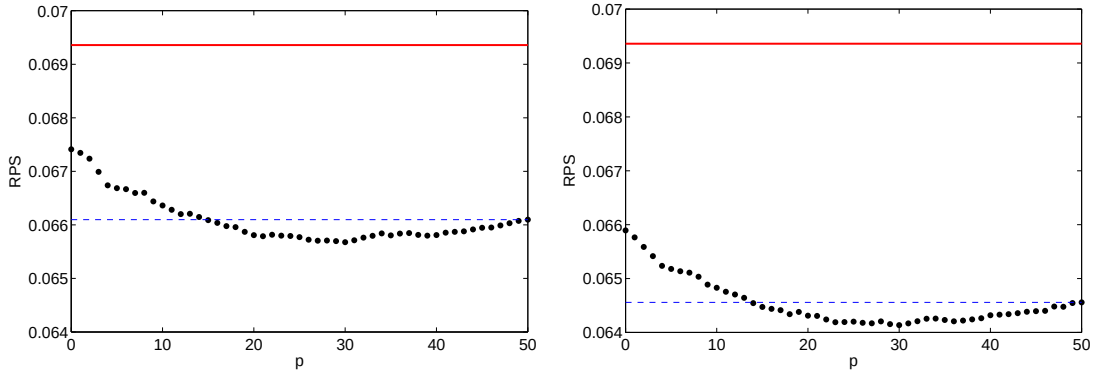


Figure 5.25: Ranked probability score for the time series with improved stationarity (left panel) and the original version (right panel - see Fig. 4.20) using $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3w,p)})$ and first passage times of up to 30 days. The red horizontal line shows the score of the benchmark forecast, the blue dashed line the score of the forecast using $\Delta T^{(10)}$.

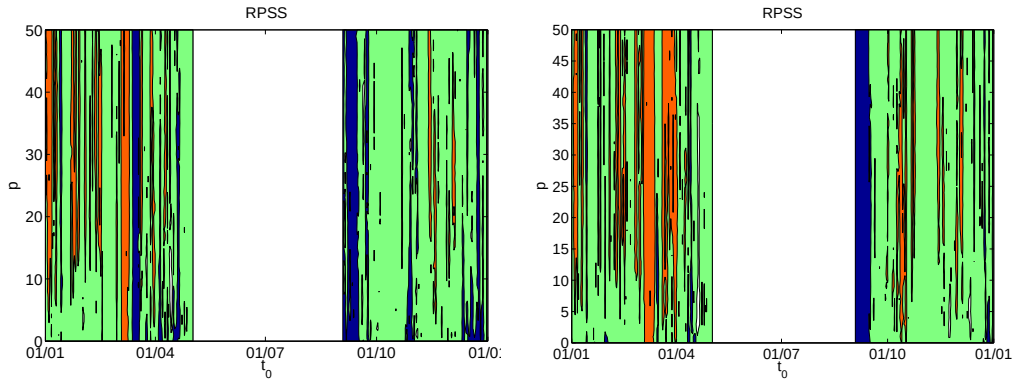


Figure 5.26: Contour plot of the ranked probability skill score using $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3w,p)})$ for the time series with improved stationarity (left panel) and for the original time series (right panel - see Fig. 4.22). White denotes perfect forecast and benchmark, dark blue corresponds to negative RPSS, green to positive RPSS that is smaller than 10% and red to positive RPSS that exceeds 10%.

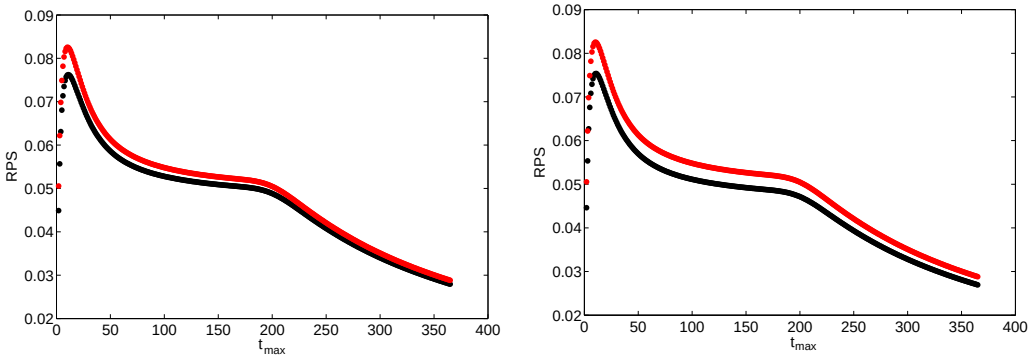


Figure 5.27: Ranked probability score for the time series with improved stationarity (left panel) and the original time series (right panel - see Fig. 4.21) using $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3w,30)})$ for different maximal first passage times.

Using the Kolmogorov-Smirnov test to analyse where the distributions $P_{\text{data}}(t_{\text{frost}}|t_0, y, \Delta T^{(3)})$ are significantly different for the lower and upper initial anomaly tercile, we found that the detrending reduced this number of initial days t_0 - a clear indication that the potential for predictability had diminished with the improved stationarity.

Issuing actual forecasts confirmed this initial impression. In the deterministic case, the forecasts calculated from this new version of the anomaly time series scored worse on average. Only if considering the proportion correct of the forecasts, some improvement was achieved. The binary forecast also performed worse with a clear increase in the number of days for which the benchmark actually scored better than the forecast. Only for t_0 around April 9^{th} the forecasts improved with the new version of the anomaly time series. The full probabilistic forecasts also scored worse than before, especially for t_0 in March where the benchmark is now better than the forecast while before the forecast managed an improvement on the benchmark by over 10%.

This lack of improvement might be due to the additional sources of errors introduced into the forecasting procedure by the detrending and deseasoning of the variance, without the benefit of a truly stationary time series. As improving the stationarity of the anomaly time series did not help improving the quality of the predictions of first frost, we will try a different possibility of improving the forecast schemes in the following.

6 Adding information on the North Atlantic Oscillation (NAO)

6.1 Introduction

Chapter 4 showed that we were able to issue forecasts directly from the data that scored better than the respective benchmark forecasts for several different prediction tasks. However, the absolute quality still left room for improvement. In Chapter 5, we improved the stationarity of the temperature anomaly time series and thus hoped to reduce the errors introduced by coarse-graining the condition on the initial date t_0. This did not succeed in improving the forecasts correspondingly except for the forecast accuracy in the deterministic case. In this chapter, we will therefore analyse another possible improvement of the forecasts: Adding information from a second relevant variable, as the fraction of variability accounted for by a single factor is generally not very large at least in the extratropics[29]. As the improved stationarity of the anomalies led to slightly worse forecasts, in this chapter we will revert back to the first and simpler way of defining the temperature anomalies as given by Eq.(3.2).

6.1.1 Choosing the second variable

First, though, we need to choose an appropriate second variable. Considering the large inter-dependence of meteorological variables, the temperatures are influenced by a great number of factors. Foremost is the current "Großwetterlage" (general weather situation) in the vicinity of the station for which the forecasts are to be issued. In fact it has been found previously that the memory in surface air temperatures is due to "regime-like behaviour" in the atmosphere[134]. The second variable should therefore characterise the current general weather situation on the initial date t_0.

As described in Sec. 2.4, the most apparent atmospheric pattern that influences European surface temperatures is the North Atlantic Oscillation (NAO). The NAO is constituted by fluctuations in the atmospheric pressure at sea level between the low pressure system in the vicinity of Iceland and the high pressure system that is usually located between the Azores and southern Spain. Indeed, it has been affirmed previously that surface air temperatures are driven by the passage of low and high pressure systems[134], pointing to the use of the NAO. Moreover, it has been found that changes in the NAO are reflected markedly in surface air temperatures especially over north-eastern Germany[110] and it should therefore influence the temperatures in Potsdam. As the NAO is still not well understood, current global circulation models used for operational weather forecasts do not incorporate it correctly[98, 99]. It is therefore a good candidate for data-driven prediction schemes and we will focus in the following on the NAO as second input variable for our predictions.

6.1.2 Choosing the specific index

While Sec. 2.4 described the mechanisms behind the NAO's influence on European temperatures and the properties of this atmospheric pattern, it also listed a wide variety of possibilities for the characterisation and quantification of its current state. We therefore first need to choose an appropriate index to represent the NAO for our forecasts.

Our goal is to compare the possible merits of statistical forecasts based directly on time series analysis to the much more complicated model-based forecasting approaches. We therefore evidently need a "model-free" quantification of the actual state of the NAO that does not require too much additional computational effort, i.e. a measurement-based index. We will choose a NAO index constructed only from pressure data, in keeping with Jones et al.[108]. Luckily, correlations between different indices, station-based or more elaborate ones, are very high especially in winter[107]. The exact choice of index should therefore not have an overly large impact on the following analysis.

The only choice remaining are the measurement stations chosen for the index computation as several different stations have been used to characterize the state of each of the action centers of the NAO. For the northern low pressure system, the most common choices are the Reykjavik, Stykkisholmur and Akureyri stations, all on Iceland. While Reykjavik offers the longest time series duration[108], Stykkisholmur also has long time series and is part of the most commonly used index in climate research[101] which was proposed by Hurrell. Tinz found that using the surface air pressure in the city of Bergen on the Norwegian west coast characterises the Potsdam winter temperatures best, while in summer any station on Iceland would be appropriate[96]. The lack of conclusive evidence for any one station is due to the fact that the temporal variability in the North is much larger than the spatial variability[94, 108], with the correlation coefficient between the Stykkisholmur and Akureyri time series as large as 0.98[112]. The choice of the northern station for the index is therefore not so important and can be made on ready availability of the time series.

For the southern station there also exist three common choices: Ponta Delgada, Lisbon and Gibraltar. The traditional choice has been Ponta Delgada on the Azores, a station that is close to the center of action especially in summer when the NAO systems move westward[112]. It does provide a reasonably long time series[108], but stopped operating in 1997. Moreover, it was found that more easterly stations might be better in winter[108]. Indeed, Stephenson remarks that easterly stations such as Gibraltar give the strongest correlations with Central England temperatures in winter[98]. Other studies have also found that Lisbon or Gibraltar correlate better with temperature records over Europe than the Azores station[108, 111] and are in fact highly correlated with each other in winter[108]. Lisbon has a longer time series than Ponta Delgada and also a higher signal-to-noise ratio at least in winter[112]. Tinz found that for the Potsdam temperatures in winter, Gibraltar offers the best characterisation, while in summer a station near the center of the Baltic Sea such as Stockholm would be best[96]. The choice of a southern station is therefore not clear-cut either, especially since it was pointed out that the southern station does make a difference especially if one does not consider winter[109].

Fortunately, it was also found that indices based on sea-level pressures from different pairs of stations are highly correlated[101]. Indeed, the correlation coefficient between the station pair Lisbon and Stykkisholmur and the station pair Ponta Delgada and Akureyri for the whole joint time series length of 1894 to 1995 is as high as 0.93[112]. Therefore, we will make our choice based upon the ready availability of the relevant time series and settle on the monthly pressure differences between the stations of Gibraltar and Reykjavik. The pressure measurements are normalised separately for each station by the monthly values of the reference period 1951-1980.[1] This has the added benefit that it correlates best with the more involved Eulerian index found to be most appropriate to describe temperature anomalies over Germany[107].

[1] This time series was first compiled by Jones[109] and is now regularly updated by Osborn at the Climate Research Unit[140].

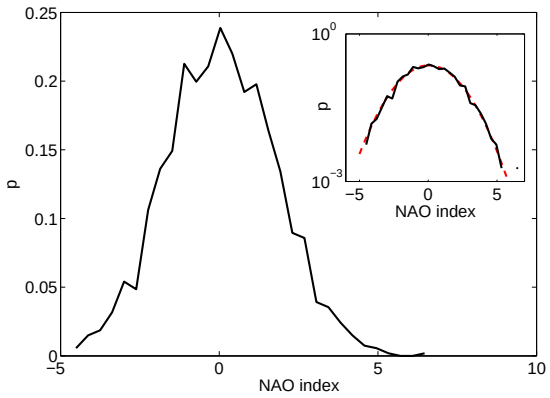


Figure 6.1: Probability distribution of the NAO index following Jones[109] for the years 1893 to 2010. The inset shows the same on a semilogarithmic scale, with the red dashed line representing the closest fit with a normal distribution.

6.2 NAO data preparatory analysis

6.2.1 NAO data properties

Before we can issue forecasts containing information from this second variable, we will first analyse the properties of the NAO time series as provided by the Climate Research Unit[140].

The first year without any missing values in the original dataset provided by Jones[109] is 1825, it only extends until the end of 1999 but is being continuously updated by Tim Osborn. Therefore monthly values are available for the whole period of our temperature time series. As described above, the index was constructed from pressure measurements that were normalised as follows:

$$\Delta P_i = \frac{P_i - \langle P \rangle}{\hat{\sigma}}, \tag{6.1}$$

where P_i denotes the pressure measured at time i, $\langle P \rangle$ is the mean pressure in the time series during the reference period 1951 - 1980 and $\hat{\sigma}$ the corresponding standard deviation. The index NAOI is then constructed for each month from the stations Gibraltar (G) and Reykjavik (R):

$$\mathrm{NAOI}_i = \Delta P_{G,i} - \Delta P_{R,i}. \tag{6.2}$$

A positive NAO index therefore means that the pressure difference between the two stations is larger than average. A negative NAO index represents the case where the pressure difference is smaller than average. A true reversal, i.e. a northern high-pressure system with a southern low-pressure system occurs only very rarely[101].

For the years 1893 to 2010 the probability distribution of the NAOI is very close to normal, as can be seen in Fig. 6.1. The tail of high NAO indices is, however, slightly longer with two instances of a positive index exceeding 3σ in magnitude. To verify that the index indeed contains spectral components that show a time dependence on seasonal scales, we look at an estimate of the power spectral density, following Welch's method as explained in detail in Sec. 2.1.3. As can be seen from the zoom into the smaller frequencies in the right panel of Fig. 6.2, it appears as if there are significant components on a yearly basis (which is to be expected since the index was not deseasoned), but also for time scales of slightly over a year, 1,5 and 2,5 years, and around 6 years. The left panel of the same figure shows the yearly and twice-yearly frequency peaks also inherent in the temperatures, as well as reflecting the four seasons in a year.

However, the significant noise inherent in the estimate precludes forming any significant con-

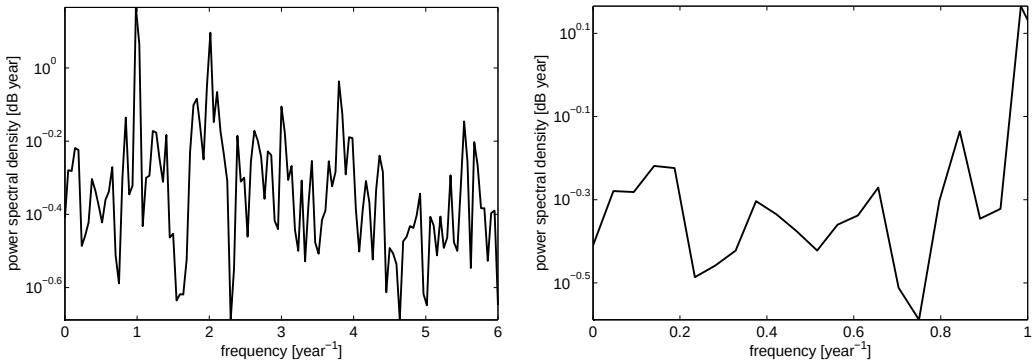


Figure 6.2: Power spectral density estimate through an ensemble average of 11 periodograms of time series segments of length 256 data points with a 50% overlap between consecutive segments. The right panel shows a zoom into the small frequencies.

clusions as to specific frequencies involved in the dynamics of the NAO index. The smaller frequency components are nevertheless an encouraging find that indicates that additional conditioning on the index might well improve the seasonal forecasts.

6.2.2 Conversion to a daily time series

The major problem this NAO time series poses is the lack of daily information. On one hand, the power spectral density is close to a constant so that it resembles white noise. This hints at serious undersampling of the measurements. On the other hand, we want to issue forecasts on a daily basis. We therefore need to convert the monthly to a daily time series to obtain initial condition information on the NAO for every day. Simply using the monthly NAOI value for every day in the month is equal to using a step function to interpolate. This procedure introduces significant discontinuities at the end of each month. To prevent this, we could use more sophisticated interpolation schemes to convert the time series to daily values. However, there are many different ways to achieve this. We will choose three of them in addition to the step function and compare the resulting forecast quality: Linear, cubic and cubic spline interpolation.[2]

Fig. 6.3 illustrates the differences in these interpolation schemes. The spline interpolation results in the smoothest curve thus appearing most realistic. It is the only scheme that over- and undershoots the NAO index, broadening its range. The step function on the other hand is the only interpolation scheme that conserves the mean NAOI for each month. The forecast quality using each of the interpolation schemes will be assessed in the following. For clarity in these analyses, we will always keep note of the different interpolation schemes being used, denoting them as $\mathrm{NAOI}_{\mathrm{spline}}$ for example.

Having thus obtained a daily time series of NAO index values as a second input to our forecasting schemes, we will now proceed to analyse its impact on the predictability effects previously looked at in Sec. 4.2.

[2]Since we have additional values of the NAOI both before the period covered by the temperature time series and after 2010, we can escape the need for extrapolation at the ends. This avoids possible artefacts caused by a strong change in NAOI between the first two or the last two data values. For actual forecasts into the future, the daily pressure measurements would be needed.

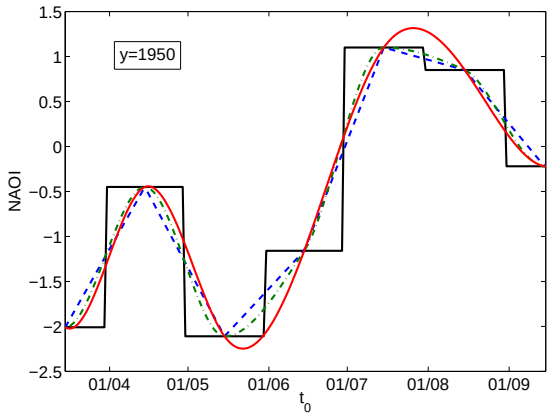


Figure 6.3: Example of the NAO index time series with the monthly mean values taken to be the index at the middle of the month. The time series was constructed using a step function (black), linear interpolation (blue), cubic interpolation (green) and spline interpolation (red).

6.3 Potential predictability

6.3.1 Change of first passage time distributions under conditioning

In order to issue forecasts incorporating the additional information provided by the NAO index, we will now have to estimate $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T, \text{NAOI})$. Previously, we had already coarse-grained the conditioning on the initial date t_0 by always contemplating a time window $[t_0 - w_t, t_0 + w_t]$ with $w_t = 45$ days. We had also analysed different coarse-graining schemes for the initial temperature anomalies, namely $\Delta T^{(3)}$, $\Delta T^{(3w,p)}$ and $\Delta T^{(10)}$, i.e. a separation into terciles, into three bands of unequal weight where the two outer bands each contain $p\%$ of all initial anomalies, and into deciles respectively.

To be able to incorporate the additional variable while still retaining an adequate number of data points from the training set for each distribution estimation, we will also have to coarse-grain the condition on the NAO index. In order to keep the statistical fluctuations small, we will only consider a separation into the positive and negative phase of the oscillation, denoted as $\text{NAOI}^{(2)}$.[3]

We analyse the differences between the estimates of $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3)}, \text{NAOI}^{(2)}_{\text{step}})$, but now for the six different categories of initial conditions. Fig. 6.4 plots the change of both the median and the standard deviation of these conditional first passage time probability distributions across the calendar year. For clarity, it only shows three categories of initial conditions, namely high initial anomalies with a positive NAO index, average initial anomalies with a negative NAOI and low initial anomalies with a negative NAOI. As can be seen, the differences in the median that were evident in autumn already before have been significantly enhanced by the additional incorporation of the NAOI. Even stronger is the change in the standard deviation, where the differences both in winter and also in late spring have been enlarged. This is a very promising first analysis of the influence of the NAOI on the predictability effects.

In the earlier chapters, we also looked at the full conditional probability distributions for different initial conditions, as well as the change of their summary measures with increasing and more finely splitted initial anomalies. Here, however, we do not only condition on the initial anomalies but also on the NAO index. The predictability analysis has therefore gained

[3] As there are consistently more values with NAOI > 0 than NAOI < 0 for every interpolation scheme, we will consider the negative phase to also contain the cases with NAOI $= 0$.

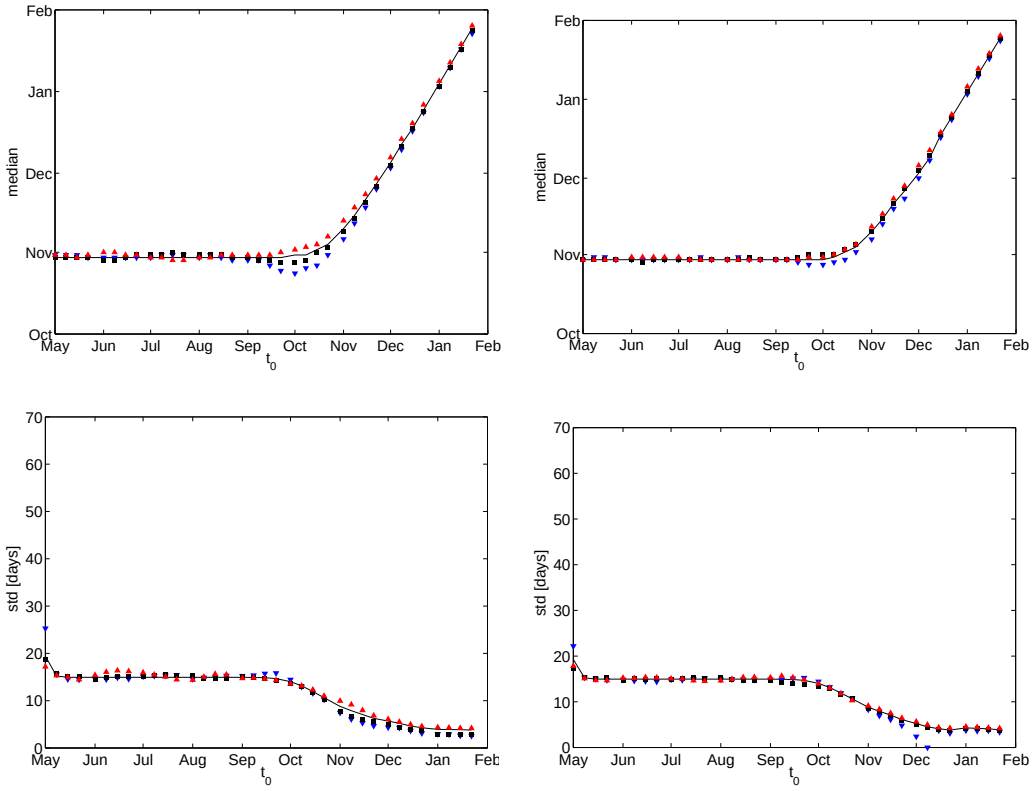


Figure 6.4: Median and standard deviation of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(3)}, \mathrm{NAOI}^{(2)}_{\mathrm{step}})$ (left panels). The continuous line denotes $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0)$, red upper triangles show the results for high initial anomalies and positive NAO index, black squares those for average initial anomalies and negative NAO index and blue downward triangles represent lower initial anomalies and negative NAO index. The right panels reproduce the corresponding results for $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(3)})$ as seen before in Fig. 4.3.

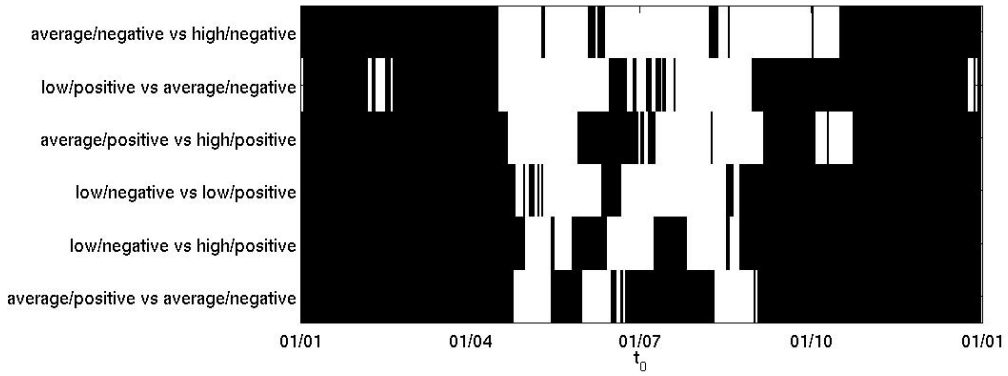


Figure 6.5: Results of Kolmogorov-Smirnov hypothesis tests comparing the distributions $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(3)}, \mathrm{NAOI}^{(2)}_{\mathrm{step}})$ for different combinations of initial anomaly terciles/NAO indices. Black denotes the cases in which the null hypothesis of equal underlying probability distribution was rejected with confidence $(1-\alpha) = 95\%$.

an additional dimension, making direct evaluations as done before even more unwieldy. As the increase in predictability can best and most significantly be judged using the scores of actual predictions, we will leave out the other analyses in this chapter and concentrate on the analysis of the statistical significance and the two predictability examples examined previously in Secs. 4.2.4 and 5.3.4 before proceeding to the actual first passage time predictions to frost in the next section.

6.3.2 Statistical tests of significance in distribution differences

Using the Kolmogorov-Smirnov test as introduced in Sec. 2.2.5 to evaluate the significance of the observed differences in the distributions $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(3)}, \mathrm{NAOI}^{(2)}_{\mathrm{step}})$ for different combinations of initial anomaly terciles and NAO indices confirms our previous conclusions: The additional variable significantly enhances the observed predictability effects. Fig. 6.5 shows several examples of the extent across the calendar year of statistically significant distribution differences (compare to Fig. 4.11 for the case without the NAO). In fact, the number of days n_{sig} for which the Kolmogorov-Smirnov test rejected the null hypothesis of equal underlying probability distribution increases from 200 out of 366 days for the two distributions with negative NAOI and average or high initial anomalies respectively, to 306 days for the case with average initial anomalies and positive versus negative NAOI. Testing all other pairs of initial conditions combinations results in values for the total number of days n_{sig} that lie between the two displayed extrema.

In order to better evaluate the increase in N_{sig} when compared to the results in Chapters 4 and 5, we listed its mean value and extrema in Table 6.1. As can be seen, with the incorporation of the NAO index the conditional first passage time distributions can be distinguished over two weeks longer on average than for the other cases. The most different cases of initial conditions are now distinguishable for over 10 calendar months, compared to the 9 months we achieved before. The least distinguishable case with $n_{\mathrm{sig}} = 200$ days appears worse if one incorporates the NAO. However, if one considers that we have 15 different values of n_{sig} in this Chapter and that the next highest value was $n_{\mathrm{sig}} = 238$ days, this cannot be interpreted as a deterioration in predictability overall.

Fig. 6.6 shows the change of the p value of the Kolmogorov-Smirnov tests with intial date

	original	improved	NAO
$\min(n_{\text{sig}})$ [days]	225	210	200
$\text{mean}(n_{\text{sig}})$ [days]	251	238	266
$\max(n_{\text{sig}})$ [days]	274	275	306

Table 6.1: Characteristic values of the number of calendar days n_{sig} for which the null hypothesis of equal underlying probability distribution for $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3)})$ and different combinations of initial conditions (original - Chapter 4), $P_{\text{data}}(t_{\text{frost}}|t_0, y, \Delta T^{(3)})$ (improved - Chapter 5) and $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3)}, \text{NAOI}^{(2)}_{\text{step}})$ (NAO - Chapter 6) was rejected by a Kolmogorov-Smirnov test with significance level $\alpha = 5\%$. All comparisons with $P_{\text{data}}(t_{\text{frost}}|t_0)$ were omitted here.

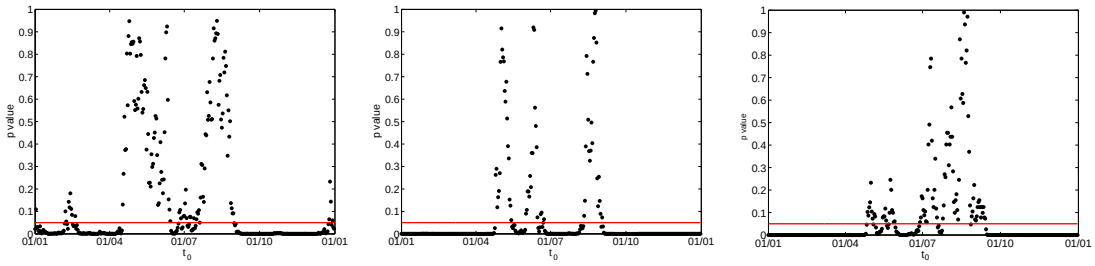


Figure 6.6: P value of the Kolmogorov-Smirnov test comparing $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3)}, \text{NAOI}^{(2)}_{\text{step}})$ for low initial anomaly with positive NAOI and average initial anomaly with negative NAOI (second worst case in terms of n_{sig} - left panel) and for average initial anomaly with positive NAOI and average initial anomaly with negative NAOI (best case in terms of n_{sig} - middle panel), as well as $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3)})$ with high and low initial anomaly (right panel). The horizontal lines represent the 5% confidence level.

t_0 for the second-worst and best case in terms of N_{sig} as well as the previous result without incorporating the NAOI for comparison. As can be seen, the distinction between periods with distinguishability and those without has become much clearer, i.e. there are now very few values that fluctuate near the confidence level. Other differences are even more striking, however. The increase in initial condition categories has introduced non-distinghuishable distributions for periods in time that had none of them in the cases without the NAO index, namely at the end of December and in February. Other periods in time have now lengthened stretches for which the distributions remain distinct, especially for t_0 in July and September. As the correlations between the NAO and the surface air temperatures have been found to vanish in July[96], this is especially surprising.

The Kolmogorov-Smirnov test has therefore shown that the incorporation of the NAO index has greatly enhanced the number of initial dates t_0 for which there are statistically significant differences between the conditional first passage time distributions to frost for different initial conditions.

6.3.3 Predictability examples

Before proceeding to the actual predictions of the date of next frost, we will again look at the two predictability examples already considered previously in Secs. 4.2.4 and 5.3.4.

First, we will again consider an initial date of April 1$^{\text{st}}$. Then the conditional first passage time distribution is bimodal and we have seen before that the relative weight of the two peaks

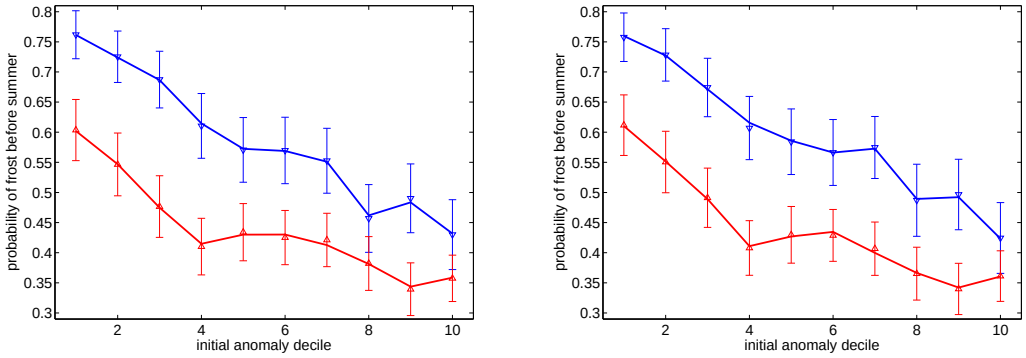


Figure 6.7: Probability of frost occurring before summer, depending on the initial anomaly decile, for NAOI > 0 (red) and NAOI ≤ 0 (blue). The left panel shows the results for $\mathrm{NAOI}^{(2)}_{\mathrm{step}}$, the right panel for $\mathrm{NAOI}^{(2)}_{\mathrm{spline}}$. The continuous lines represent the mean value obtained through 1000 bootstrap calculations, the error bars show double the resulting standard deviation.

of next frost before and after summer depended significantly on the initial anomaly decile. We now again calculate the probability of next frost before summer, i.e. the weight of the earlier peak, for each initial anomaly decile, but separately for a positive and a negative NAO index on the initial date.

Fig. 6.7 shows that there is a difference in probability of around 15% between the positive and the negative NAO case. Indeed, frost before summer is much more likely if not only the initial temperature is low as seen before, but also if the pressure difference over the Atlantic is smaller than normal. Again calculating error bars by using the bootstrap method, we can see that this difference between positive and negative NAO phase appears to be significant for 8 out of 10 initial anomaly deciles. If we use more sophisticated interpolation methods on the monthly NAO index, namely splines, then the statistical significance even extends across 9 initial anomaly deciles, although the differences between these two methods are otherwise rather slight. Using the NAO in addition therefore refines the predictability markedly in this case.

Considering a different initial date, namely October 15^{th}, and looking again at the date on which in 10% of all cases first frost had already occurred, the difference between the two NAO phases is somewhat smaller than in spring as can be seen in Fig. 6.8. Indeed, the estimated date varies by one to four days between the two cases and is rarely statistically significant if the step function version of the NAOI time series is used. In the case of the splines, the differences are slightly larger.

Incorporating the North Atlantic Oscillation Index as an additional variable to condition the first passage time distribution to frost on has therefore a significant impact on the predictability effects visible in the distributions. Now it remains to be seen whether this also translates into a marked improvement in the forecast quality of actual predictions to first frost.

6.4 Actual predictions

6.4.1 Deterministic prediction

We start again by issuing deterministic predictions of the day on which first frost will occur. To do this, we use the mean value of the estimated full conditional probability distribution

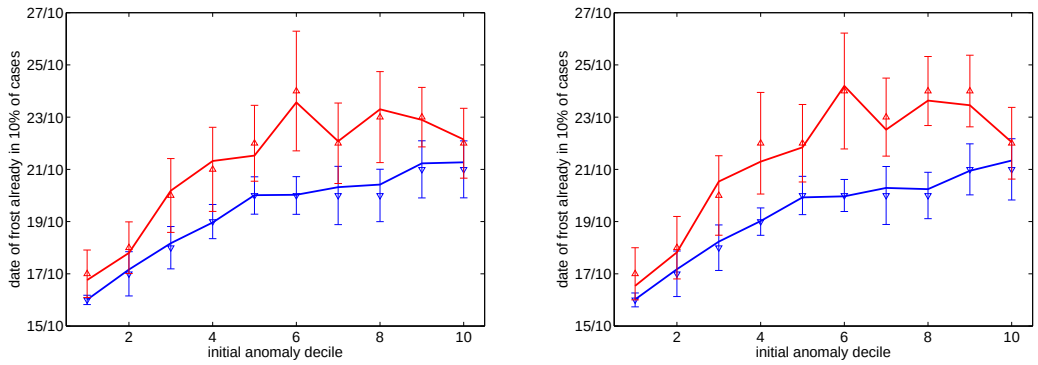


Figure 6.8: Date for which in 10% of cases first frost has already occurred, depending on the initial anomaly decile, for NAOI > 0 (red) and NAOI ≤ 0 (blue). The left panel shows the results for $\mathrm{NAOI}^{(2)}_{\mathrm{step}}$, the right panel for $\mathrm{NAOI}^{(2)}_{\mathrm{spline}}$. The continuous lines represent the mean value obtained through 1000 bootstrap calculations, the error bars show double the resulting standard deviation.

forecast scheme	no ΔT	$\Delta T^{(3)}$	$\Delta T^{(3w,10)}$	$\Delta T^{(10)}$
no NAOI information	31.1	30.8	31.0	30.8
$\mathrm{NAOI}_{\mathrm{step}}$	30.6	30.4	30.5	30.4
$\mathrm{NAOI}_{\mathrm{linear}}$	30.8	30.5	30.6	30.5
$\mathrm{NAOI}_{\mathrm{cubic}}$	30.7	30.4	30.6	30.4
$\mathrm{NAOI}_{\mathrm{spline}}$	30.7	30.5	30.6	30.5

Table 6.2: RMSE in days for deterministic forecasts of the first day of frost using different combinations of information about the initial conditions.

$P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T, \mathrm{NAOI})$ for different coarse-graining schemes for both the initial anomalies ΔT and the NAO index. We also use the probability distributions where we only condition on the initial date and one further variable, i.e. $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T)$ and $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \mathrm{NAOI})$.

The benchmark is again defined as forecasting $t_{\mathrm{frost}} = t_0 + 1$ if the climatological temperature $\widetilde{T(t_0)} \leq 0$ °C, and the day on which it drops below 0 °C in all other cases. The root mean square error (RMSE) for this benchmark was 73.5 days (see Table 4.1).

Table 6.2 lists the results from our forecasts. The error obtained as before by calculating the standard error of the mean was 0.4 days in all cases. As can be seen, all forecasts perform much better than the benchmark. However, the additional information brought by the NAO index did not significantly improve the forecasts. While the RMSE is systematically lower for all forecasts containing the NAO index, this difference is not sufficient to be distinguishable when taking the errors into account. As separate confidence intervals for each score value is a conservative method of establishing significance[76, 75], this is still a nice result.

Comparing the different interpolating methods used to convert the monthly NAO data into a daily time series, one can see that the differences in RMSE are very small. However, it appears as though the coarsest scheme, namely the step function, might be scoring best. This is most likely due to the fact that all other interpolations lead to a different monthly mean of the NAO index than given by the measurement time series, thus distorting the NAO information.

The RMSE rewards those forecasts that are closest on average to the true first day of frost. However, a RMSE of roughly one month is a rather poor forecast overall. As before, we will

forecast scheme	no ΔT	$\Delta T^{(3)}$	$\Delta T^{(3w,10)}$	$\Delta T^{(10)}$
no NAOI, improved time series	7.1	7.2	7.2	7.1
no NAOI, original time series	6.0	6.7	6.6	6.9
$\text{NAOI}_{\text{step}}$	6.0	6.8	6.4	7.0
$\text{NAOI}_{\text{linear}}$	6.0	6.8	6.5	6.9
$\text{NAOI}_{\text{cubic}}$	6.0	6.8	6.5	7.0
$\text{NAOI}_{\text{spline}}$	6.0	6.8	6.5	6.9

Table 6.3: PC in % for deterministic forecasts of the first day of frost using different combinations of information about the initial conditions.

forecast scheme	no ΔT	$\Delta T^{(3)}$	$\Delta T^{(3w,10)}$	$\Delta T^{(10)}$
no NAOI	7.1	7.2	7.2	7.1
$\text{NAOI}_{\text{step}}$	6.6	7.3	7.0	7.2
$\text{NAOI}_{\text{linear}}$	6.7	7.2	7.0	7.1
$\text{NAOI}_{\text{cubic}}$	6.7	7.2	7.0	7.2
$\text{NAOI}_{\text{spline}}$	6.7	7.2	7.0	7.2

Table 6.4: PC in % for deterministic forecasts of the first day of frost using different combinations of information about the initial conditions. The initial anomalies were gathered from the time series with improved stationarity.

therefore also evaluate how often the forecasts were exactly accurate, i.e. how large is the proportion of correct forecasts of the day of first frost?

To do this, we change the score to the proportion correct (PC). Additionally, we do not use the mean values of the conditional probability distribution estimates as the forecast day, but rather again the date on which in the training set frost had already occurred in 10% of all cases. This had proved to be a better method when considering the PC of a forecast. In this case, the anomaly time series with improved stationarity as analysed in Chapter 5 actually proved better than the original simpler version. Table 6.3 shows the percentage of correctly forecast days of first frost for the different forecasting schemes. The benchmark PC was 3.4%, the standard error of the mean proportion correct was 0.2% for all forecasts.

As can be seen, no forecast containing the NAO index outperforms the forecast without the NAOI but containing the improved anomaly time series. Indeed, within the error bars, the forecasts containing the NAO index are indistinguishable from the previous forecasts using the original anomaly time series. Moreover, the different interpolation methods for the NAO index are also indistinguishable. The gain of adding the NAO index is therefore not so clear.

Since the anomaly time series with improved stationarity outperformed all other forecast attempts in terms of proportion correct, in this case the use of the improved time series in conjunction with the NAO index should also be analysed. As Table 6.4 shows, the same conclusions as for the original time series are valid: The additional use of the NAO index does not improve the PC of the deterministic forecasts and the choice of interpolation scheme does not change the outcome in any significant way. The PC is only influenced by the choice of initial anomaly time series, where the improved stationarity leads to an improved forecast quality.

The additional conditioning on the NAO index did therefore not truly help improving the deterministic forecasts of the next day of frost.

forecast scheme	no ΔT	$\Delta T^{(3)}$	$\Delta T^{(3w,10)}$	$\Delta T^{(10)}$
original anomalies	0.139 ± 0.003 (0%)	0.134 ± 0.003 (3.3%)	0.136 ± 0.004 (2.1%)	0.134 ± 0.003 (3.4%)
improved anomalies	-	0.139 ± 0.004 (0.3%)	0.140 ± 0.004 (−0.8%)	0.138 ± 0.004 (0.6%)
$\mathrm{NAOI}^{(2)}_{\mathrm{step}}$	0.132 ± 0.003 (5.3%)	0.130 ± 0.003 (6.7%)	0.131 ± 0.003 (5.9%)	0.130 ± 0.003 (6.8%)
$\mathrm{NAOI}^{(2)}_{\mathrm{linear}}$	0.133 ± 0.004 (4.3%)	0.131 ± 0.003 (5.9%)	0.132 ± 0.003 (5.2%)	0.131 ± 0.003 (6.1%)
$\mathrm{NAOI}^{(2)}_{\mathrm{cubic}}$	0.132 ± 0.003 (5.1%)	0.130 ± 0.003 (6.8%)	0.131 ± 0.003 (6.1%)	0.129 ± 0.003 (7.0%)
$\mathrm{NAOI}^{(2)}_{\mathrm{spline}}$	0.132 ± 0.003 (4.9%)	0.130 ± 0.003 (6.5%)	0.131 ± 0.003 (5.8%)	0.130 ± 0.003 (6.7%)

Table 6.5: Brier score (Brier skill score) applied to initial dates $t_0 \in [70, 123]$ for different forecasting schemes obtained from the anomaly time series with improved stationarity, the original anomaly time series and the original time series with additional conditioning on the NAO index. The errors were determined using a bootstrap procedure with 1000 samples.

6.4.2 Binary forecasts

For initial dates t_0 in spring, an interesting forecasting quantity is the probability of still observing frost before summer, i.e. the weight of the first peak in the bimodal distribution. Here, we chose to use the forecast obtained from $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0)$ as benchmark. Its mean Brier score over the whole period of applicable initial dates, i.e. $t_0 \in [70, 123]$ which lead to the bimodality of the distribution, was $\mathrm{BS} = 0.139 \pm 0.003$ (see Table 4.2).

Table 6.5 shows the Brier scores for the other forecasting schemes, as well as the corresponding skill scores that measure the improvement of the forecasts over the benchmark. As can be seen, the additional use of the NAO index leads to a statistically significant improvement of the forecast skill of up to 7% when compared to the benchmark and certainly outperforms our previous efforts. Even discarding any information on the initial temperature anomalies leads to skillful forecasts. Considering that individual error bars for the scores are a more conservative estimate of statistical significance than confidence intervals for the score difference[75, 76], this should even be a significant result as well.

Interestingly, the best forecast scores are achieved with the largest number of categories for the initial conditions, namely two for the NAO index and ten for the temperature anomalies. This indicates that, while fluctuations are rather high, the number of data points in each category seems to be sufficient not to detract skill in this most coarse-grained of our prediction targets.

Comparing the different interpolation schemes converting the monthly NAO time series into a daily one, we can see that the differences are very small. However, the cubic interpolation performs better or equally to the others, with the simplest interpolation using a step function a close second except in the absence of initial temperature anomaly information, when it is actually best. Both the linear and the spline interpolation are somewhat worse, even though none of these differences are statistically significant considering the magnitude of the errors.

The improvement in the Brier skill score over the previous forecast schemes has shown that the NAO index does indeed have a significant impact on the morning temperatures measured at the Potsdam station.

Considering the variation of the Brier skill score (BSS) with the initial date t_0, we can see in Fig. 6.9 that the NAO index increases the forecast skill especially in the first half of April when fluctuations are small. Then, the NAO index adds significant resolution.

As we could see in earlier forecasts, separating the initial anomalies into weighted terciles could be a more successful scheme than using exact terciles. Analysing the influence of the size of the outer initial anomaly bands p on the Brier skill score, we can see in Fig. 6.10 that in this case, exact terciles seem to be close to the optimum, with the local maxima in the BSS

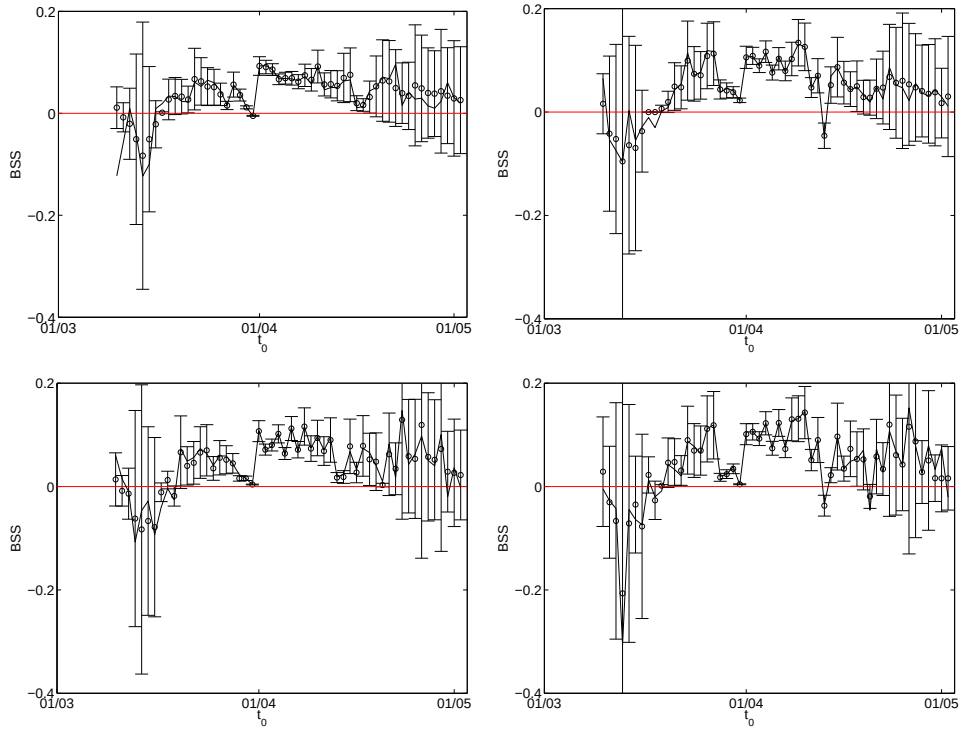


Figure 6.9: Brier skill score comparing the forecasts using $P_{\text{data}}(t_{\text{frost}}|t_0, \text{NAOI}^{(2)}_{\text{step}})$ (top left), $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3)}, \text{NAOI}^{(2)}_{\text{step}})$ (top right), $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3w,10)}, \text{NAOI}^{(2)}_{\text{step}})$ (bottom left) and $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(10)}, \text{NAOI}^{(2)}_{\text{step}})$ (bottom right) to the benchmark. The error bars were obtained using twice the standard deviation of 1000 bootstrap samples, the horizontal red line denotes no forecast improvement over the benchmark.

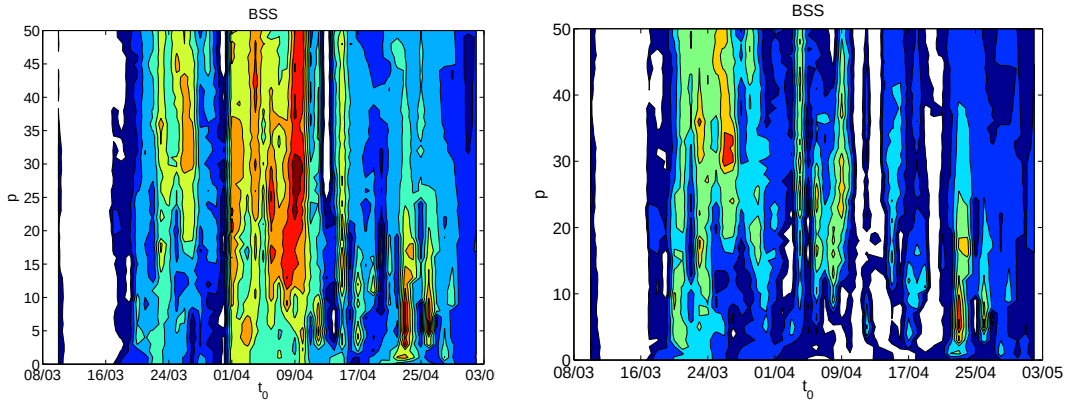


Figure 6.10: Contour plot of the Brier skill score comparing the binary forecast of frost still occurring before summer using $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T^{(3w,p)}, \text{NAOI}^{(2)}_{\text{step}})$ to the benchmark. The right panel again shows the previous results without using information on the NAO for comparison (see Fig. 4.19). For both figures, white denotes the region of negative skill scores, the other colours show increasing skill scores in increments of 2% from dark blue for BSS $\in [0\%, 2\%]$ to dark red for BSS $\in [12\%, 14\%]$.

all touching the region with slightly less weight in the outer bands (around 30% of all initial anomalies in each outer band, i.e. $\Delta T^{(3w,30)}$). The maximum in skill is reached around the beginning of April for $\Delta T^{(3w,27)}$.

Comparing these results to those obtained previously when using only the information about the initial anomalies, it is immediately visible that the regions with skill scores exceeding 10% cover much longer time intervals than before. Moreover, the regions of negative skill have almost all vanished, except around mid-March. Overall the forecast has indeed been significantly improved by incorporating the NAO index additionally.

6.4.3 Full probability prediction

Finally, we will consider the full probabilistic forecast of the next day of frost with a maximal resolved first passage time $t_{\max}$. Fig. 6.11 shows that the improvement over the benchmark forecast has been increased from around 7.5% for the best fraction of anomalies p in the outer bands of $\Delta T^{(3w,p)}$ to almost 9.5% when additionally incorporating the NAO index with a step function interpolation into the prediction procedure[4]. Interestingly, as for the binary forecast, here the weighted terciles with 30% of anomalies in each outer band again score best.

Looking at the change of the skill score across the initial dates t_0 as displayed in Fig. 6.12, we can see that compared to the influence of the initial date, the slight variations in the RPSS with p are almost negligible, again resulting in almost vertical bands of score values. Of course, this is mostly due to the strong coarse-graining of the score values in this figure. Comparing the results with and without the additional conditioning on the NAO index, we can see a larger prevalence of skill scores that exceed 10%, thus also showing that for the full probabilistic forecast of the next date of frost, the NAO index is a helpful addition to the prediction scheme.

Finally analysing the dependence of the RPSS on the maximal fully resolved first passage time, it is evident that the NAO index not only increases the skill score overall, it also leads to an asymptotic skill score that is higher than without the NAO index. The additional variable therefore also helps when one needs a fully detailed forecast, even though it is based on information that was only monthly to begin with.

[4]The other interpolation schemes produce very similar results with ranked probability skill scores that are slightly worse than with the step function but markedly better than without incorporating the NAO index.

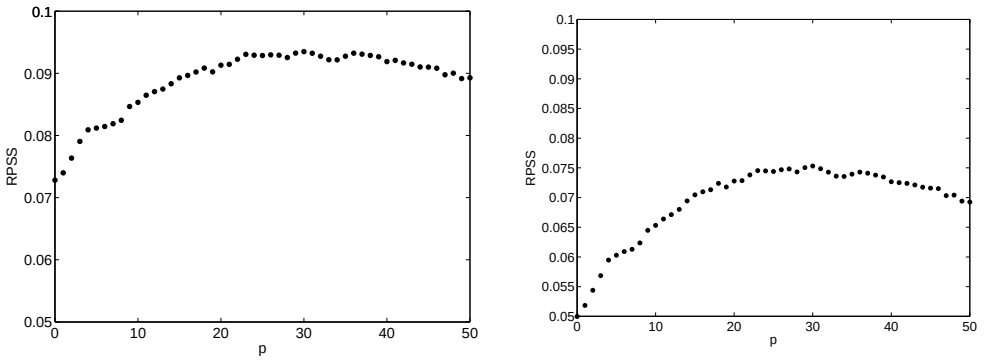


Figure 6.11: Ranked probability skill score using $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(3w,p)}, \mathrm{NAOI}^{(2)}_{\mathrm{step}})$ (left panel) and $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(3w,p)})$ (right panel) with first passage times fully resolved up to 30 days.

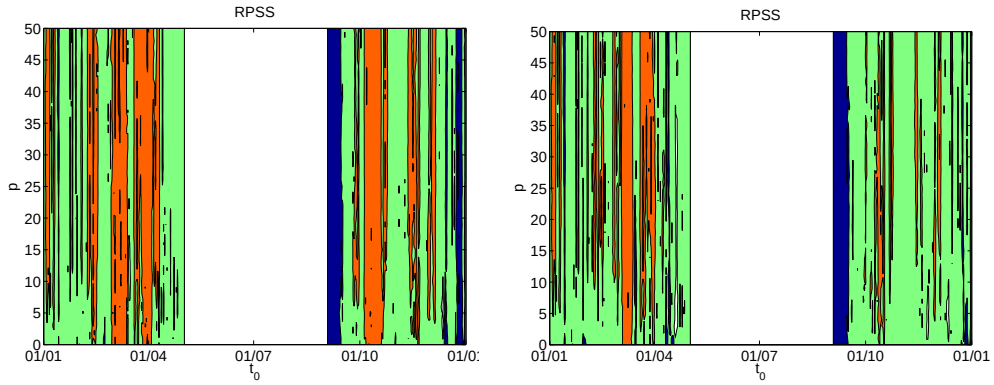


Figure 6.12: Contour plot of the ranked probability skill score using $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(3w,p)}, \mathrm{NAOI}^{(2)}_{\mathrm{step}})$ (left panel) and $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(3w,p)})$ (right panel). White denotes perfect forecast and benchmark, dark blue denotes forecasts worse than the benchmark, green denotes forecasts better than the benchmark by less than 10% and red those better by more than 10%.

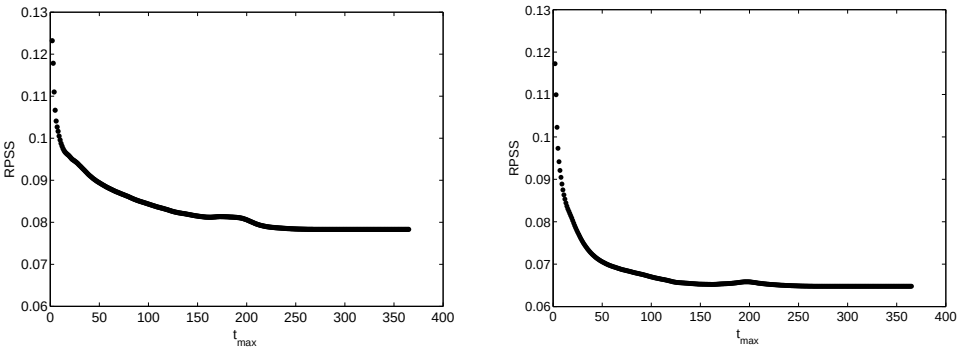


Figure 6.13: Ranked probability skill score for the forecast using $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(3w,30)}, \mathrm{NAOI}^{(2)}_{\mathrm{step}})$ (left panel) and $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(3w,30)})$ (right panel) for different maximal first passage times t_{max}.

	$\mathrm{NAOI}_{\mathrm{step}}$	$\mathrm{NAOI}_{\mathrm{linear}}$	$\mathrm{NAOI}_{\mathrm{cubic}}$	$\mathrm{NAOI}_{\mathrm{spline}}$
$\mathrm{NAOI}^{(2)}$	30.6	30.8	30.7	30.7
$\mathrm{NAOI}^{(3)}$	30.7	30.4	30.4	30.4
$\mathrm{NAOI}^{(10)}$	31.0	30.4	30.5	30.6
benchmark	31.1			
best otherwise	30.4 [using $P_{\mathrm{data}}(t_{\mathrm{frost}}\|t_0, \Delta T^{(3)}, \mathrm{NAOI}^{(2)}_{\mathrm{step}})$]			

Table 6.6: Root mean square error of the deterministic prediction of the first date of frost using the mean of $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \mathrm{NAOI})$, as well as the RMSE for the benchmark and the best result when incorporating also initial anomalies for comparison.

	$\mathrm{NAOI}_{\mathrm{step}}$	$\mathrm{NAOI}_{\mathrm{linear}}$	$\mathrm{NAOI}_{\mathrm{cubic}}$	$\mathrm{NAOI}_{\mathrm{spline}}$
$\mathrm{NAOI}^{(2)}$	6.0	6.0	6.0	6.0
$\mathrm{NAOI}^{(3)}$	6.2	6.4	6.4	6.3
$\mathrm{NAOI}^{(10)}$	6.4	6.4	6.4	6.5
benchmark	3.4 ± 0.2			
best otherwise	7.2 [using $P_{\mathrm{data,improved}}(t_{\mathrm{frost}}\|t_0, \Delta T^{(3)})$]			

Table 6.7: Proportion correct of the deterministic prediction of the first date of frost using the date t_{forecast} with $P_{\mathrm{data}}(t_{\mathrm{frost}} \leq t_{\mathrm{forecast}}|t_0, \mathrm{NAOI}) = 0.1$, as well as the proportion correct for the benchmark and the best result when incorporating also ΔT for comparison.

	$\mathrm{NAOI}_{\mathrm{step}}$	$\mathrm{NAOI}_{\mathrm{linear}}$	$\mathrm{NAOI}_{\mathrm{cubic}}$	$\mathrm{NAOI}_{\mathrm{spline}}$
$\mathrm{NAOI}^{(2)}$	0.132	0.133	0.132	0.133
$\mathrm{NAOI}^{(3)}$	0.132	0.129	0.129	0.129
$\mathrm{NAOI}^{(10)}$	0.134	0.129	0.129	0.130
benchmark	0.139 ± 0.003			
best otherwise	0.129 ± 0.003 [using $P_{\mathrm{data}}(t_{\mathrm{frost}}\|t_0, \Delta T^{(10)}, \mathrm{NAOI}^{(2)}_{\mathrm{cubic}})$]			

Table 6.8: Brier score of the binary prediction of frost still before summer using $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \mathrm{NAOI})$, as well as the Brier score for the benchmark and the best result when incorporating also initial anomalies for comparison.

	$\mathrm{NAOI}_{\mathrm{step}}$	$\mathrm{NAOI}_{\mathrm{linear}}$	$\mathrm{NAOI}_{\mathrm{cubic}}$	$\mathrm{NAOI}_{\mathrm{spline}}$
$\mathrm{NAOI}^{(2)}$	0.0643	0.0646	0.0645	0.0645
$\mathrm{NAOI}^{(3)}$	0.0647	0.0641	0.0643	0.0644
$\mathrm{NAOI}^{(10)}$	0.0650	0.0647	0.0645	0.0646
benchmark	0.0694			
best otherwise	0.0629 [using $P_{\mathrm{data}}(t_{\mathrm{frost}}\|t_0, \Delta T^{(3)}, \mathrm{NAOI}^{(2)}_{\mathrm{step}})$]			

Table 6.9: Ranked probability score of the probabilistic prediction of the first date of frost using $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \mathrm{NAOI})$ with first passage times fully resolved for the first 30 days, as well as the RPS for the benchmark and the best result when incorporating also ΔT for comparison.

6.5 Using a finer subdivision of the NAO index

As seen in Sec. 6.4.2, it appears as though in some situations, the additional information on the NAO index has a much larger influence on the forecast skill than the initial temperature anomalies. Since we only used a very coarse-grained way of incorporating the NAO information, namely only positive versus negative index, changing this to terciles or even deciles of the NAO index might further improve the resulting forecast quality. In order to retain a sufficient amount of data points in each category, we will in the following only condition the first passage time distribution on the NAO index and not on the initial temperature anomalies.

Tables 6.6 and 6.7 show the forecast quality using different coarse-graining and interpolation schemes for the NAO index for deterministic forecasts of the first day of frost. As can be seen, using terciles or even deciles for the index is indeed a further improvement for all interpolation schemes except for the step function when considering the root mean square error (RMSE). The forecast quality in terms of RMSE obtained previously using both ΔT and NAOI can be reproduced by leaving out the initial temperature anomalies.

Considering instead the proportion correct of the deterministic forecasts, no scheme containing the original temperature anomaly time series and a version of the NAO index that we analysed could outperform the anomaly time series with improved stationarity as analysed in Chapter 5. However, leaving out the condition on the initial temperature anomalies while maintaining the information about the NAO index does not change the prediction quality much when considering the original anomaly time series. Using only the less coarse-grained NAO information instead of both the NAOI and the initial anomalies is therefore an improvement in forecast speed and simplicity without a corresponding loss in the quality when considering deterministic forecasts.

Moving on to the binary forecasts of frost still before summer, Sec. 6.4.2 showed that the influence of the NAO index is much larger than that of the initial temperature anomalies. Table 6.8 confirms this and shows that the forecast quality can be matched and even augmented to BSS = 7% by using only information on the NAOI and changing its coarse-graining to terciles or deciles and using a linear interpolation instead of a step function. While this is not a significant improvement in quality considering the magnitude of the errors, it is still an improvement in simplicity.

Tables 6.6 to 6.9 generally show that when using more than two different categories for the initial NAO index, more sophisticated interpolation methods than the step function result in better forecast quality.

For the full probabilistic forecasts, we can see in Table 6.9 that conditioning both on the NAO and on the initial anomalies performs best. Using only one of the two variables to condition the first passage time probability distribution on leads to slightly worse forecasts, where the forecast quality is independent of the variable that was used. Here, the finer subdivision of the NAO index into terciles again constitutes a slight improvement of the forecast quality for the more sophisticated interpolation schemes.

From this analysis, it appears as though a subdivision of the NAO index into terciles in general improves the forecasts. However, as the drop in quality for the full probabilistic forecast when changing from terciles to deciles already indicates, the number of data points remaining in each category needs to be kept large enough even when adding the initial temperature anomaly information to avoid a decrease in skill due to fluctuations, precluding too large a subdivision.

One surprising result of this section is that using the NAO index as the only input information on the initial condition for the predictions does not detract forecast skill in every case, even though the NAO is not a variable measured locally close to the Potsdam station. For deterministic predictions, it performs similarly to using the NAO index in addition to the original temperature anomaly time series, but results in a quicker and simpler forecasting scheme. For

full probabilistic forecasts, on the other hand, the quality analysis yields the expected result: Using the two input variables is better than only using one which is better than using none. However, for binary forecasts of frost still before summer, using only information about the initial NAO index actually performs just as well if not better than using both variables.

6.6 Conclusion

Even though the forecasts in the previous chapters already had significant quality, there was still room for improvement. In this chapter, we therefore incorporated a second variable into our forecast schemes to further increase the success of our predictions. We chose the North Atlantic Oscillation (NAO) since it is the dominant atmospheric circulation pattern for Europe and therefore a good representation of the general weather state[5]. As it is not well represented by the current global circulation models, its incorporation into a data-based prediction scheme might provide a significant improvement over model-based seasonal temperature forecasts.

The time series of the NAO index only provides monthly mean values. Since the forecasting schemes have a daily resolution, we needed to use interpolation to convert the NAO time series to a daily resolution. There are many different possible interpolation methods and we used four different ones, comparing the resulting forecast quality.

First analysing the predictability effects, i.e. the differences in $P_{\text{data}}(t_{\text{frost}}|t_0, \Delta T, \text{NAOI}^{(2)})$ for different combinations of initial conditions, coarse-graining and interpolation schemes, we found that they were significantly enhanced both in the median and in the standard deviation when compared to the previous case without the NAO index. These differences were statistically significant for more than 14 additional initial calendar days t_0 compared to before. There was therefore a significant improvement in the predictability obtained through incorporation of this second variable.

When issuing actual predictions, we found that in the deterministic case, the root mean square error (RMSE) was systematically lowered by incorporating the NAO index, but considering the errors involved, this finding was not statistically significant. In terms of the proportion correct of the forecasts, using the initial anomalies with improved stationarity still proved better than using the original initial anomaly time series and adding information on the NAO index.

Both for the binary prediction and when issuing full probabilistic forecasts, there was a significant improvement in the Brier score when using the NAO index in addition to the initial anomalies.

For most of these cases, using the step function to interpolate the NAO index to a daily resolution resulted in the best forecasts, only in the binary case, the cubic interpolation performed even better. However, this difference was not statistically significant.

Surprisingly, even though the NAO is not a variable measured locally close to the Potsdam station, using it as the only input to our data-based prediction schemes led to significant forecast skill that matched or even exceeded the results reached with both variables for the binary prediction and for the deterministic prediction evaluated using the RMSE.

The inclusion of the NAO index into the forecast procedure has therefore led to an increased forecast performance especially in the binary and full probabilistic forecasts, despite the need for interpolation resulting from the monthly resolution of the time series. Since the spectrum of the NAO index closely resembled white noise, monthly values undersample the dynamics of the North Atlantic Oscillation. A daily time series of the NAO index values should therefore result in an even further improvement, if it could be obtained.

[5] The NAO is described in this chapter by the normalised pressure difference between the stations of Gibraltar and Reykjavik.

7 Conclusion and outlook

7.1 Conclusion

Current weather forecasts are based on numerical models of $\mathcal{O}(10^8)$ degrees of freedom. As the underlying climate dynamics are chaotic, model and data assimilation errors in this high-dimensional system grow exponentially fast. One consequence is a maximum feasible forecast lead time of currently around 7 days for conventional weather forecasts and up to 14 days for medium-range efforts made for instance at the ECMWF.

Yet agriculture, the energy sector, risk and finance markets and the health sector would profit significantly from accurate specific seasonal forecasts. However, all corresponding efforts made up to date for the extratropics provide only time-averaged forecast outputs and still have extremely low skill despite the highly complex and expensive models they are based on.

We therefore wanted to see whether computationally cheap, low-dimensional prediction efforts that are based purely on the observational time series themselves could provide forecasts of comparable or even better quality. In order to do this, we analysed predictions of the first passage time to frost as a crucial example that is more coarse-grained than current short-term weather prediction but still provides information on the precise timing.

To generate the data-based forecasts, we chose to use a time series of daily Potsdam morning temperature measurements provided by the DWD. It contains the years 1893 to 2010, i.e. 118 years' worth of complete, good quality, and homogeneous data.

We first transformed the measurements into temperature anomalies by subtracting the seasonal cycle. This was done with two different levels of sophistication: First, we only chose to use a sinusoidal model of two frequencies (yearly and twice-yearly) to smooth the average temperatures for each calendar day. To improve the stationarity of the anomaly time series, we then generated another version by first detrending the temperatures and also normalising the variance of the anomalies by their average for each calendar day.

As even such long time series do not provide adequate statistics when stratifying according to the calendar day on which they were recorded, we aggregated the data by considering all points within a gliding window of 3 months' width around the calendar day in question. This procedure needs a high level of stationarity in order to avoid introducing systematic errors. However, even in the detrended and renormalised anomaly time series, both the skewness and the autocorrelation still exhibited seasonality.

In order to reduce such systematic errors, we attempted to discard the aggregating in favour of generating more data points by modeling the time series with an autoregressive process. For the simple temperature anomaly time series, the AR(1)-process provided a reasonable fit at least in winter and summer but not in spring, mostly due to the non-stationarities such a model cannot reproduce. Using the more sophisticated version of the anomalies, an AR(8)-process fit the time series very well, but when comparing both the unconditioned and the conditional first passage time distributions generated by the model and the data, there were systematic differences. We therefore discarded this approach.

Calculating the probability distributions of the first passage time to frost for each calendar day, $P(t_{\text{frost}}|t_0)$, we found three regimes: a roughly exponential decay with the maximum probability of first frost within the next two days for t_0 in winter, a bimodal distribution for t_0 in spring, and a roughly normal distribution for t_0 in summer and autumn.

We then analysed the dependence of the conditional first passage time distributions on the initial temperature anomalies as a first indication of the potential predictability of our data-based schemes. We found a non-trivial dependence on the initial conditions that was enhanced for initial dates in autumn and spring when using the more sophisticated version of the temperature anomalies. A Kolmogorov-Smirnov test confirmed that the differences in the conditional first passage time distributions for initial temperature anomalies in the low and high terciles were significant for initial dates in the winter half of the year and in June. However, the more sophisticated version of the temperature anomalies actually showed less calendar days for which the differences were significant than the simpler one.

To further improve the predictability, we added a second measured time series to represent the large-scale atmospheric weather regimes, namely the North Atlantic Oscillation (NAO) index that is based on the sea-level atmospheric pressure differences between Reykjavik and Gibraltar. Using the NAO index as further input considerably enhanced the differences in the conditional first passage time distributions and extended the period over which they were statistically significant, improving the potential predictability.

After these preliminary analyses, we proceeded to issue data-driven out-of-sample forecasts of the first passage times to frost with the previously studied varying degrees of sophistication and evaluated the forecast skill. We started by addressing deterministic predictions of the next date of frost. An evaluation of the root mean square error measuring the average distance between the forecast date and the verification showed that while the forecast always performed much better than the benchmark, the error was quite large overall. The more sophisticated temperature anomalies led to worse forecasts than those obtained from the simpler version, while using the NAO index as input either on its own or in addition to the simple temperature anomalies performed best.

We additionally evaluated the deterministic forecasts using the proportion correct to focus on the cases in which the forecasts were actually perfect to reduce the influence of large deviations from the verification that occurred regularly in spring when the underlying first passage time distribution is bimodal. In this aspect, all the issued forecasts were again better than the benchmark, but here the more sophisticated temperature anomaly time series performed better. Even incorporating the NAO index did not improve the forecasts further.

With both these evaluations, the deterministic predictions issued in autumn performed best, while the forecast quality was worst in spring. Then, a different and more coarse-grained forecast target is appropriate, namely the total probability of still observing frost before summer. We showed that this probability is heavily influenced by the initial conditions. Evaluating the corresponding forecasts using the Brier score, we found that only between March 20^{th} and April 18^{th} they were better than the benchmark. Here again, the more sophisticated temperature anomaly time series actually led to worse forecasts except for very few initial days in April, while adding the NAO index significantly improved the predictions. Even using only the NAO index information as input reached similarly high skill levels of around 7%.

The final forecast target was a fully probabilistic prediction of the next date of frost with daily resolution for the first 30 days into the future. Our forecasts beat the benchmark by around 7.5% when using only the simple temperature anomalies, with the least well-predicted initial dates in autumn. The quality became worse when relying on the more sophisticated initial temperature anomalies but the NAO index provided a sizable improvement over the benchmark of up to 9.4%.

Table 7.1 summarises the best data-based prediction schemes for each case and the specific scores obtained using them. As can be seen, the more sophisticated temperature anomalies generally performed worse except for the proportion correct of the deterministic forecasts, despite the improved stationarity of the time series which should have reduced the systematic errors introduced when coarse-graining the initial condition on t_0. This is most probably due

	forecast score	benchmark score	skill score	distribution used	input variables
RMSE	30.4 days	73.5 days	58.6%	$P(t_{\text{frost}}\|t_0, \text{NAOI}^{(3)}_{\text{linear}})$	NAO index
PC	7.2%	3.4%	3.9%	$P(t_{\text{frost}}\|t_0, y, \Delta T^{(3)})$	sophisticated T anomalies
BS	0.129	0.139	7.2%	$P(t_{\text{frost}}\|t_0, \text{NAOI}^{(3)}_{\text{linear}})$	NAO index
RPS	0.0629	0.0694	9.4%	$P(t_{\text{frost}}\|t_0, \Delta T^{(3)}, \text{NAOI}^{(2)}_{\text{step}})$	NAO index, simple T anomalies

Table 7.1: Root mean square error (RMSE), proportion correct (PC), Brier score (BS) and ranked probability score (RPS) of the best data-based prediction schemes, as well as the benchmark score and the corresponding skill scores.

to the fact that the variability between the calendar years in the temperature time series was much larger than the trend so that the detrending could only provide a small contribution while adding a further source of errors. Moreover, deseasoning the variance left the non-stationarities in the higher moments of the temperature anomaly distribution unchanged.

The NAO index also appears to have a much larger influence on the results than the initial temperature anomalies, especially in spring. This is probably due to the persistence of the sea surface temperature anomalies which influence the NAO index and whose state therefore carries more long-term information than the land surface temperatures themselves.

In this thesis, we have managed to issue purely data-driven forecasts of the first passage time to frost that proved more skillful than the corresponding benchmark even on longer lead times than those usually provided by the operational weather forecasts. Even though we calculated the forecast skill scores, a direct comparison with the corresponding skill of the dynamical models is still missing[1]. A first comparison will be given in the next section.

7.2 Comparison with a dynamical model ensemble forecast

One fundamental problem for an objective comparison between data-based and model-based prediction quality is the continued evolution and changes of the dynamical models that lead to a non-stationary forecast skill. In order to base a forecast skill comparison on a sufficient amount of data to ensure robust statistics, we therefore chose to use the National Center for Environmental Prediction (NCEP) Medium Range Forecast output from the 1998 version of their global atmospheric circulation model, for which a long reforecast was done by the National Oceanic and Atmospheric Administration (NOAA) from 1979 to the present[141]. While there exist newer models with better quality, the NOAA data has the advantage of length and easy availability.

Similar to other global dynamical models it uses the fluid dynamical equations of atmospheric motion to generate possible time evolution scenarios on a $2.5^\circ \times 2.5^\circ$ grid with 28 vertical levels. The model is initialised daily at 0:00 UTC and provides forecasts every twelve hours for up to 15 days into the future using 15 ensemble members. These are generated through one control forecast initialised with the assimilated data provided by the NOAA reanalysis and seven pairs of varied initial conditions using Bred perturbations[2].

[1] While one can look up specific scores and their range for the operational seasonal models, such as BSS $\leq 5\%$ for temperatures in Europe[25] or BSS $\in [-5\%, 4\%]$ for the multimodel temperature forecasts at the ECMWF[22], those forecasts address different targets than this thesis. As the base rates of the forecast targets are not equal, any differences in skill scores are directly reflecting the difference in base rates rather than the difference in actual forecast skill.

[2] For more detail on the reforecast see [141].

Since the global model resolution is very coarse - the grid point distance in mid-latitudes is around 200km - there is no grid point located at Potsdam, the station for which the previous data forecasts in this thesis were generated. To avoid artefacts in the comparison that result exclusively from the disparity in geographic location between model and data or from an attempt at model output interpolation to a different location, we will issue new data-based forecasts using temperature measurements made at Hannover Langenhagen airport. It is situated at 52°26'N 9°44'E, with the next model grid point at 52°30'N and 10°00'E, roughly 20km away.

As the model output is generated for 0:00 UTC and 12:00 UTC, we will use the noon measurements of the temperature instead of the morning ones as in the previous forecasts for this comparison. The minimum forecast lead time is therefore 24 hours for the data-based predictions and 36 hours for the model-based predictions as they are initialised only once per day at 0:00 UTC. The difference generously accounts for the time lag between measuring the initial data and obtaining the model-based forecast results after the full data assimilation and model computation procedure.

While this is not quite the same prediction task as in the main parts of this thesis - the morning measurements are close to the daily minimum temperature especially in winter, while the noon measurements should be close to the daily maximum leading to a different base rate of frost events - the Hannover noon temperature time series offers the closest possible comparison to a dynamical model.

We repeated the detailed analysis as described previously in Sections 3.2 and 5.2 also for the noon measurements made at the Hannover station and obtained from the DWD[1] to confirm that they are of similar quality and homogeneity as the Potsdam morning temperatures (results not detailed here). We then proceeded to construct the two different temperature anomaly time series as in the main part of the thesis. Finally, the probabilistic forecasts of the first passage time to frost were calculated as before with $t_{\max} = 15$ days, i.e. with fully resolved frost probabilities for the first 14 days into the future.

For the model-based predictions, first the climatology was subtracted from the ensemble members to convert them into temperature anomaly forecasts. Then, the fraction of ensemble members falling below 0 °C was counted for the first 14 days into the future, with the 15^{th} category containing the residual probability weight. This gave rise to a raw ensemble forecast.

However, in operational dynamical weather forecasting, the ensemble members routinely undergo post-processing, called model output statistics. In fact, the most notorious error of such forecast ensembles that is also present in the NOAA ensemble is a systematic bias. It is given by the average distance between the forecast ensemble mean and the corresponding verification and shows a seasonal dependence. For the comparison, this bias was calculated for the two previous forecast years and then subtracted from the current prediction year to keep all forecasts out-of-sample. This procedure led to a debiased ensemble forecast.

In the following, all raw and post-processed first passage time predictions to frost based on the NOAA forecast ensemble were generated by Stefan Siegert.

Constructing the benchmark forecast in the same way as before (see Sec. 4.3.3), we used the ranked probability score (RPS) and its corresponding skill score to again evaluate the forecasts.

The resulting scores for the model-based and a selection of data-based first passage time to frost forecasts are listed in Table 7.2. All different forecast schemes - not only those shown in the table - performed significantly better than the benchmark. While data-based forecasts that use only one input variable, i.e. either initial temperature anomalies or the NAO index, mostly perform worse than even the raw model ensemble, one can outperform the dynamical model and achieve skill scores of up to 21% by relying on both input variables. However, additional statistical post-processing on the model output proves to be the best strategy overall if one resolves only the first two weeks into the future. Beyond this lead time, the model does not offer any output, while the data-based predictions still show significant skill (see Fig. 6.13 for

forecast scheme	RPS	RPSS [%]	comments
debiased model ensemble	0.0445	24.6	
$P(t_{\text{frost}}\|t_0, \Delta T^{(10)}, \text{NAOI}^{(2)}_{\text{step}})$	0.0467	20.9	best data-based forecast
$P(t_{\text{frost}}\|t_0, \Delta T^{(3)}, \text{NAOI}^{(2)}_{\text{step}})$	0.0473	19.9	data-based forecast expected to be best (Chapter 6)
$P(t_{\text{frost}}\|t_0, \Delta T^{(3w,10)})$	0.0475	19.6	best forecast without using the NAO index
$P(t_{\text{frost}}\|t_0, y, \Delta T^{(10)}, \text{NAOI}^{(2)}_{\text{step}})$	0.0476	19.5	best forecast using more sophisticated T anomalies
raw model ensemble	0.0479	18.9	
$P(t_{\text{frost}}\|t_0, y, \Delta T^{(3w,10)})$	0.0495	16.2	best forecast using sophisticated T anomalies but no NAO
$P(t_{\text{frost}}\|t_0, \text{NAOI}^{(2)}_{\text{step}})$	0.0503	14.9	best forecast using no initial T anomalies
benchmark	0.0590	0	

Table 7.2: RPS and corresponding skill score for the benchmark, for several data-based forecasts with different levels of sophistication, and for two model-based forecasts with and without post-processing for the Hannover noon temperatures. All predictions were verified on their common forecast period only, namely on the years 1981 to 2010.

the Potsdam morning temperatures).

Interestingly, contrary to our expectations after the analysis of the Potsdam morning temperatures in the earlier chapters, here a separation of the initial temperature anomalies into deciles leads to a better score than the coarser separation into terciles. This is most probably due to the different frost base rate in the noon temperatures. Now, 92.7% of the temperature data points lie above the threshold of 0 °C and can therefore be used as initial conditions for the forecast generation, while in the case of the Potsdam morning temperatures, only 80.2% of all data points were available. Considering that the Hannover measurements only start in 1946, this does still seem to imply a smaller absolute number of data points than in the Potsdam morning case. However, in winter the absolute number of days with temperatures above freezing in the time series is almost equal between the two stations. Especially in early spring the Potsdam morning time series even has less data points with positive temperature than the Hannover noon time series. The statistics are therefore indeed severely impacted by this change in time series, explaining why the less coarse-grained approach might lead to better results.

Even more striking is the difference in score values between the Potsdam morning temperatures and the Hannover noon time series. Indeed, while the best data-based prediction obtained a ranked probability skill score of RPSS= 9.4% in Sec. 6.4.3, we are now contemplating data-based forecasts with ranked probability skill scores as high as RPSS=20.9%. This can also be explained by the different frost base rates between the two time series. Indeed, this directly influences the forecast uncertainty: The smaller the base rate, the smaller the uncertainty as the situation is easier to predict. The score decomposition shows that this results in a smaller RPS and a higher RPSS[89, 86], i.e. better forecast scores, as indeed observed here.

As we have seen, the purely data-based forecasts perform quite well for lead times of only two weeks, where they manage to keep up with purely model-based forecasts even though they involve much less computational effort and costs. They are only beaten by statistically post-processed model output.

There exist, of course, better models and it should be possible to further improve even the

model-based forecasts analysed here by additionally correcting the ensemble underdispersiveness or using more ensemble members. However, the data-based predictions come quite close to the quality provided by the dynamical model on shorter lead times and their skill remains in evidence much further into the future than just the first two weeks that are provided by the dynamical model.

We have therefore shown here that the data-based prediction schemes actually provide an adequate alternative to seeking ever further model refinement if one wants to target specific coarse-grained prediction tasks over longer time scales for practical applications.

7.3 Outlook

While the data-based prediction schemes have therefore shown promising skill, there is always room for further improvement.

The most obvious weakness up to date lies in the specific way the NAO index is currently incorporated into the forecasts. Indeed, since the variations in atmospheric circulation can explain up to 70% of all variations in the surface air temperatures[142], it is well worth expending effort to improve how the information about the current state of the North Atlantic Oscillation is used. The first priority would be to procure good quality daily measurements as Chapter 6 showed that the monthly mean values constitute a badly undersampled time series.

Further improvements to the NAO incorporation might come from a different index better suited to the specific prediction target analysed in this thesis. Previous studies have shown that the normalised pressure difference between Paris and London explains the temperature variations in Berlin quite well[142], while pressure differences between Gibraltar and Bergen (Norway) in winter and between Stockholm (Sweden) and Reykjavik in summer[3] explain the Potsdam temperature variations[96]. As these other possibilities for NAO indices are essentially correlated with the same temperature variations, these studies show that there are several good choices for the NAO index thus offering potential for further optimisation.

Incorporating a third complementary input variable might also help to further improve the forecast quality as only a small fraction of climate variability can be accounted for by a single factor, especially when considering the extratropics[29]. We have already enumerated a large number of variables influencing the European temperatures on seasonal time scales in Chapter 1. Some of them, such as soil moisture, might provide suitable additional information. However, by employing this strategy one needs to be careful not to follow the dynamical model path of increasing complexity thereby negating the specific advantages of data-based predictions in terms of small computational cost.

While generating additional data points using the simplest possible stochastic model, namely an AR(p)-process, failed to reproduce the observed first passage times to frost, this does not mean that there might not be a different simple model that could succeed. One avenue would be a slightly more long-range correlated process with a power spectrum corresponding to that of the temperature anomalies.

A different and to our knowledge new approach to seasonal predictions would be a hybrid model- and data-based forecast scheme that incorporates the model output for the next 14 days into the selection of analogues for the data-based forecast, i.e. for the construction of the conditional first passage time distributions.

In the broader picture, it would also be interesting to look into other prediction targets on seasonal time scales, especially first passage times for different variables, in order to establish just how specific our results are to frost forecasts and whether there are other situations in which data-based predictions could provide a true advantage over model output.

[3]The changing stations across the calendar year reflect the spatial shift of the NAO action centers.

A Appendix: Different possible definitions of the climatology

A.1 Introduction

Due to seasonal solar forcing, temperature records show a seasonal cycle that is the cause of the largest part of the temperature variance throughout a year. Any temperature prediction scheme therefore already scores quite well just by reproducing this cycle that is also called *climatology*. Since the fluctuations around it are usually of more interest, one first needs to determine the annual cycle as accurately as possible in order to subtract it from the measured data. However, there is no universal agreement on the best method to be employed.

As Vecchio and Carbone already pointed out[143], the climatology needs to be suitably defined to lead to a good analysis of different aspects of temperature time series, such as persistence in their case. In fact it has been found that the scaling of temperature power spectra changes significantly with a change in the determination of the annual cycle[124].

The most common defintion of the climatology throughout the literature is the mean temperature for each calendar day, averaged over all years in the time series[123, 32, 133, 34, 124]. There are also some minor variations of this definition such as an additional division by a seasonal standard deviation in order to deseason also the variance of the time series[131]. In some cases, the calculation of the average is restricted to a reference period that is much shorter but common to all time series considered within one study and which does not contain missing data points - often 1961 to 1990[126].

One further possibility is to smooth the sequence of average daily temperatures. Moberg et al.[126] use an 11-term binomial filter or Fourier methods to allow no influence from the particularly steep parts of the seasonal cycle to leak into the season-corrected anomalies. May et al.[123] use the mean temperature and a sinusoidal model of the first six frequencies in the time series with decaying and arbitrarily chosen weights to fit the daily mean values.

The problem with all these ways of calculating the climatology is their lack of validity for non-stationary (or rather non-cyclo-stationary) processes[1]. Since climate is a highly nonlinear system with external forcing and strong evidence towards some sort of climate change (see also Sec. 3.3.1), an assumption of stationarity is rather questionable[143].

Using the Potsdam morning temperatures, we will check in the following how well the different definitions of the climatology are suited to the time series and to our chosen prediction task.

A.2 Mean temperature for each calendar day

A.2.1 Definition and model fit

We start with the simplest definition, where the climatological value of the temperature for a specific date is just the mean value of the temperatures recorded on this same date throughout all the years in the time series.

[1] Of course, in some cases an additional term to represent e.g. the trend or other non-stationarities can be added to ensure validity of the decomposition.

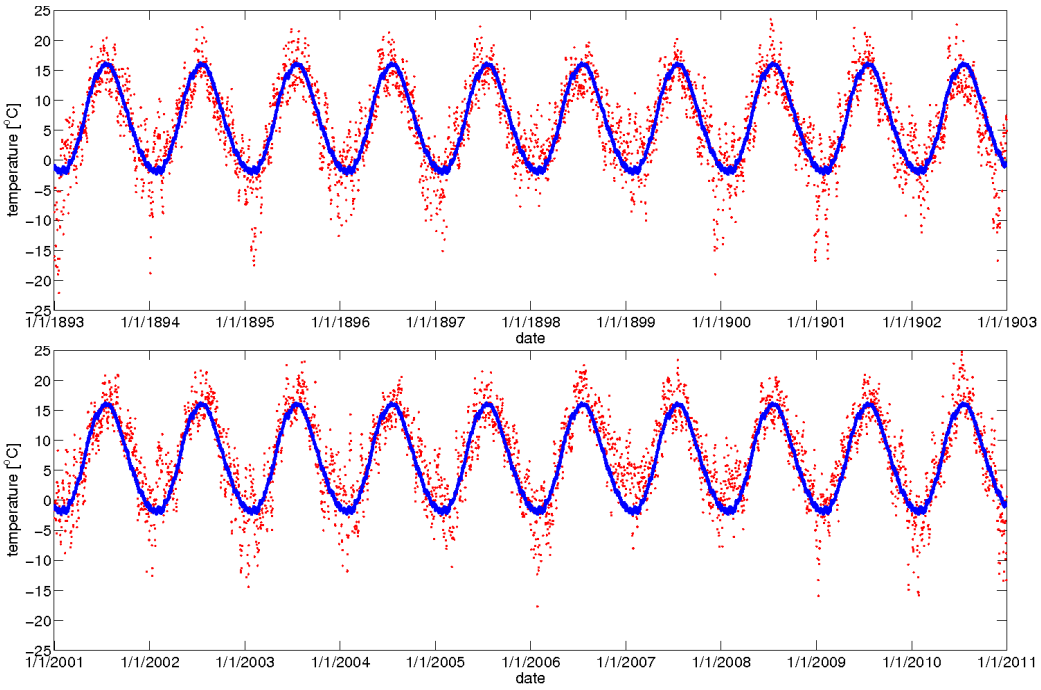


Figure A.1: Potsdam morning temperatures over the course of 10 consecutive years (red dots) and the corresponding mean values for each calendar day, averaged over the whole 118 years of the time series (blue line).

For this case, Fig. A.1 shows that the resulting climatology seems to fit the corresponding data points quite nicely, both for the first and the last ten years of our time series, implying that there is no large trend in the temperatures that would not be represented well by the mean value over the whole time span.

A measure of how well the measured data is represented by this definition of the climatology is the distribution of the residuals after subtracting the climatological value, also called *anomalies*. This is shown in Fig. A.2. As can be seen, the temperature anomalies appear to follow a normal distribution, with the exception of a fat tail for very small values[2]. This leads not only to an asymmetry of the observed range of values but also to a larger probability of small values even within the interval that is assessed on both sides of the range. This means that while the mean for each calendar day represents the temperatures quite well, there is a slight bias towards negative anomalies.

A.2.2 Deseason the variance?

In order to assess whether a deseasoning of the variance would help in this case, we look at the spread of the distribution for different calendar months in Fig. A.3. As can be seen, neither the mean or the median, nor the quartiles change much over the course of the year. The maximum is also reasonably constant, but the minimum changes significantly, thus influencing the variance of the anomalies.

Looking at Fig. A.3, it becomes clear that the anomalies cannot be looked upon as a stationary time series with the previous definition of the climatology. It might therefore indeed be good to deseason also the variance of the anomaly distribution.

[2]For an introduction of normal probability plots see Sec. 2.2.4.

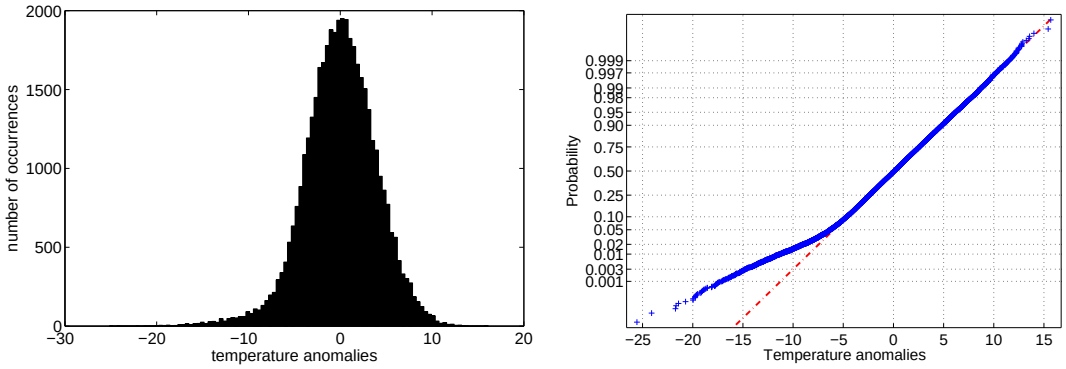


Figure A.2: Histogram and normal probability plot of the deviations from the mean temperatures for each calendar day. The red dashed line in the right panel represents the expected plot for a normal distribution.

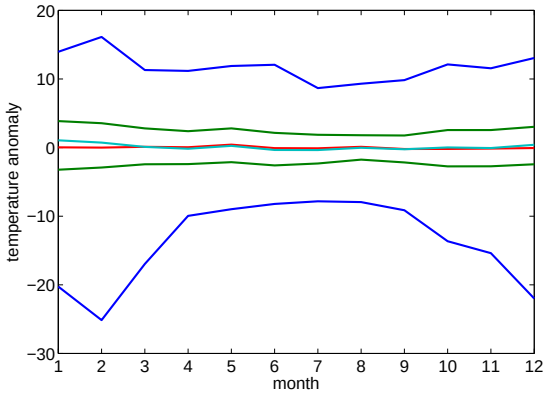


Figure A.3: Extremal values (blue lines), quartiles (green lines), mean (red line) and median (cyan line) of the anomaly distribution.

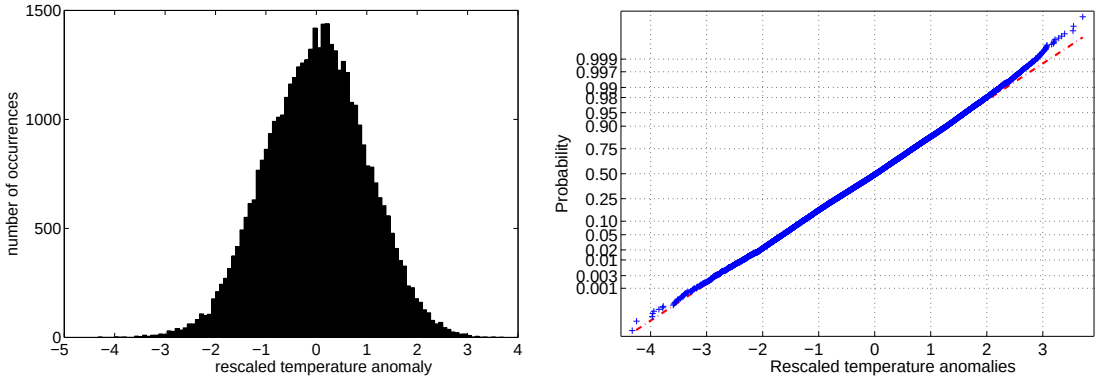


Figure A.4: Histogram and normal probability plot of the rescaled deviations from the climatology. The red dashed line in the right panel represents the expected plot for a normal distribution.

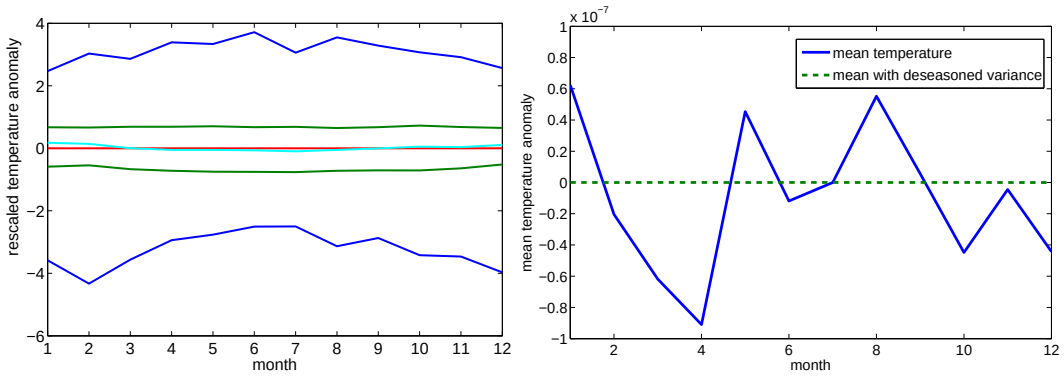


Figure A.5: Left panel: Extremal values (blue lines), quartiles (green lines), mean (red line) and median (cyan line) of the rescaled anomaly distribution. Right panel: Mean temperature anomaly for each calendar month, where the climatology is defined as simply the mean temperature for each day (blue line) and as the rescaled mean (green dashed line).

To this end, we compute anomalies in the following way:

$$\Delta T_i = \frac{T_i - \langle T_{j(i)} \rangle}{\hat{\sigma}_{j(i)}}, \tag{A.1}$$

where ΔT_i denotes the anomaly on day i, $j(i)$ denotes the calendar day corresponding to i, $\langle T_{j(i)} \rangle$ the mean temperature on calendar day $j(i)$, and $\hat{\sigma}_{j(i)}$ the corresponding standard deviation. Now we repeat our distribution analysis of the resulting anomalies.

Fig. A.4 shows that the rescaled anomalies now follow a normal distribution very closely and the left panel of Fig. A.5 illustrates the constant spread. The right panel confirms that the mean temperatures for each month are much better represented by the rescaled climatology, as the residual anomalies are much closer to the zero line. However, the deviations are very small in magnitude even without deseasoning.

Taking a closer look at the left panel of Fig. A.5, it becomes evident that even though the rescaled climatology appears to be much better, the total range of the anomalies still seems to shift to slightly higher values over the summer months. To evaluate how important this shift is in practice, we look at the warmest 1% and the coldest 1% of anomalies and their distribution across the different calendar months. Fig. A.6 shows that almost all extremely cold deviations occur in winter, while the extremely warm deviations occur in summer, making it obvious that even the deseasoned variance does not manage to transform the time series of anomalies into a completely stationary one.

Concluding, one can say that the mean temperature for each calendar day is an adequate representation of the temperatures. The resulting fatter tail for negative extreme anomalies can be eliminated by rescaling the anomalies by the daily standard deviation of the temperatures. However, not even the deseasoning of the variance is enough to lead to an entirely stationary time series: The distribution changes in skewness throughout the year so that despite the constant mean and variance, the probability for extreme deviations from the mean to happen in a fixed direction is not constant over time.

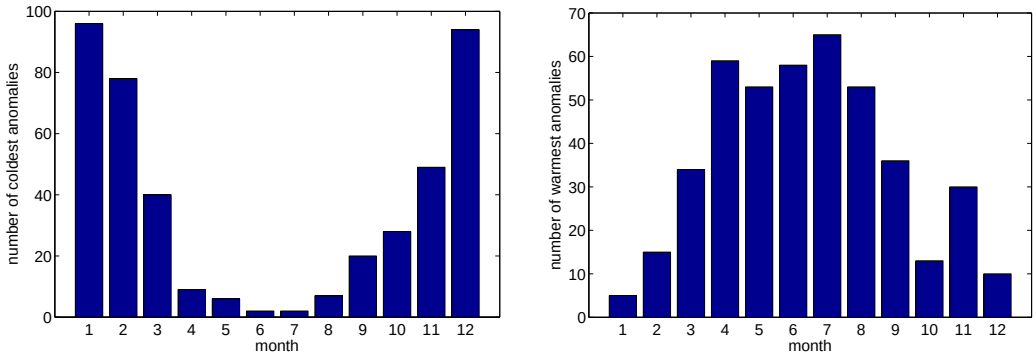


Figure A.6: Histogram of the occurrence of the 1% coldest anomalies across the different calendar months (left) as well as of the 1% warmest anomalies (right).

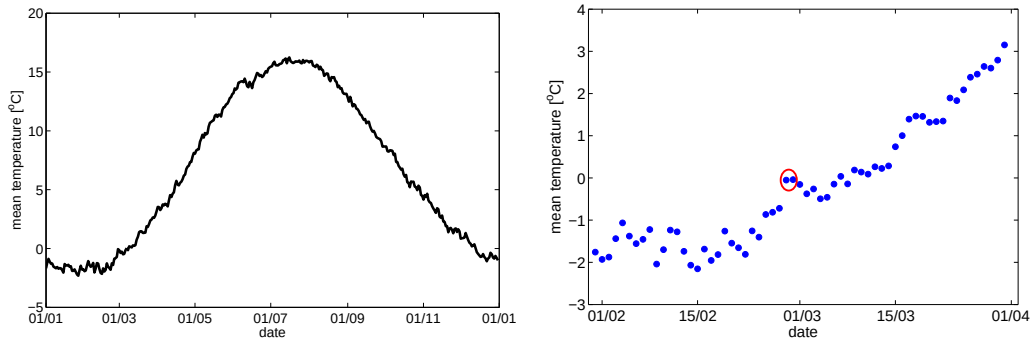


Figure A.7: Mean temperatures calculated for each calendar day separately, as well as a zoom into the months of February and March. The red circle singles out the 28th and 29th of February.

A.3 Smoothing the climatology

A.3.1 How to treat leap years?

As can be seen in the left panel of Fig. A.7, the mean temperatures for each calendar day still form quite a noisy curve, due to the relatively small number of years in our time series in terms of good statistics. To avoid carrying over too many particularities from the specific time series into the climatology, such as outliers exerting too much influence on the mean temperature of their respective calendar day, a simple smoothing model might be applied to the resulting curve. Normally, the whole time series is fitted with such a model. In our case however, we need a unique climatological value for each calendar day irrespective of the year in order to be able to enhance the statistics as explained in Sec. 3.2.2. This then carries an added difficulty: the existence of leap years.

In fact, for any calculation of the average temperature, February 29th simply has a worse statistic than the other calendar days. There are also no other effects from the changing numbering of days in the year visible in the mean temperature and central moments of February 29th and March 1st with and without conditioning on it being a leap year or not. This means that the mean temperature for March 1st does not depend on the existence of February 29th in that year, at least not considering the magnitude of the errors inherent in a time series of only 28 leap and 90 normal years. This is very encouraging.

However, if one were to fit a smooth model to the mean temperatures, the changing support bears thinking about. May et al.[123] simply excluded February 29th from their analysis, but they also only considered annual smoothed cycles and frequency components. If one instead uses the climatology to calculate an anomaly time series for further analysis, assuming 365 days throughout for simplicity will introduce a phase shift between the climatological model and the actual temperature time series of 28 days or almost a whole month over the total length of the time series, while using 366 days leads to three months' shift in total. This would seriously distort the resulting anomaly time series.

We will therefore calculate the smooth model with a support of 366 days and leave out the value for February 29$^{\text{th}}$ except in leap years.

The only thing left to decide now is the specific form of the model.

A.3.2 Which smooth model?

By far the most common smooth climatological model is based on a sum of the global mean of the temperature time series and fitted trigonometric functions with a yearly fundamental frequency and higher harmonics, as given by Eq.(A.2), where $\langle T \rangle$ denotes the mean temperature over the whole time span, j the calendar day and $\omega = \frac{2\pi}{365.2425 \text{ days}}$ the angular yearly fundamental frequency in the Gregorian calendar.[3]

$$\widetilde{T_j} = \langle T \rangle + \sum_{n=1}^{k} a_n \sin(nj\omega) + \sum_{n=1}^{k} b_n \cos(nj\omega) \tag{A.2}$$

There seems to be no consensus in the literature on the total number k of harmonics, which ranges from one to six and is described as 'chosen arbitrarily'[123, 144], if an attempt at justification is made at all. Epstein even claimed that 'there does not appear to be a neat, rigorous method by which to determine unambiguously an optimum number of harmonics to fit any dataset'[144].

He did, however, propose a test statistic to provide some orientation as to the optimal k. It is based on the quantity X_j as given by Eq.(A.3), where N denotes the number of calendar years in the temperature time series, σ_j the standard deviation of the temperatures recorded on calendar day j, and k the number of harmonics used in the smooth climatological model. The X_j should follow a student's t distribution with $N-1$ degrees of freedom.

$$X_j = \frac{(\langle T_j \rangle - \widetilde{T_j}^{(k)})}{\frac{\sigma_j}{\sqrt{N}}} \tag{A.3}$$

If one then computes the counts $n_m, m = 1, 2, ...10$ of the values X_j that fall within the m$^{\text{th}}$ decile of the student's t distribution with $N-1$ degrees of freedom, one would expect a uniform distribution, i.e. $n_m \approx 36.5 \, \forall m$. The corresponding sum of squared errors Y as given by Eq. (A.4) then should follow a χ^2-distribution with nine degrees of freedom, leading to $\mathbb{E}[Y] = 9$.

$$Y = \sum_{m=1}^{10} \frac{(n_m - 36.5)^2}{36.5} \tag{A.4}$$

We will therefore fit the smooth climatological model on a support of 365 days and calculate the quantity Y as well as some other common measures of model accuracy to evaluate the optimum number of harmonics to use for smoothing the climatology.

[3] Of course, as already stated by Epstein[144], there is no guarantee that harmonics are the optimal choice for smoothing the annual cycle.

term	$\langle T \rangle$	ω		2ω		3ω		4ω	
fitting function	constant	sin	cos	sin	cos	sin	cos	sin	cos
parameter value [°C]	6.6	-3.0	-8.4	-0.07	0.69	0.03	0.14	0.19	-0.21
relative error	0.45%	1.3%	0.48%	57%	5.8%	180%	39%	37%	33%
RMSE [°C]	6.3	0.58		0.30		0.29		0.28	
R^2	$-2 \cdot 10^{-16}$	0.9918		0.9977		0.9979		0.9981	
Y	923.52	272.45		41.88		50.15		59.47	

Table A.1: Table listing the first nine parameter values for the linear least squares fit of the climatological model as given by Eq.(A.2), as well as their relative errors corresponding to the 95% confidence level, the root mean square error (RMSE) corresponding to a fit with all parameters up to the one in question, and the corresponding R^2-value, as well as the value of Y as given by Eq.(A.4).

Since a climatological model as defined in Eq.(A.2) is linear in the model parameters, and trigonometric functions with different higher harmonics of the same fundamental frequency form an orthogonal set of basis functions, we can fit the parameters one after the other and add components without changing the estimated values that were obtained previously. This allows us not to fix a cutoff number of harmonics beforehand. Table A.1 shows the resulting parameter values for the first four frequencies, as well as the chosen measures of model accuracy.

As can be seen, the root mean square error (RMSE) of the model fit decreases drastically when incorporating the yearly frequency and further decreases if the twice-yearly frequency is also taken into account, but then there is only a very slight further decrease for additional parameters. This is confirmed by the behaviour of the R^2-value and also by the large relative errors of the higher order parameter values themselves. The Y value has a minimum for two harmonics, but remains much larger than the expected value for perfectly Gaussian anomalies throughout, suggesting that the choice of basis functions might be problematic. On the whole, the criteria seem nevertheless to indicate that a model with yearly and twice-yearly frequencies would be the adequate choice for our purpose.

To check the validity of this choice of harmonics, we will additionally conduct a frequency analysis on the Potsdam morning temperature data, as described in detail in Sec. 2.1.3. Fig. A.8 shows the resulting ensemble average of modified periodograms, both with low variance due to a large number of data segments in the ensemble (left panel) and with low bias due to a choice of much longer - but necessarily few - segments (right panel).

The left panel shows a single large peak that is smeared-out around frequencies of once or twice per year. The large number of short segments means that this is a very smooth estimate, but also that any frequency lines that lie too close together cannot be distinguished anymore. We therefore also need the higher variance estimate in the right panel: It provides a finer frequency resolution and actually reveals two distinct lines, one for yearly frequency and a less important one for twice-yearly influences. All other frequency components seem to be negligible and close to white noise.

The yearly and twice-yearly frequencies are therefore confirmed to be the predominant frequencies in the temperature time series. However, the twice-yearly peak is much smaller than the annual one and temperatures are not expected to have a behaviour as clearly twice-yearly as, for example, precipitation, which has maxima in both spring and autumn. Indeed it has been found that the first harmonic explains already between 89% for maritime stations and 99% for continental stations of the variability of the annual cycle[124]. Moreover, at least the Y value remained very large for all tested numbers of harmonics. We should therefore check further how many frequencies are indeed statistically relevant.

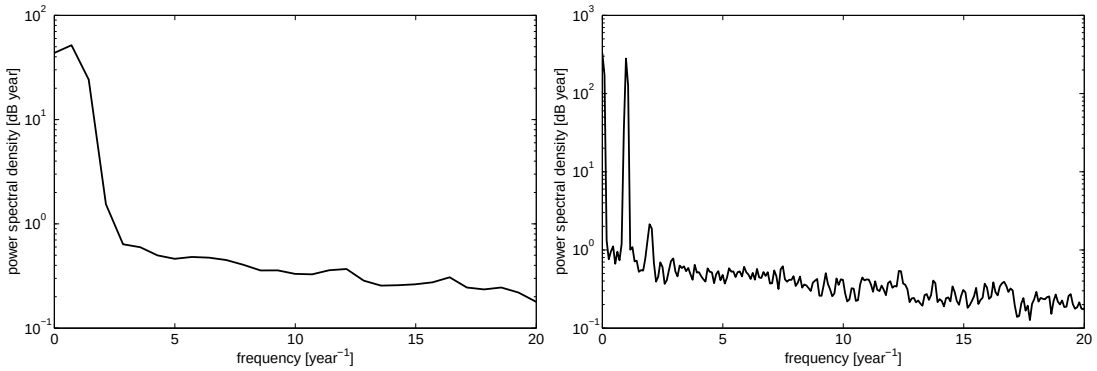


Figure A.8: Power spectral density estimate of the Potsdam morning temperature time series on a semilog scale, obtained using Welch's periodogram method with 167 segments of length 512 data points (left panel) as well as 20 segments of length 4096 data points (right panel), both with a 50% overlap between consecutive segments. For both figures, only the lower end of the frequency bandwidth is shown here.

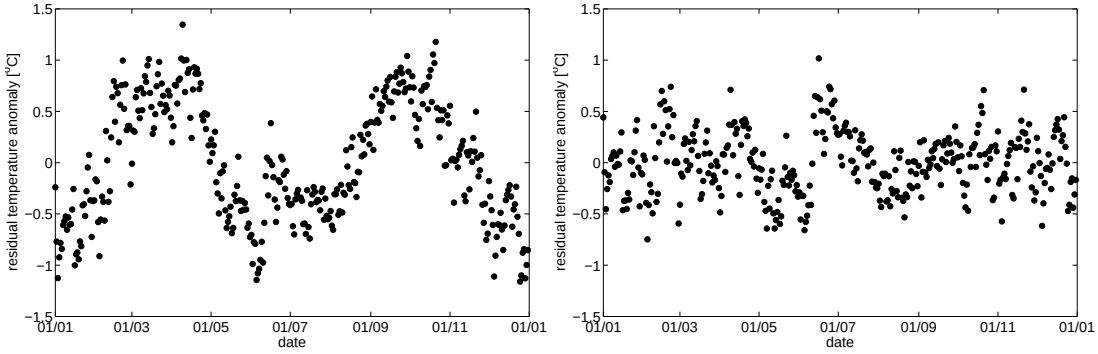


Figure A.9: Residuals after fitting the mean temperature for each calendar day with a smooth model containing only the yearly frequency (left panel) and with both the yearly and the twice-yearly frequency (right panel).

Fig. A.9 shows the residuals of a model fit with only the yearly frequency (left panel) and both the yearly and the twice-yearly frequency (right panel). If only the yearly frequency is chosen, the residuals still have a very obvious time dependence that vanishes when the twice-yearly frequency is also taken into account. This further confirms that at least two components are necessary.

If the climatological model represents the temperature time series well, then the anomalies averaged over larger samples should be very close to zero. To see whether this is the case and to illustrate whether any additional higher harmonics are needed, Fig. A.10 shows the mean anomalies for each calendar month for a climatology using up to four different frequencies. As can be seen, there is a significant temporal dependence when only using the yearly fundamental frequency, as already noted before. After adding the twice-yearly terms, there is a marked improvement but no further significant changes for additional higher harmonics.

The left panel of Fig. A.11 shows that while there is a significant difference in the curve shape between using only one frequency and adding the second, all further additions do not change the curve shape much. It also reveals that the additional frequency component takes care of the deviations from the perfect sinusoidal shape rather than representing true oscillations of double

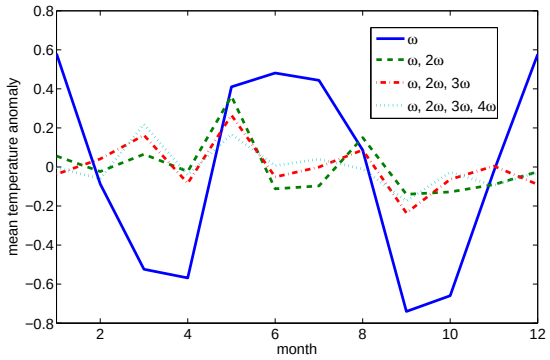


Figure A.10: Mean monthly anomalies after subtracting a climatology consisting of a sinusoidal model with a yearly fundamental frequency and up to three higher harmonics from the Potsdam morning temperature time series.

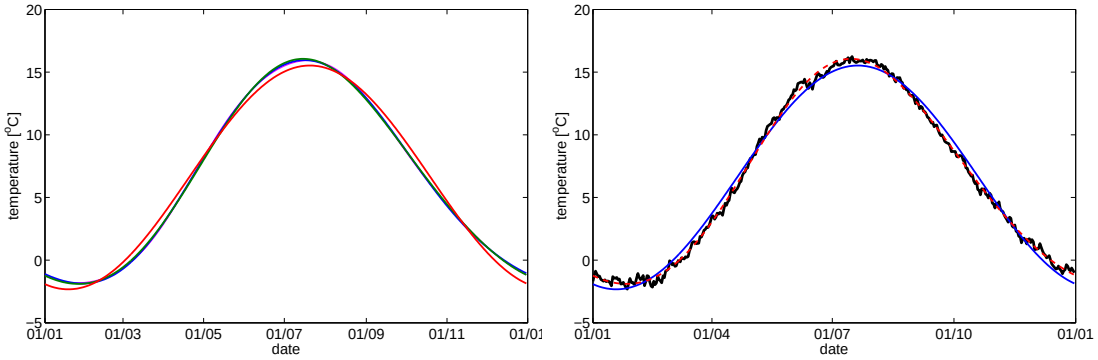


Figure A.11: Smooth climatological model as given by Eq.(A.2) with up to the first four frequencies (left panel). Mean temperatures and smooth climatological model for the first and the first two frequencies (right panel).

the fundamental frequency. The right panel illustrates that the addition of only the twice-yearly frequency does not only take care of the obvious remaining temporal dependence in the residuals but also fits the mean temperatures very well. It therefore seems that two frequencies in the climatological model is indeed the best choice.

Using this smooth model to derive the anomaly time series, we see in Fig. A.12 that the smoothing does not change the anomaly distribution for better or worse - there is no difference visible between Fig. A.12 and Fig. A.2: It still appears to be a mostly normal distribution with a much heavier negative tail. This is in fact characteristic of all such temperature anomaly distributions, and has already been noticed before[34].

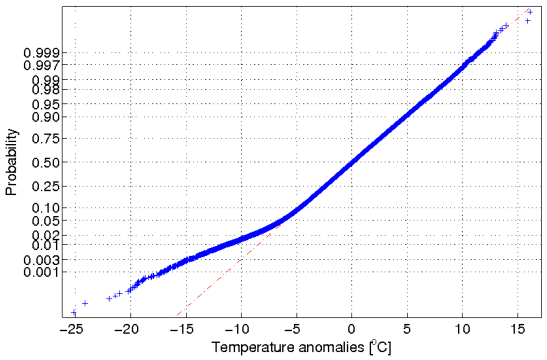


Figure A.12: Normal probability plot of the Potsdam morning temperature anomalies from the climatology computed using the smooth model with two frequencies for the long-term mean for each calendar day. The red dashed line represents the expected plot for a normal distribution.

A.4 Conclusion

Choosing the simplest possible definition of the climatology - namely the mean temperature for each calendar day - already offers a good model for the seasonal cycle even though the resulting anomaly distribution still has a negative tail that is too heavy compared to a normal distribution.

Two problems remain with this choice. The first one is the seasonality of the anomaly variance which might be remedied by normalising the anomalies by their standard deviation for each calendar day. However, while this removes the heavier tail of the distribution, it only shifts the seasonality from the variance to the skewness. We could therefore not treat the anomaly time series as truly stationary in either case. As the main concern is thus not removed through a rescaling and increased complexity of the climatological model also means increased risk of overfitting the time series, we will start our analysis using the simpler definition for the anomalies that leaves out the deseasoning.[4]

The second problem is the danger of overfitting the time series and removing some particulars of the recorded fluctuations along with the seasonal cycle. This can be avoided by smoothing the mean values with a sinusoidal model as given in Eq.(A.2) with $k = 2$, i.e. containing both the yearly and twice-yearly frequencies.

However, using this definition of the climatology one should keep in mind that there are stationarity issues in the anomaly variance and also a heavier negative tail in the probability distribution.

[4] As it is not clear how well the conclusions on the anomaly time series properties translate to first passage time problems on these anomalies, we will also use rescaled anomalies later (see Chapter 5).

B Appendix: Predictability of conditional first passage times - Figures

B.1 Introduction

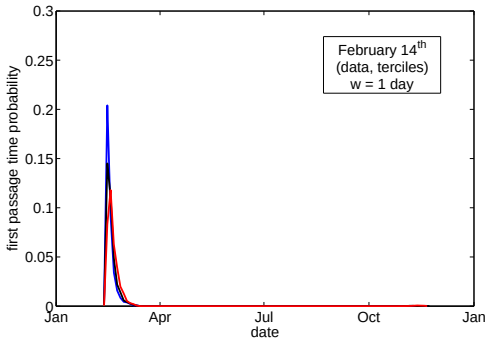
With the different time series both obtained directly from the measured data and through an autoregressive model process and predictability effects depending both on initial date and initial anomaly, there are many different conditional first passage time distributions to be studied in this thesis.

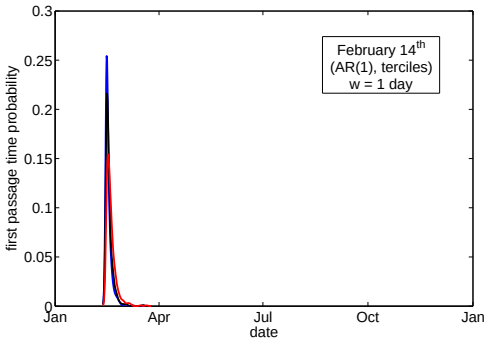
Taking into account that the predictability effects are studied not only in the full distributions but also in several different summary measures, this results in a multitude of figures of which only a few representative examples were shown within the main thesis. In order to provide the full backing for the claims made in the main part, we included those on which our analysis was based in the following appendix.

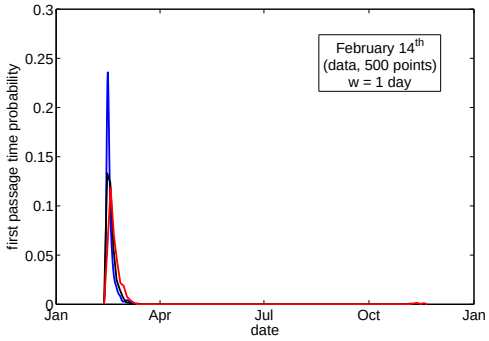
B.2 Full distribution estimates (original time series)

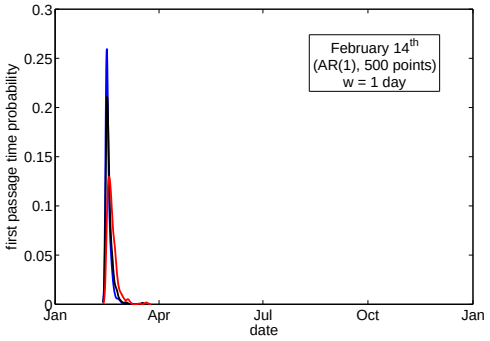
We will start with the original anomaly time series and the dependence of the resulting full first passage time distributions to frost on the initial anomalies for different initial dates as analysed in Sec. 4.2.1.

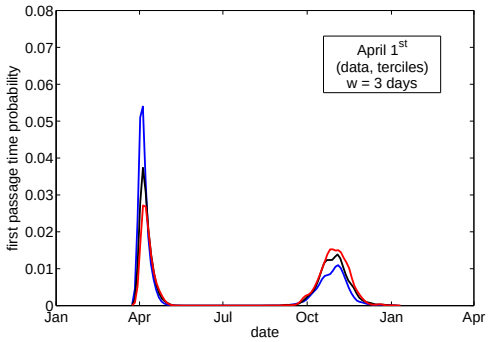
This section therefore shows the kernel density estimates of the conditional first passage time distributions for initial anomalies recorded within a 3 month window around the stated initial date. The distributions depicted in the top row for each initial date were conditioned on initial anomalies lying within a chosen initial anomaly tercile, with colder than average anomalies depicted in blue, average in black and warmer than average anomalies in red. The figures in the bottom row for each initial date show the conditioning on anomaly bands of the warmest and coldest 500 data points as well as the 500 closest to the distribution median. For each estimate, normal density kernels were used, with the width w as stated in the legend of each figure.

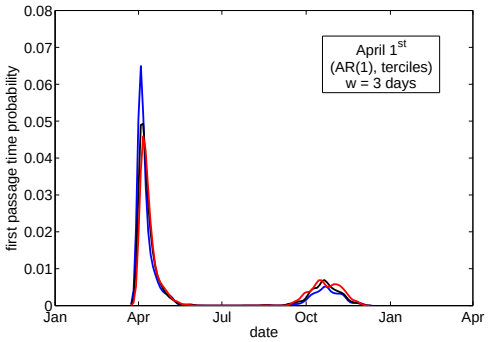

February 14th
(data, terciles)
w = 1 day
first passage time probability
0.3
0.25
0.2
0.15
0.1
0.05
0
Jan
Apr
Jul
Oct
Jan
date

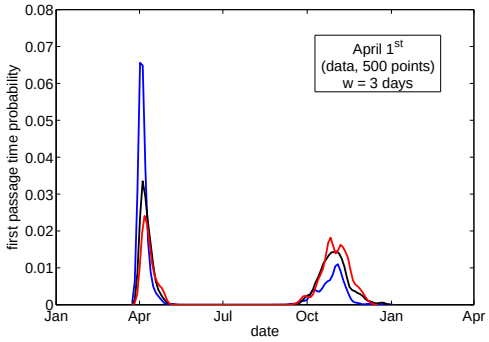

February 14th
(AR(1), terciles)
w = 1 day
first passage time probability
0.3
0.25
0.2
0.15
0.1
0.05
0
Jan
Apr
Jul
Oct
Jan
date

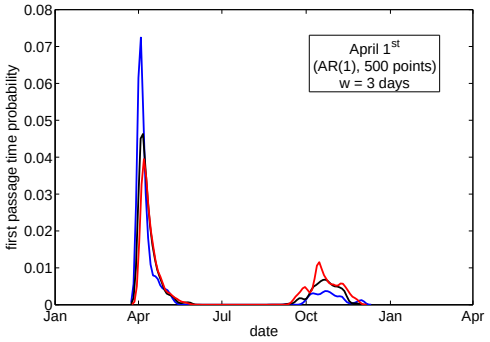

February 14th
(data, 500 points)
w = 1 day
first passage time probability
0.3
0.25
0.2
0.15
0.1
0.05
0
Jan
Apr
Jul
Oct
Jan
date

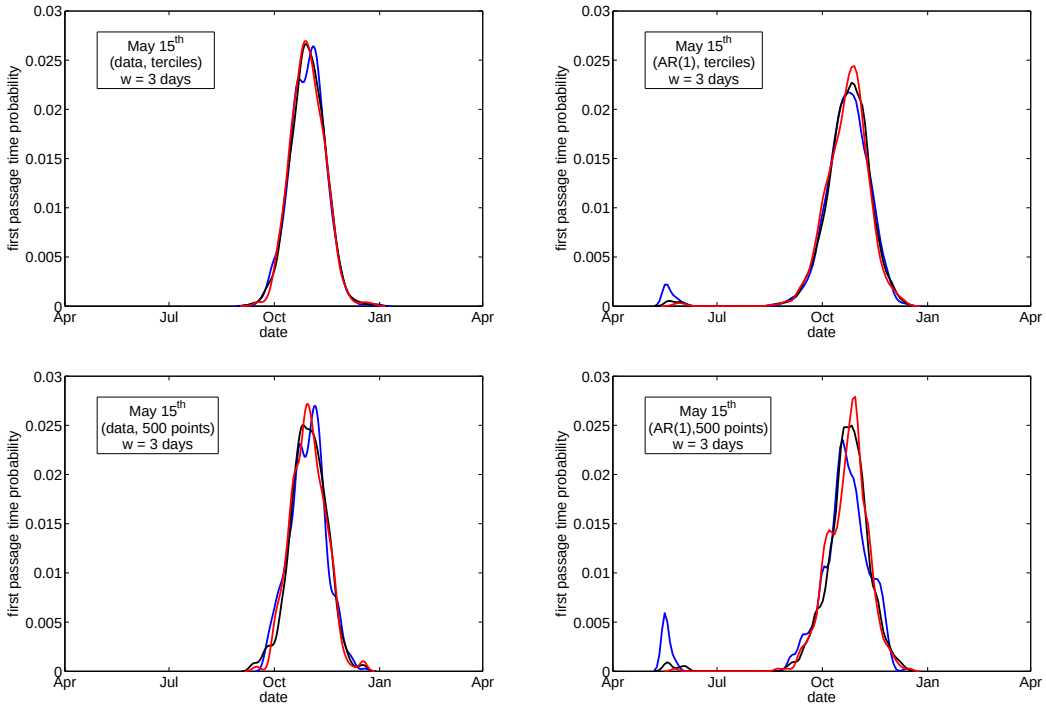

February 14th
(AR(1), 500 points)
w = 1 day
first passage time probability
0.3
0.25
0.2
0.15
0.1
0.05
0
Jan
Apr
Jul
Oct
Jan
date

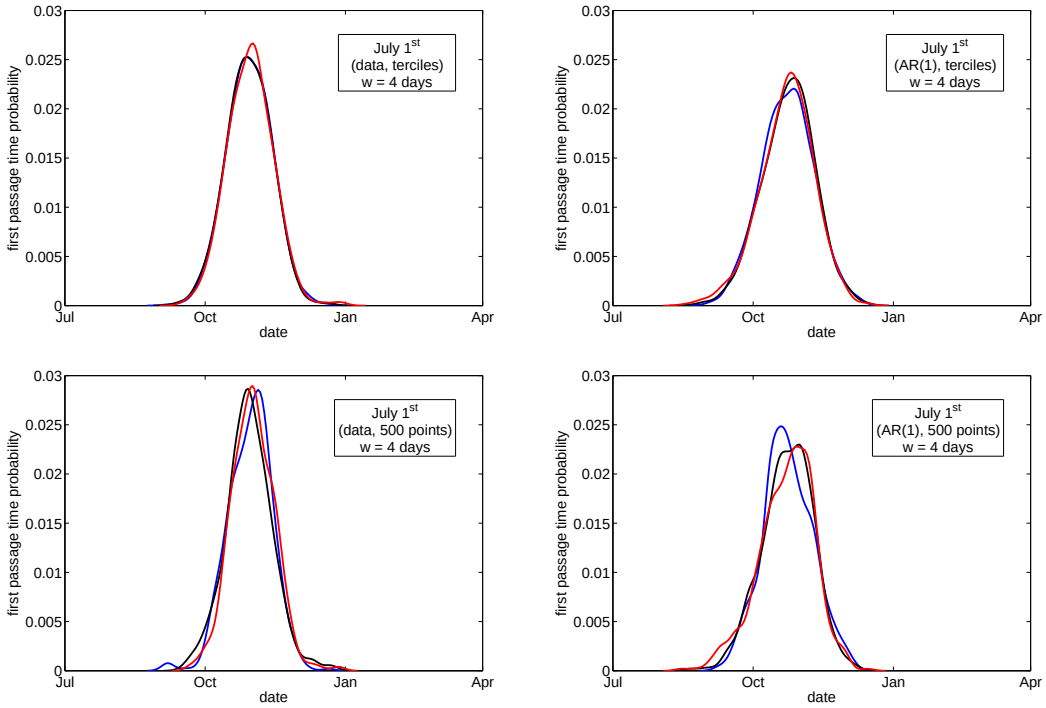

April 1st
(data, terciles)
w = 3 days
first passage time probability
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Jan
Apr
Jul
Oct
Jan
Apr
date

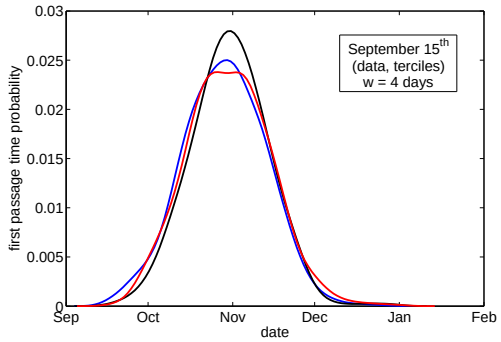

April 1st
(AR(1), terciles)
w = 3 days
first passage time probability
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Jan
Apr
Jul
Oct
Jan
Apr
date

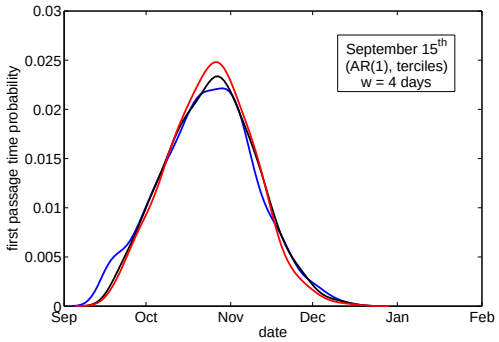

April 1st
(data, 500 points)
w = 3 days
first passage time probability
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Jan
Apr
Jul
Oct
Jan
Apr
date

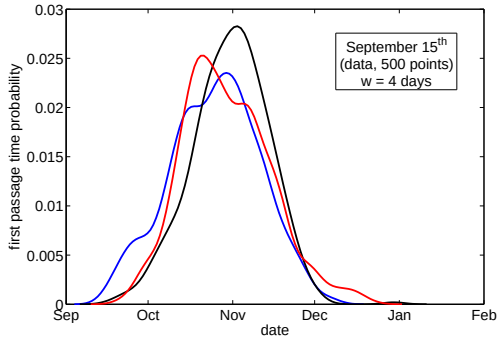

April 1st
(AR(1), 500 points)
w = 3 days
first passage time probability
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Jan
Apr
Jul
Oct
Jan
Apr
date

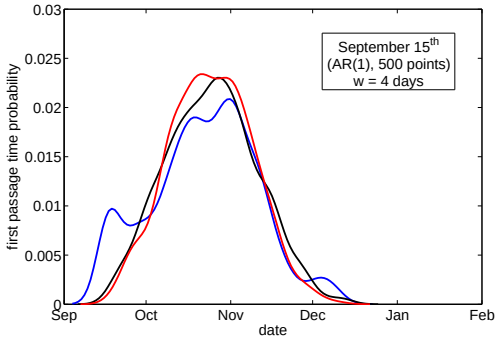

May 15th
(data, terciles)
w = 3 days
May 15th
(AR(1), terciles)
w = 3 days
May 15th
(data, 500 points)
w = 3 days
May 15th
(AR(1),500 points)
w = 3 days
first passage time probability
0.03
0.025
0.02
0.015
0.01
0.005
0
Apr
Jul
Oct
Jan
Apr
date

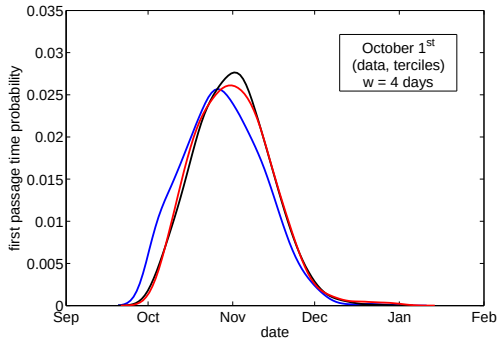

July 1st
(data, terciles)
w = 4 days
July 1st
(AR(1), terciles)
w = 4 days
July 1st
(data, 500 points)
w = 4 days
July 1st
(AR(1), 500 points)
w = 4 days
first passage time probability
0.03
0.025
0.02
0.015
0.01
0.005
0
Jul
Oct
Jan
Apr
date

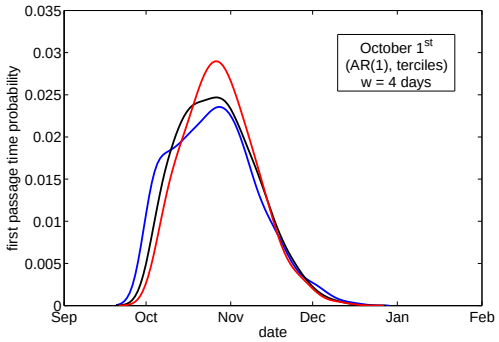

September 15th
(data, terciles)
w = 4 days
first passage time probability
0.03
0.025
0.02
0.015
0.01
0.005
0
Sep
Oct
Nov
Dec
Jan
Feb
date

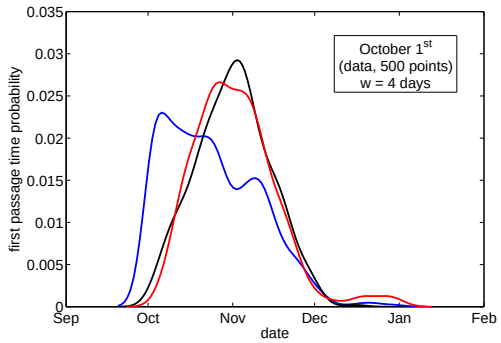

September 15th
(AR(1), terciles)
w = 4 days
first passage time probability
0.03
0.025
0.02
0.015
0.01
0.005
0
Sep
Oct
Nov
Dec
Jan
Feb
date

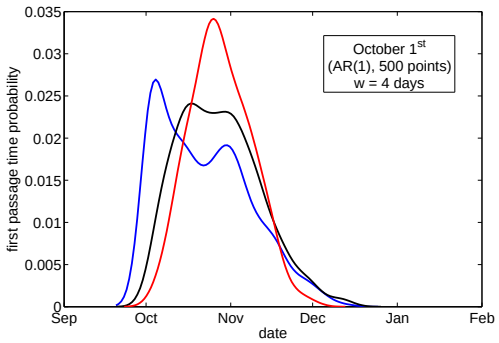

September 15th
(data, 500 points)
w = 4 days
first passage time probability
0.03
0.025
0.02
0.015
0.01
0.005
0
Sep
Oct
Nov
Dec
Jan
Feb
date

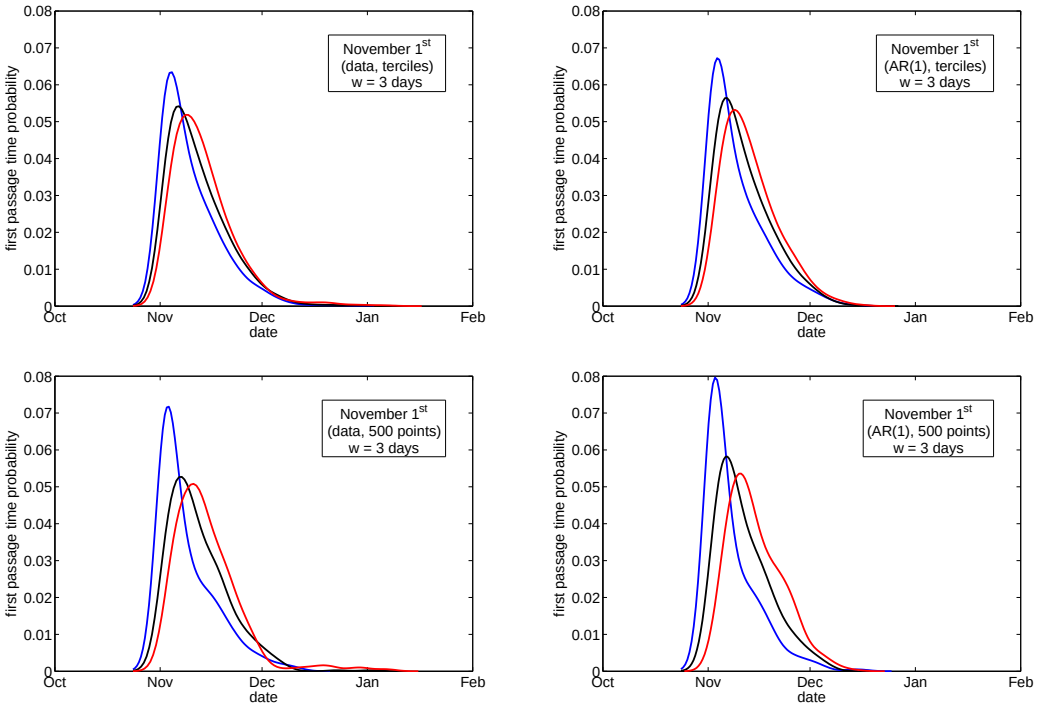

September 15th
(AR(1), 500 points)
w = 4 days
first passage time probability
0.03
0.025
0.02
0.015
0.01
0.005
0
Sep
Oct
Nov
Dec
Jan
Feb
date

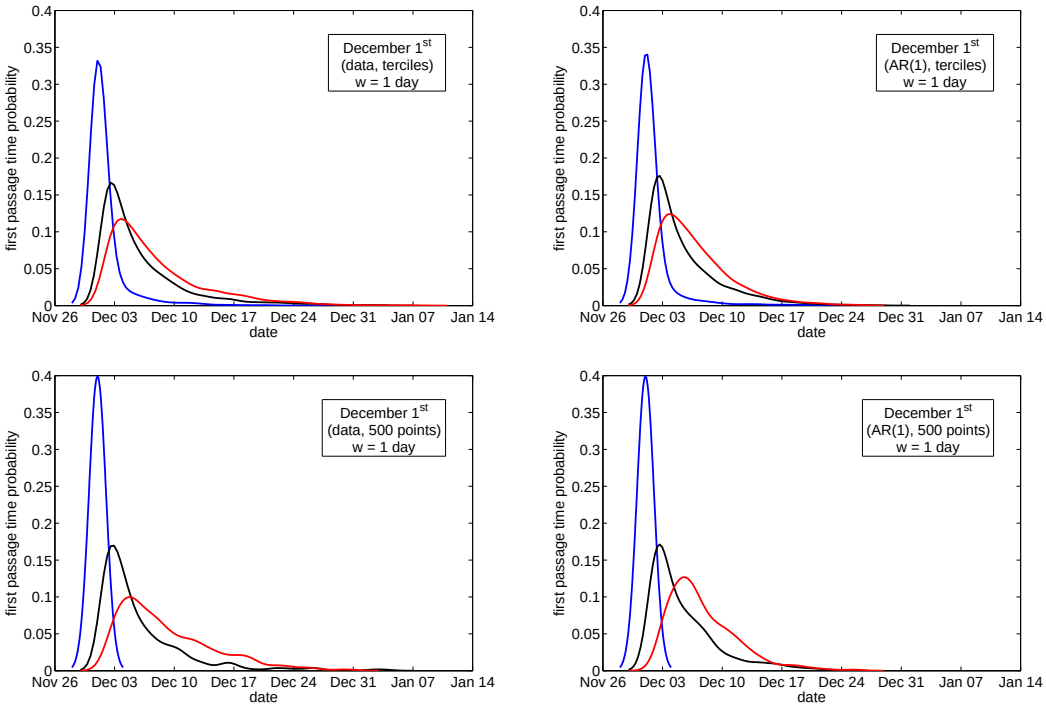

October 1st
(data, terciles)
w = 4 days
first passage time probability
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
Sep
Oct
Nov
Dec
Jan
Feb
date


October 1st
(AR(1), terciles)
w = 4 days
first passage time probability
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
Sep
Oct
Nov
Dec
Jan
Feb
date


October 1st
(data, 500 points)
w = 4 days
first passage time probability
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
Sep
Oct
Nov
Dec
Jan
Feb
date


October 1st
(AR(1), 500 points)
w = 4 days
first passage time probability
0.035
0.03
0.025
0.02
0.015
0.01
0.005
0
Sep
Oct
Nov
Dec
Jan
Feb
date


November 1st
(data, terciles)
w = 3 days
November 1st
(AR(1), terciles)
w = 3 days
November 1st
(data, 500 points)
w = 3 days
November 1st
(AR(1), 500 points)
w = 3 days
first passage time probability
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
Oct
Nov
Dec
Jan
Feb
date


December 1st
(data, terciles)
w = 1 day
December 1st
(AR(1), terciles)
w = 1 day
December 1st
(data, 500 points)
w = 1 day
December 1st
(AR(1), 500 points)
w = 1 day
first passage time probability
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Nov 26
Dec 03
Dec 10
Dec 17
Dec 24
Dec 31
Jan 07
Jan 14
date

B.3 Influence of the initial anomaly decile on summary measures (original time series)

Moving on from the full first passage time distributions to frost, we now look at distinct summary measures and their dependence on initial anomaly deciles for several different initial dates as indicated by the figure captions. These were analysed in Sec. 4.2.2.

Depending on whether the distribution is bimodal or not, we used the mean or median as a measure of the location or - in the bimodal case - of the relative weight of the two peaks, and the standard deviation (std) or interquartile range (IQR) as a measure of the distribution spread. In the following, the location is measured by the average number of days remaining until the next occasion of frost happens (mean) or the number of days until the next occasion of frost has happened in 50% of all cases. The spread is always measured in the number of days.

The third summary measure was the date on which the probability of first frost occurring had reached its maximum, or - in the case of the bimodal distributions - the locations of the two peaks. Note that, since this was obtained using a kernel density estimate of the underlying probability distribution, the support is not discrete any longer in this case.

The panels on the left side originate in $P_{\mathrm{data}}(t_{\mathrm{frost}}|t_0, \Delta T^{(10)})$, while the panels on the right side were obtained from $P_{\mathrm{AR(1)}}(t_{\mathrm{frost}}|t_0, \Delta T^{(10)})$.

In order to estimate the influence of statistical fluctuations due to the finite length of the time series, we also plotted the mean of 1000 bootstrap samples of the time series as a continuous line, as well as error bars representing the width of two standard deviations of the bootstrap calculation (see Sec. 2.2.6 for details).

Figure B.1: Initial date: February 14$^{\text{th}}$

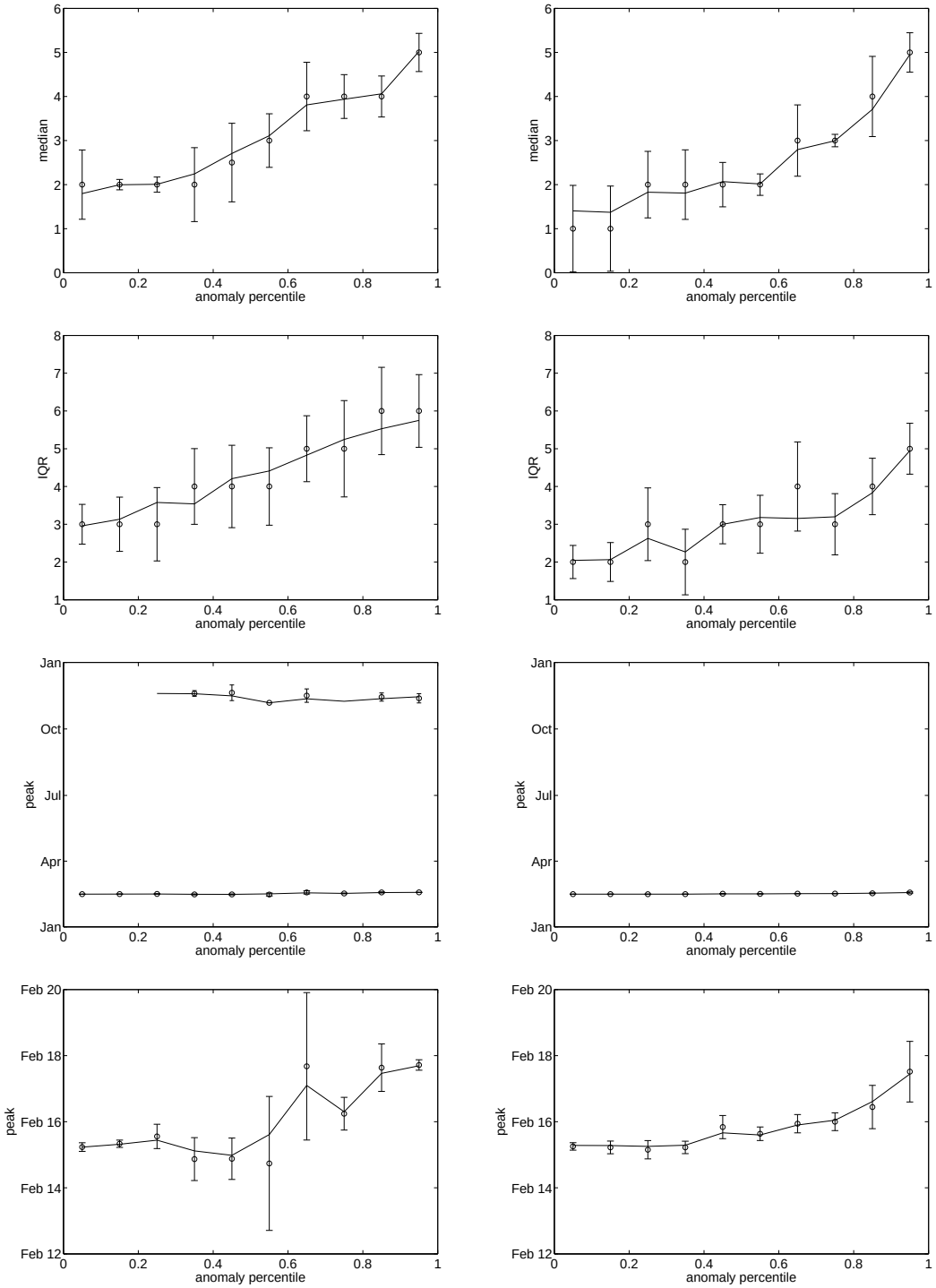


Figure B.2: Initial date: April 1st

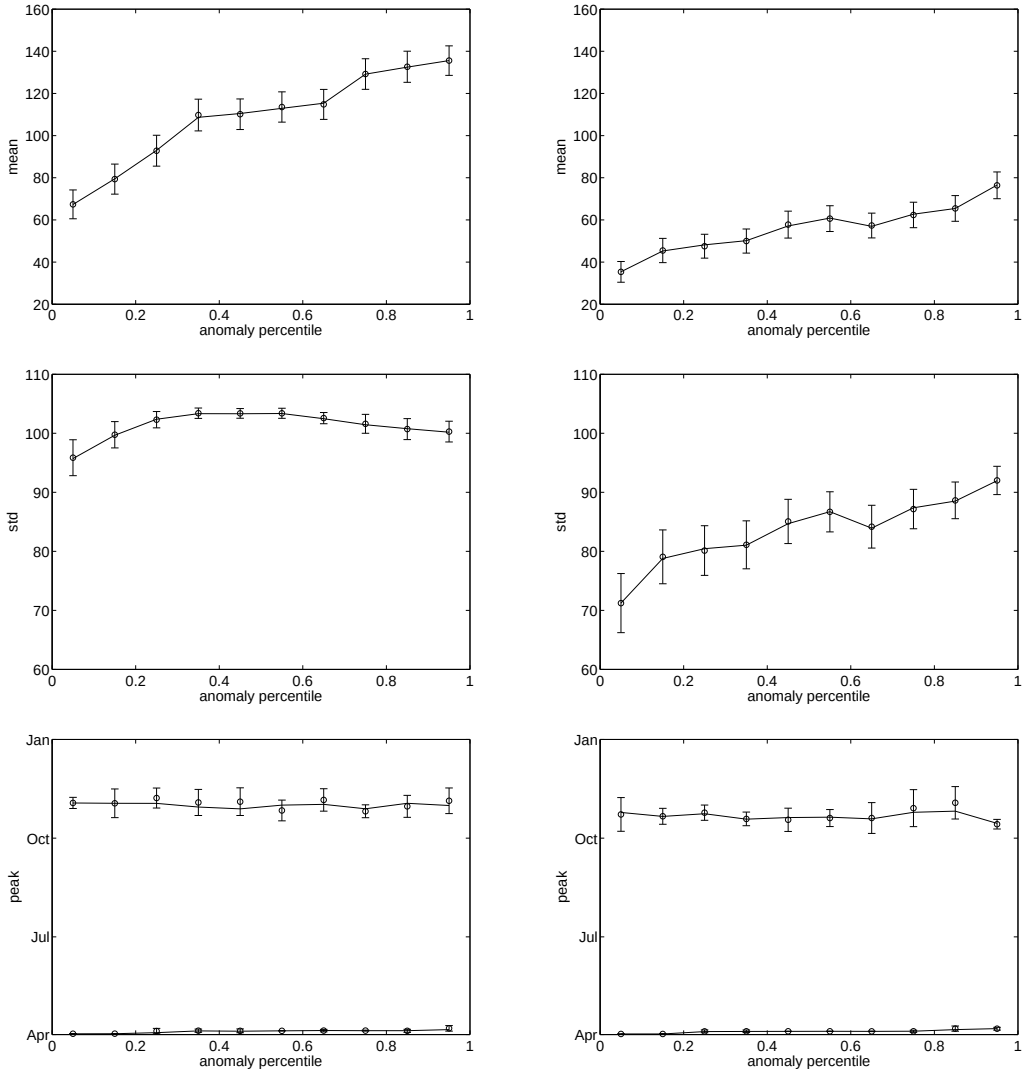


Figure B.3: Initial date: May 15^{th}

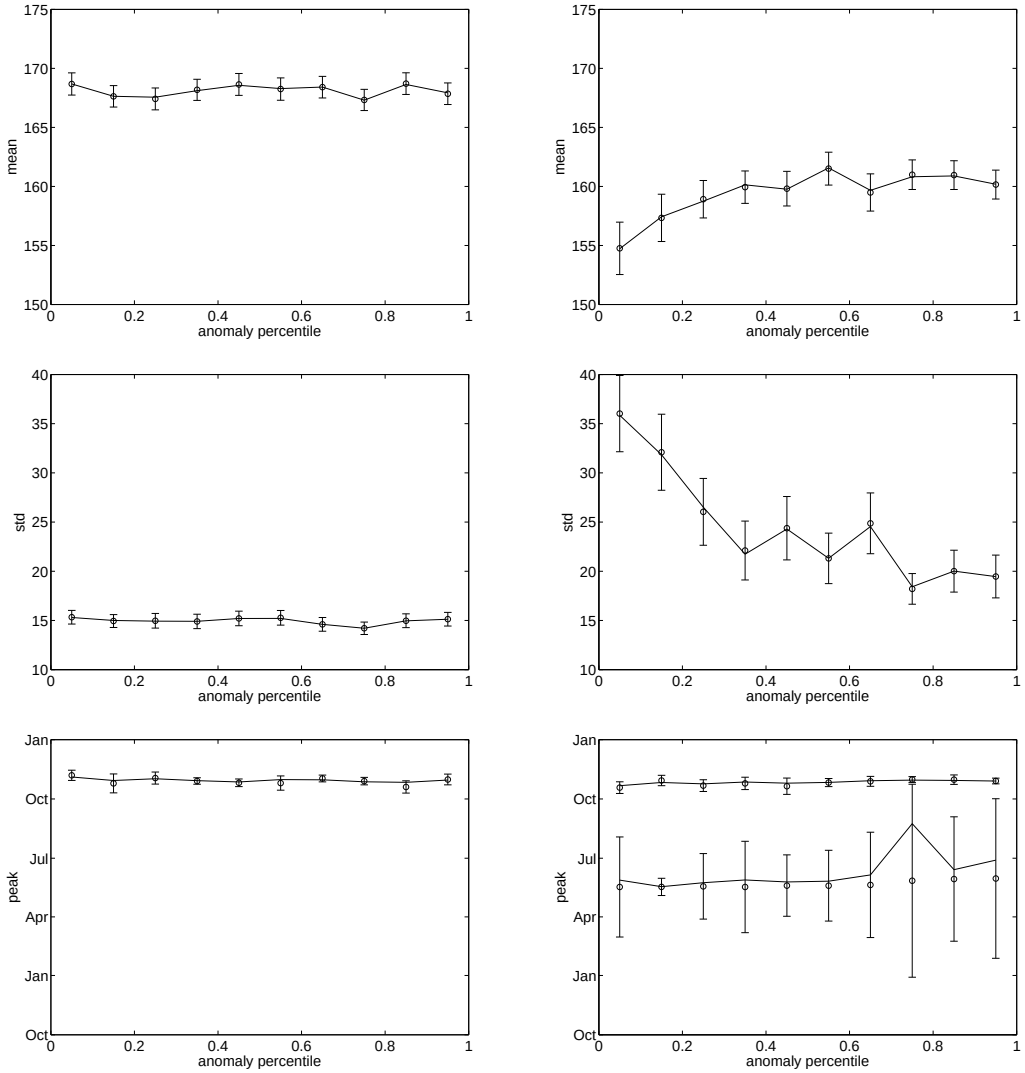


Figure B.4: Initial date: July 1st

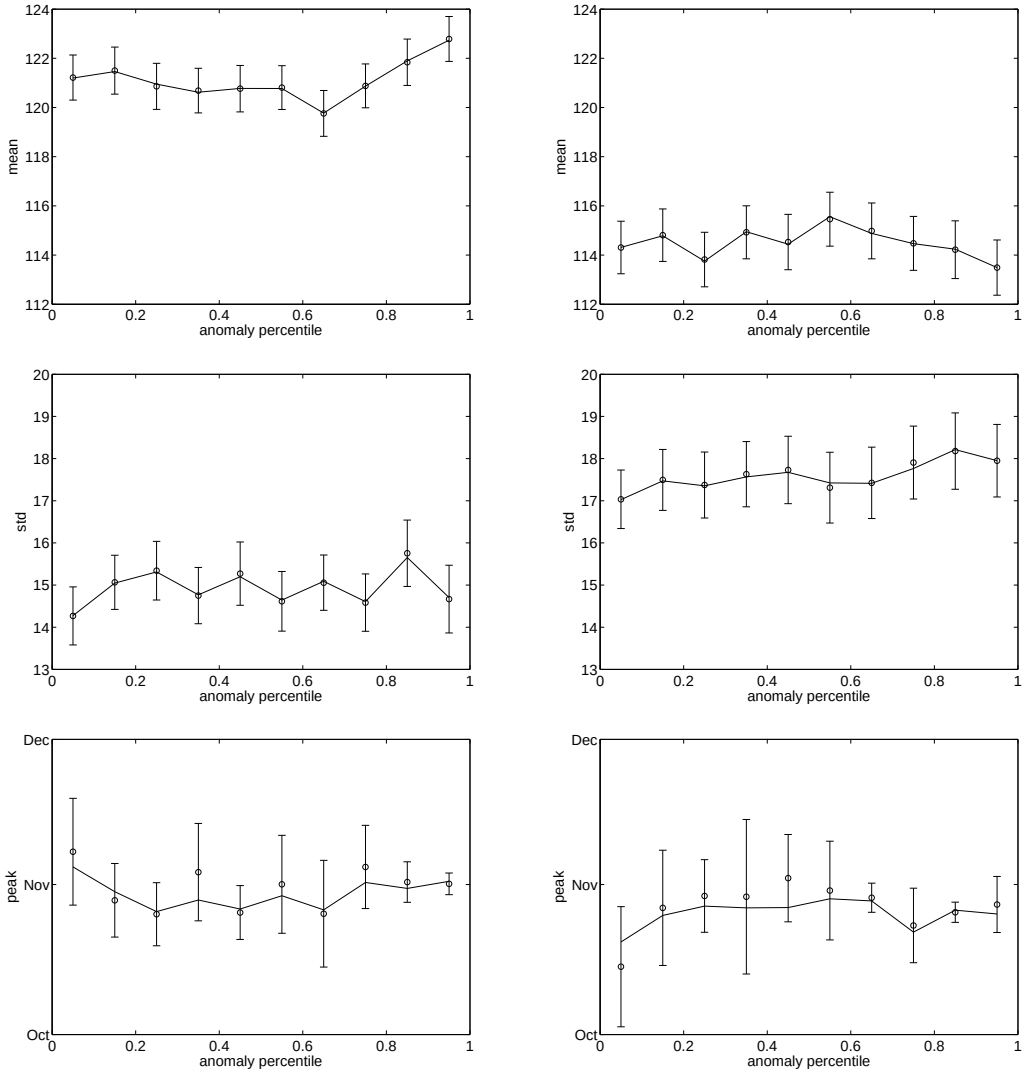


Figure B.5: Initial date: September 15$^{\text{th}}$

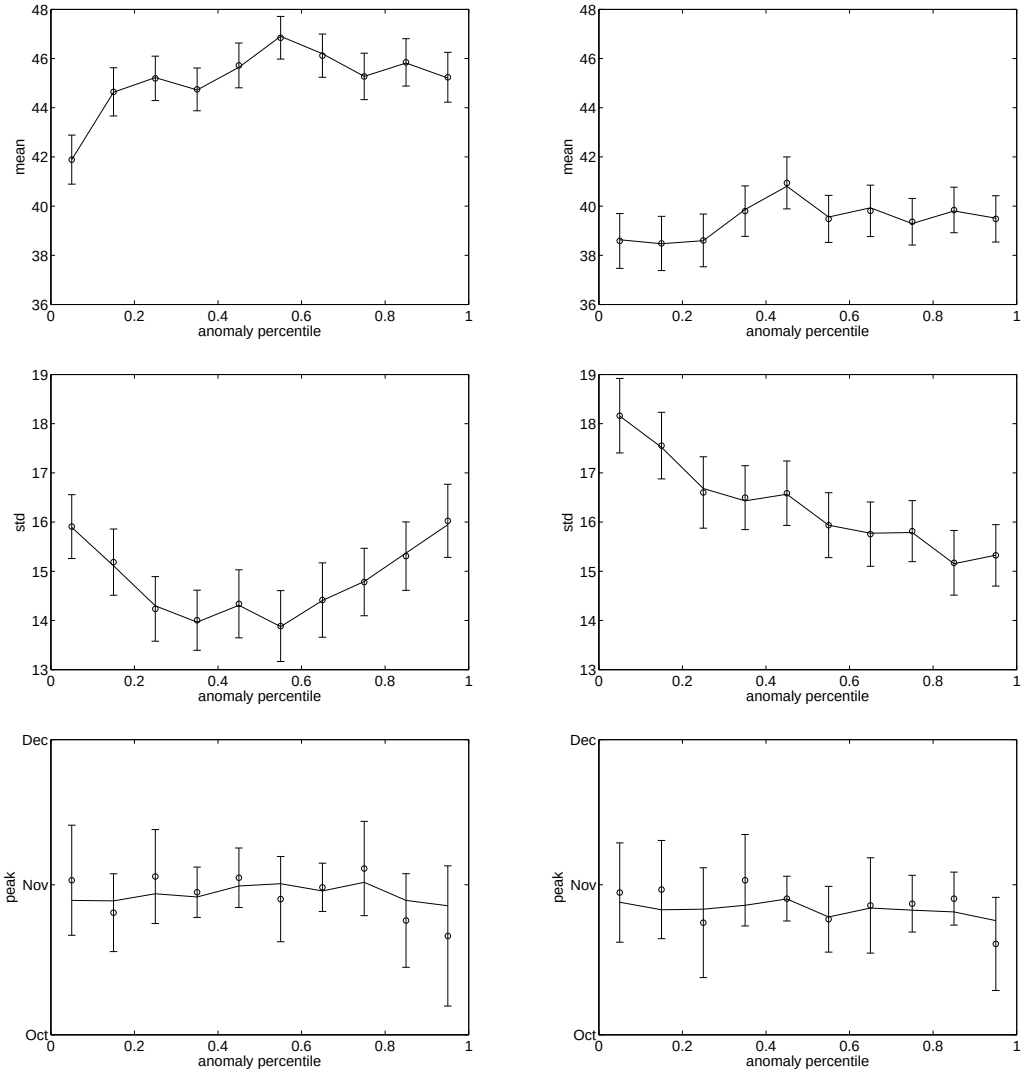


Figure B.6: Initial date: October 1st

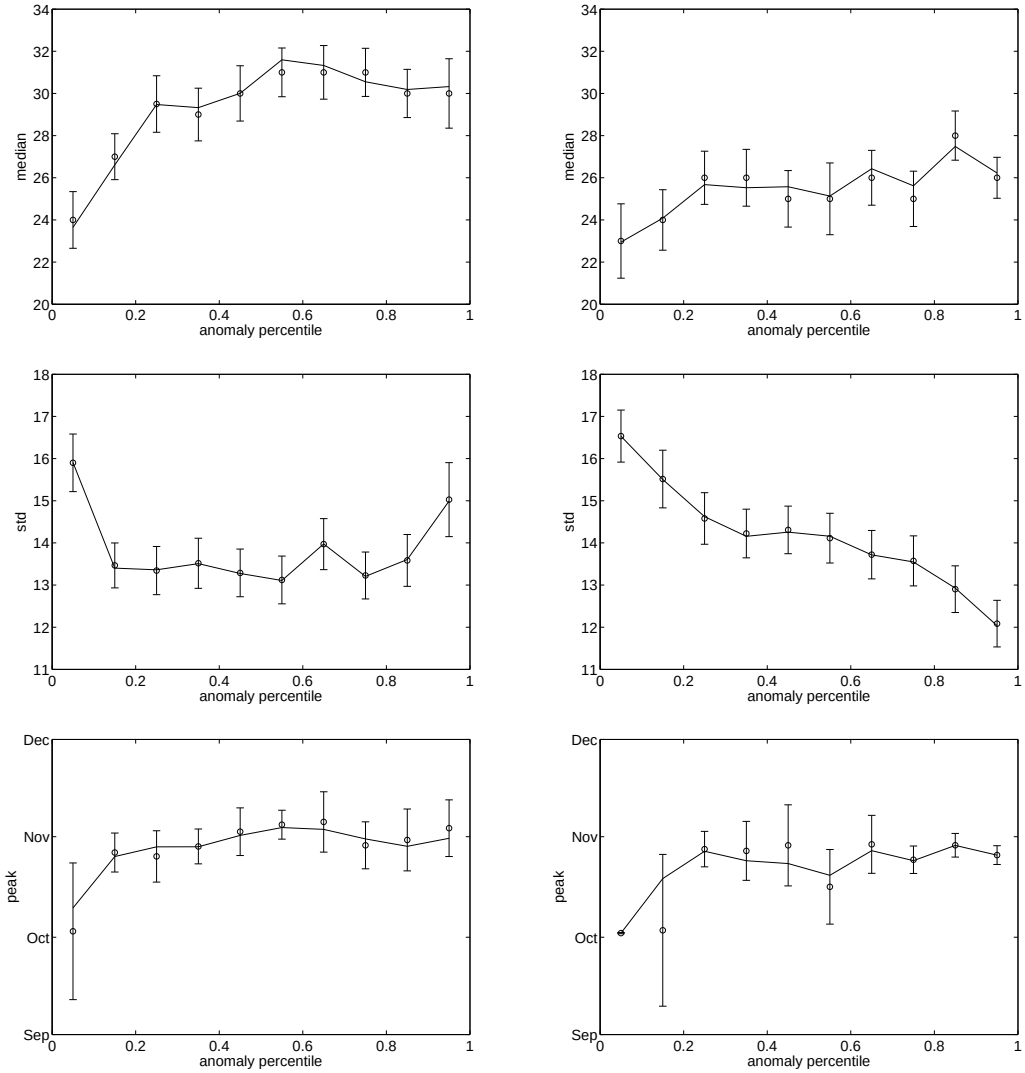


Figure B.7: Initial date: November 1$^{\text{st}}$

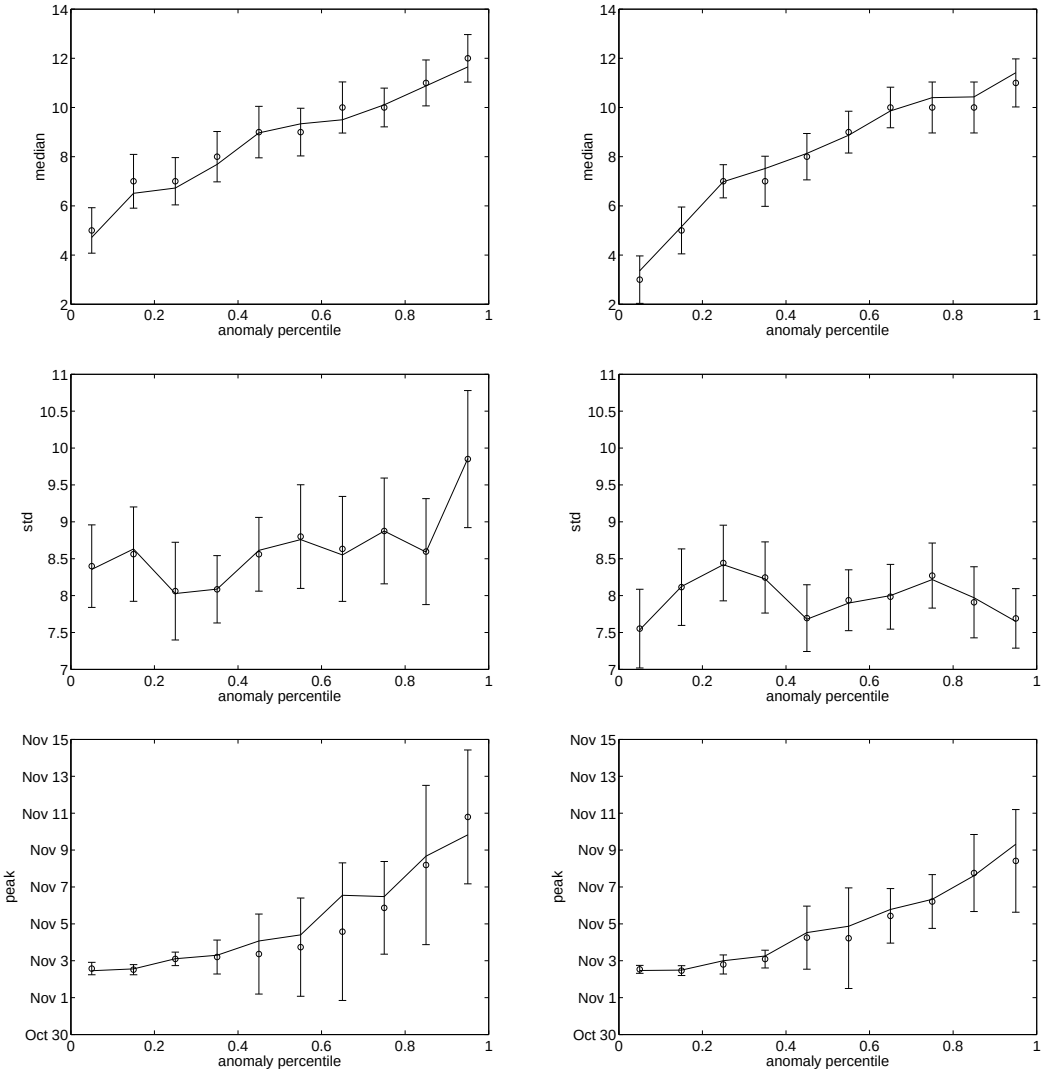
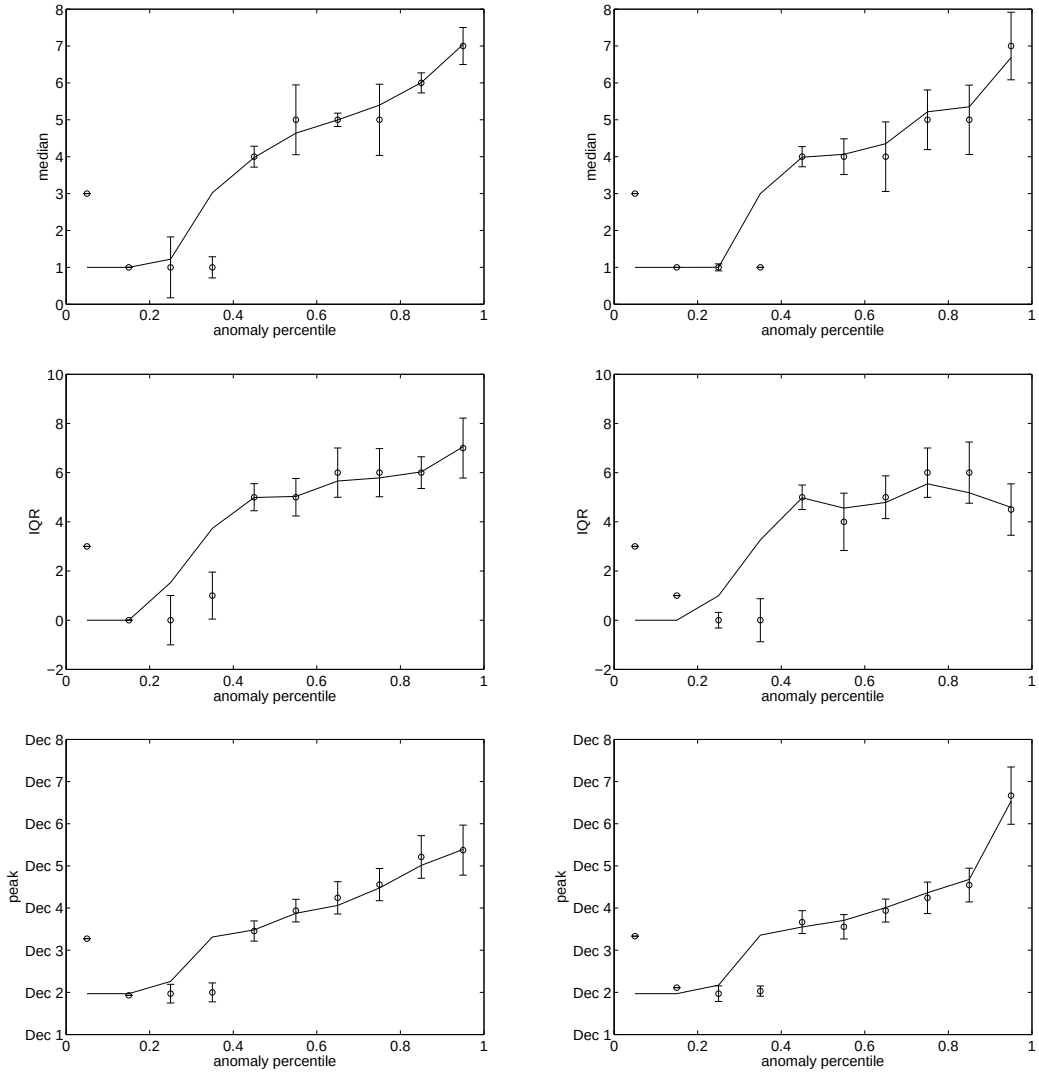


Figure B.8: Initial date: December 1st



B.4 Influence of the initial anomaly decile for more than 8 days' lead time (original time series)

In order to see how many predictability effects remain beyond a lead time of 8 days for which current operational weather forecasts are still useful[137], we show the same summary measures of the first passage time probability distributions as in Sec. B.3, excluding all first passage times of less than 8 days.

As a comparison, we included the corresponding results obtained with the AR(1) process that only has a decorrelation time of around 5 days in the right panels.

Since for the previously considered initial dates of May 15$^{\text{th}}$ and July 1$^{\text{st}}$ all recorded first passage times (at least from the temperature measurements) are larger than 8 days, they were not plotted again here.

The analysis of these figures was also done in Sec. 4.2.2.

Figure B.9: Initial date: February 14$^{\text{th}}$

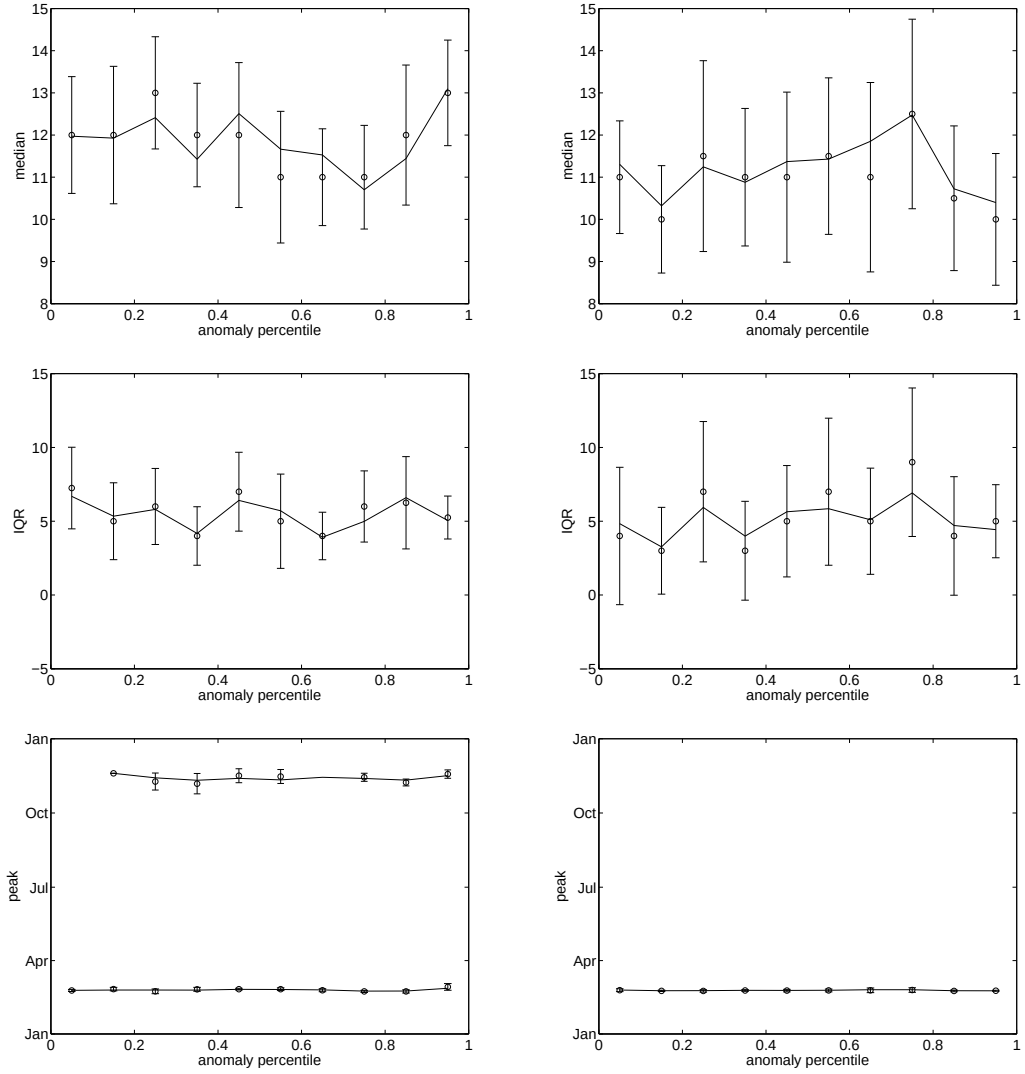


Figure B.10: Initial date: April 1st

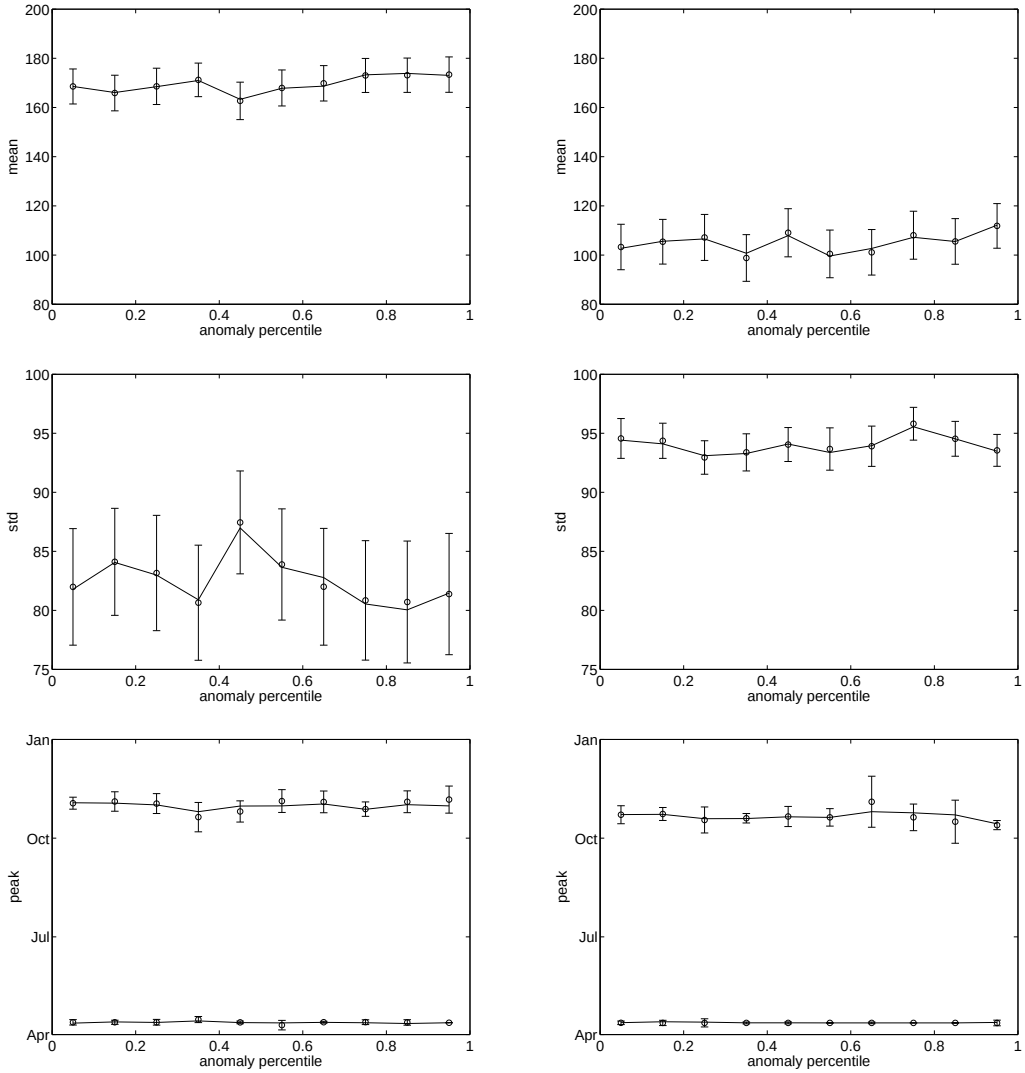


Figure B.11: Initial date: September 15^{th}

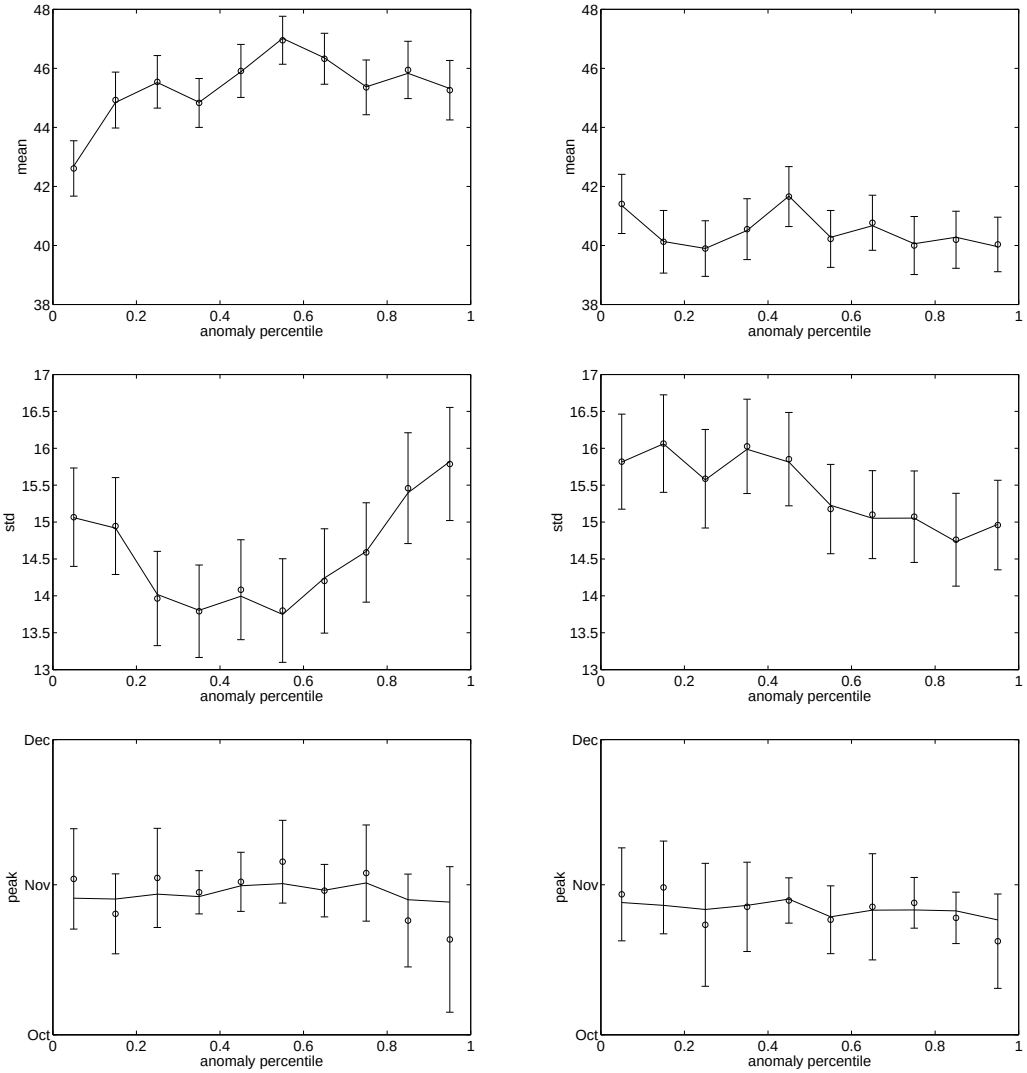


Figure B.12: Initial date: October 1st

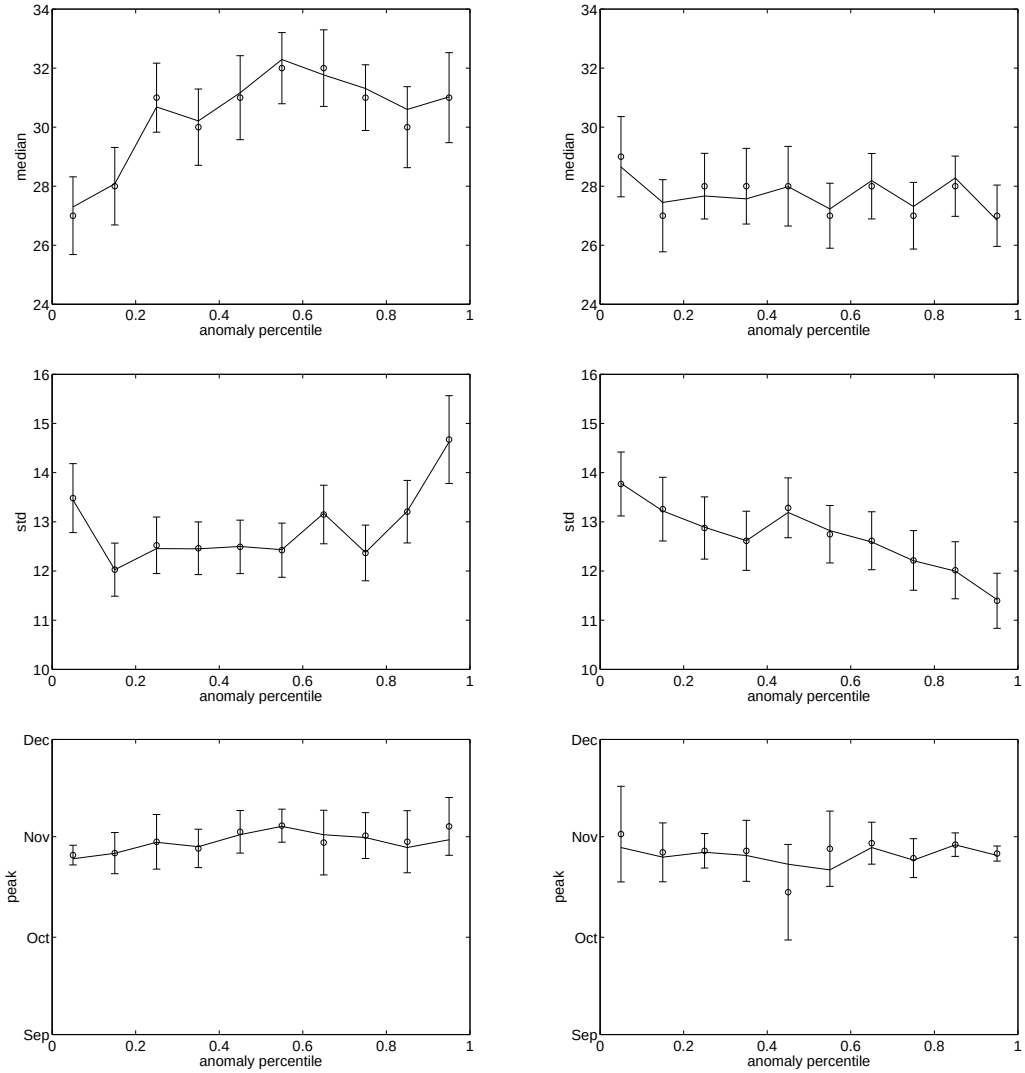


Figure B.13: Initial date: November 1st

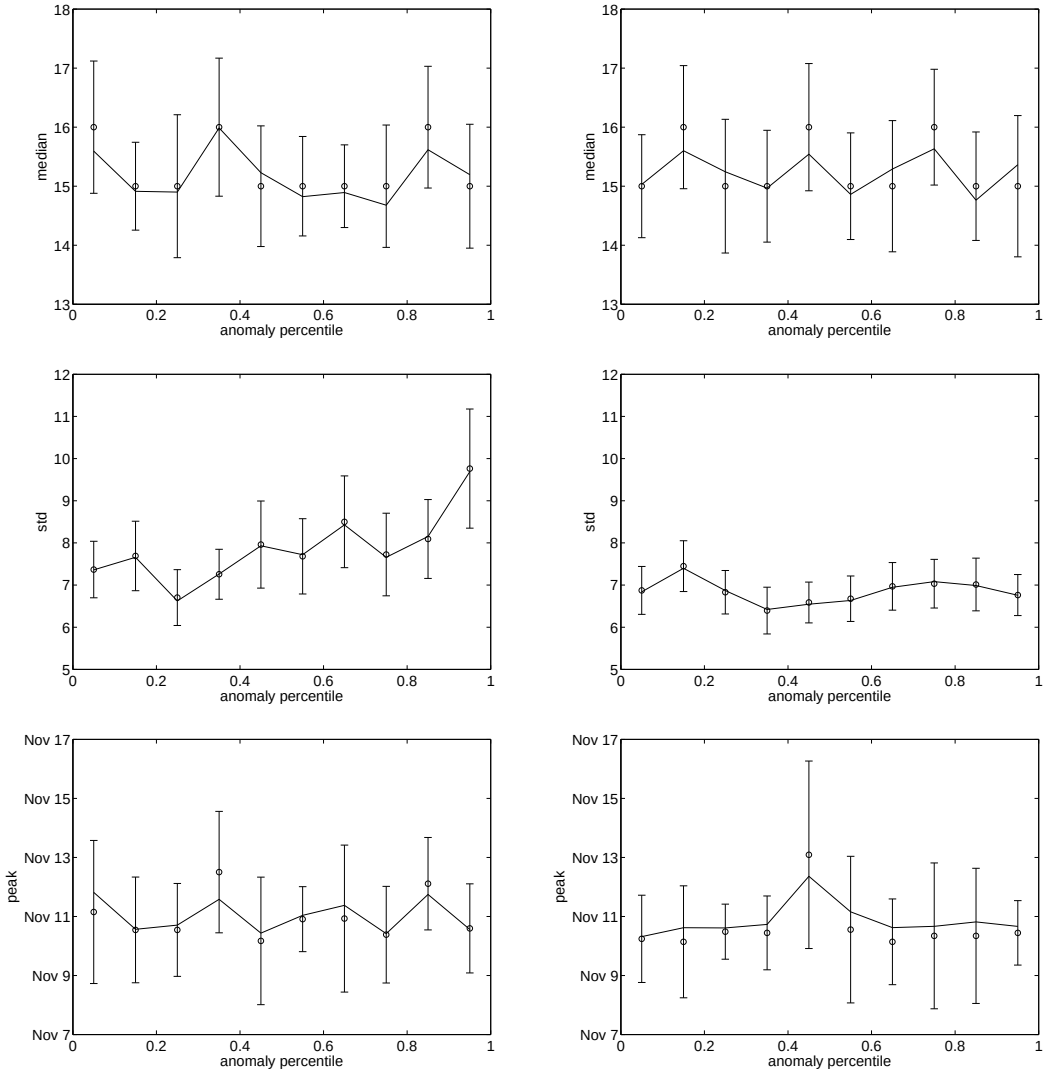
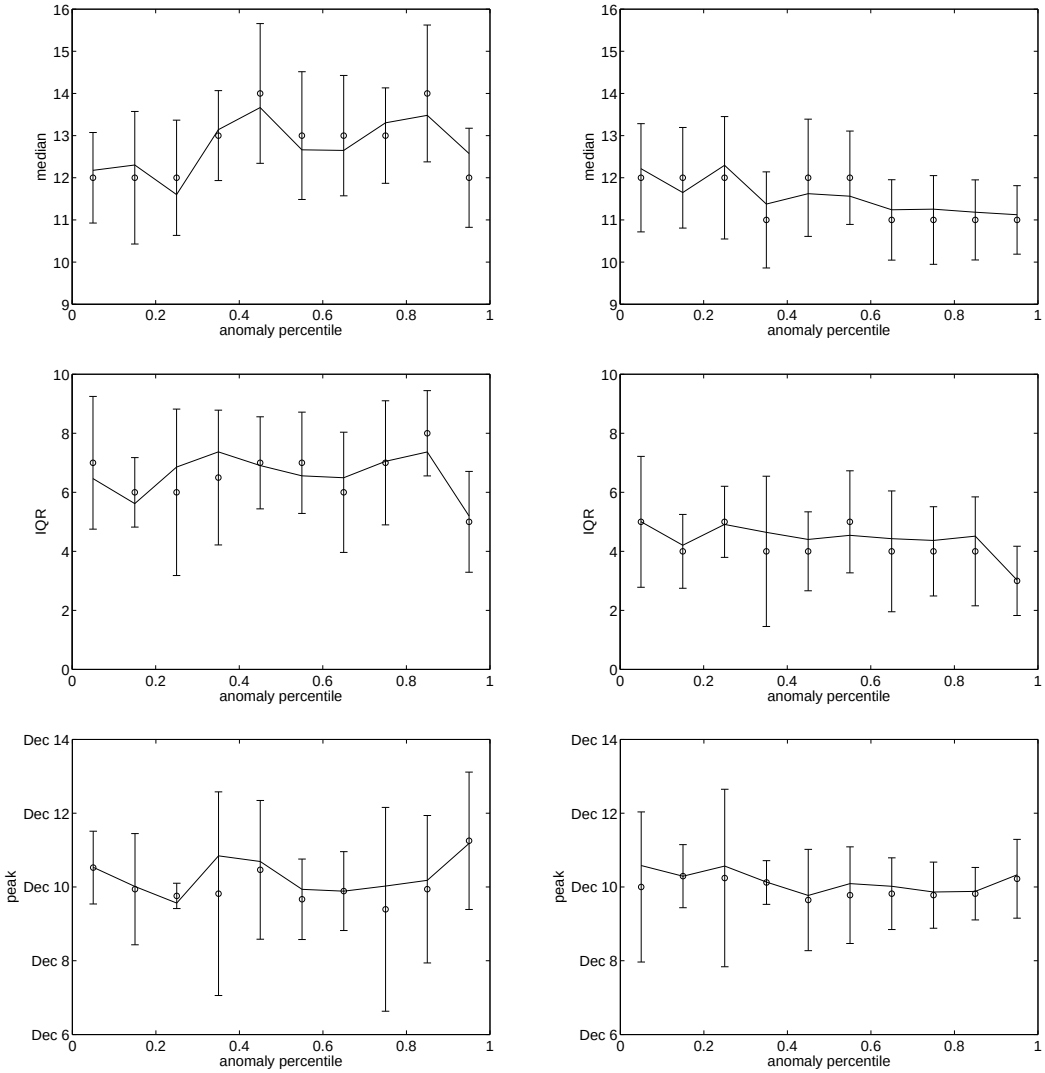


Figure B.14: Initial date: December 1st



B.5 Full distribution estimates (improved time series)

This section shows the kernel density estimates of $P_{\text{data}}(t_{\text{frost}}|t_0, y, \Delta T)$ in the left panels and from $P_{\text{AR(8)}}(t_{\text{frost}}|t_0, y, \Delta T)$ in the right panels. The model was found to be in better agreement with the measured anomalies in this case (see Sec. 5.2).

The distributions all use $\Delta T^{(3)}$, with colder than average anomalies depicted in blue, average in black and warmer than average anomalies in red. The top row displays $y = 1999$, i.e. maximum trend, the middle row $y = 1965$, i.e. an average trend, and the bottom row $y = 1941$, i.e. minimal trend.

For each estimate, normal density kernels were used, with the width $w = 1$ day for the initial dates February 14$^{\text{th}}$ and December 1$^{\text{st}}$, $w = 2$ days for the initial date of November 1$^{\text{st}}$ and $w = 3$ days for all other initial dates.

Figure B.15: Initial date: February 14$^{\text{th}}$

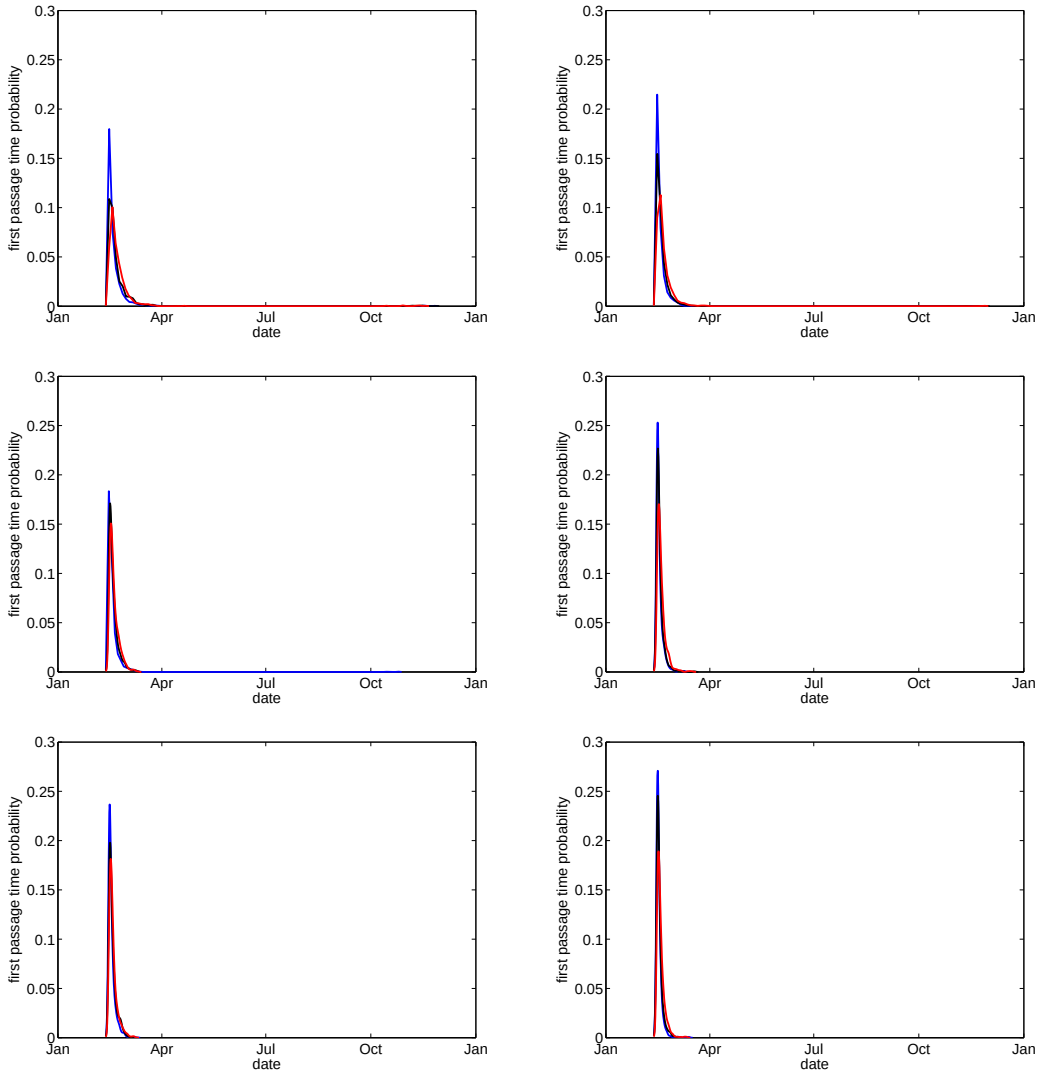


Figure B.16: Initial date: April 1st

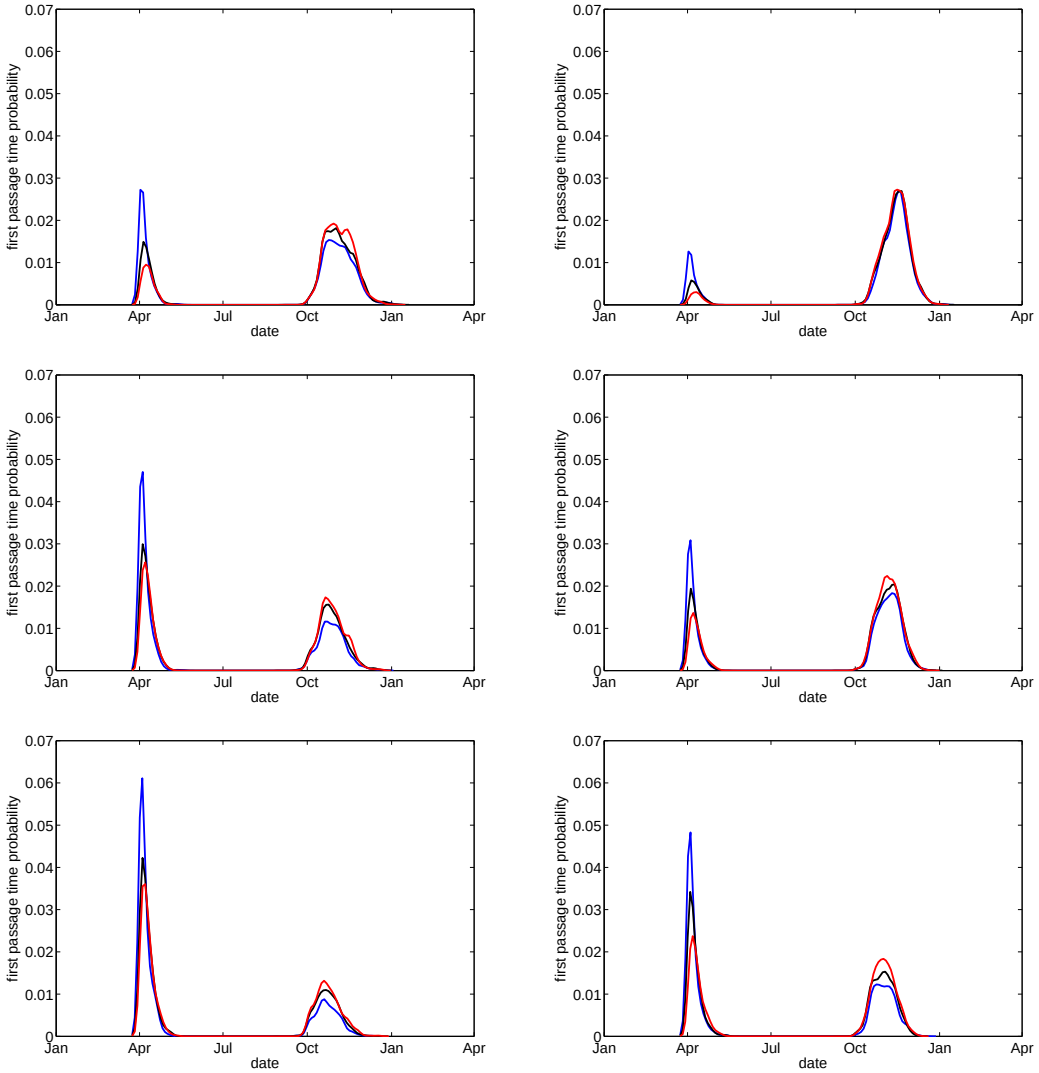


Figure B.17: Initial date: May 15^{th}

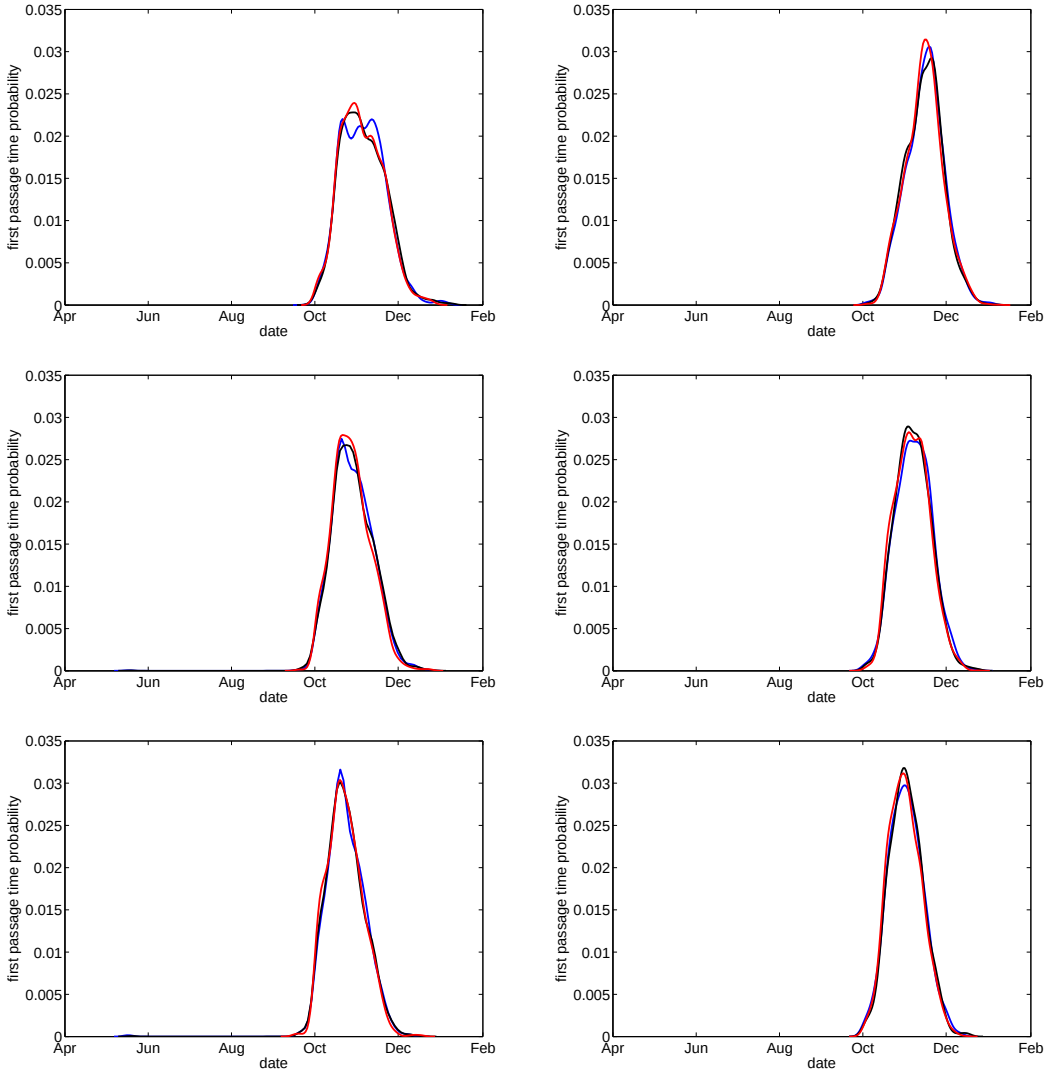


Figure B.18: Initial date: July 1st

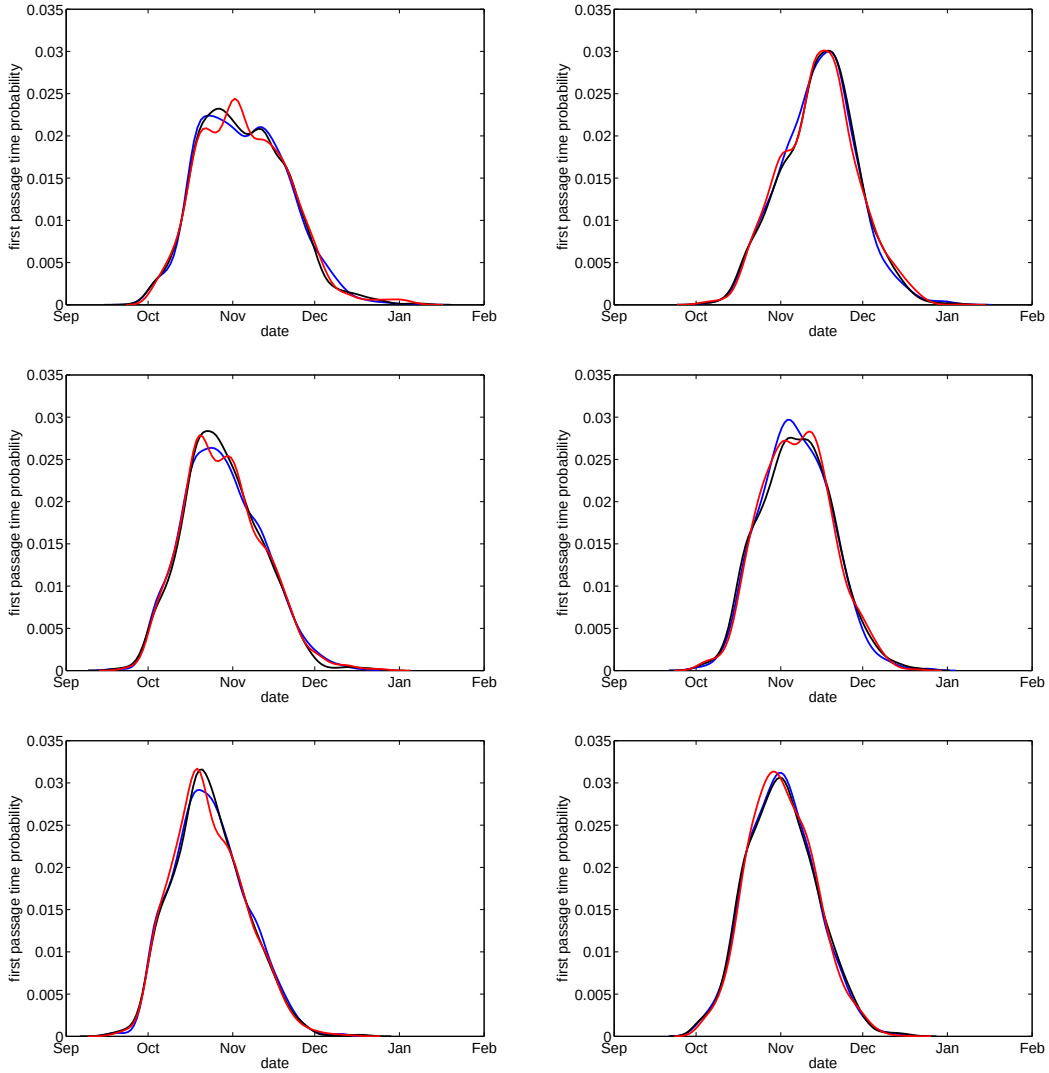


Figure B.19: Initial date: September 15$^{\text{th}}$

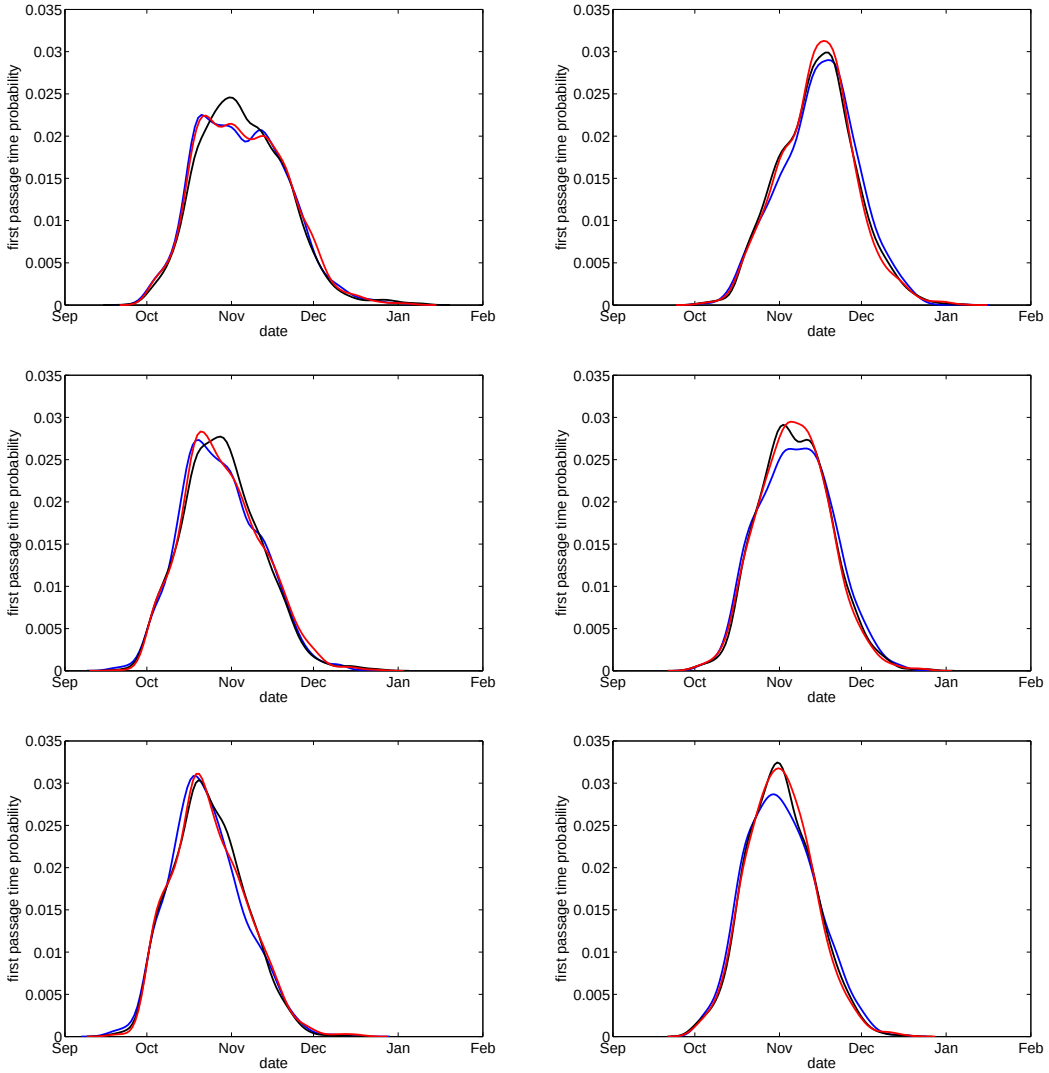


Figure B.20: Initial date: October 1st

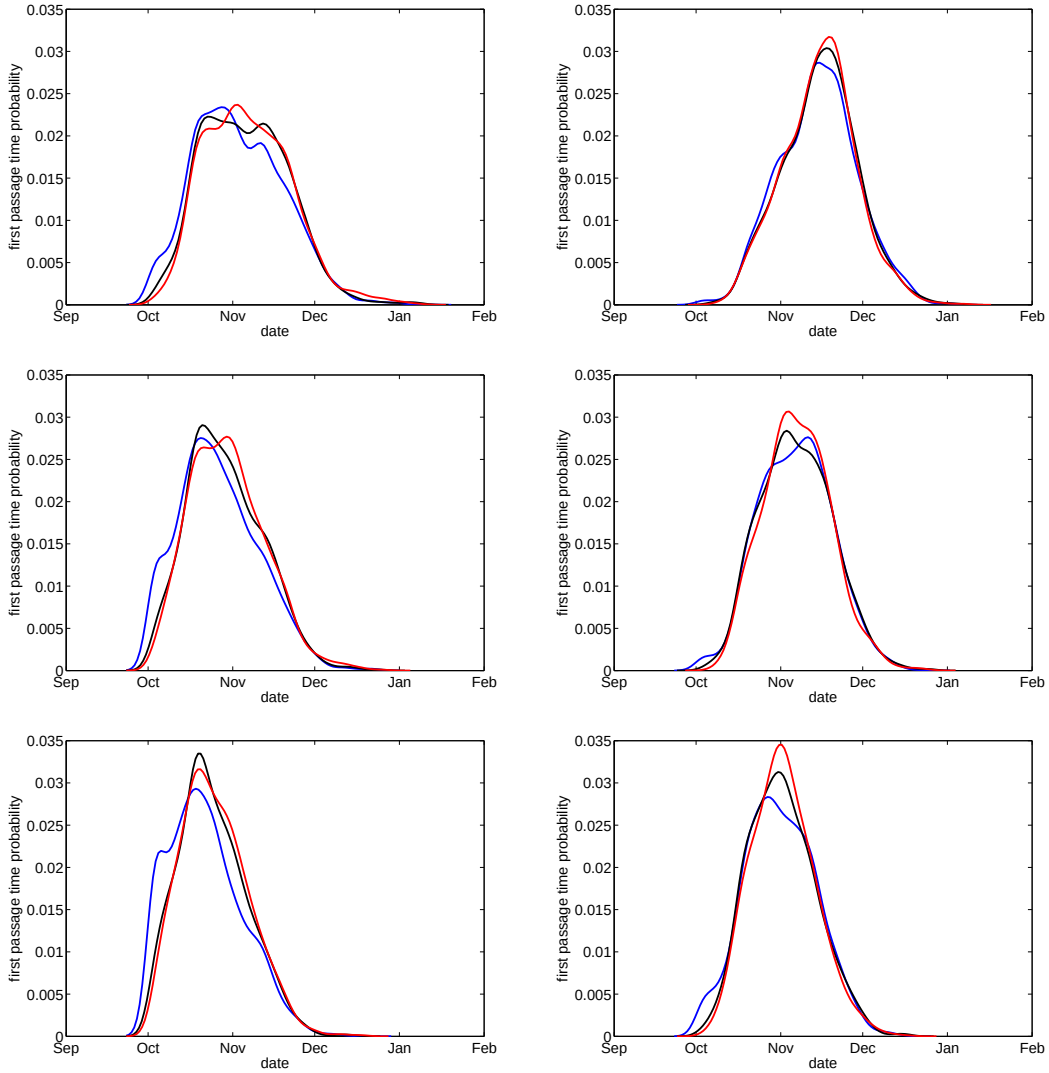


Figure B.21: Initial date: November 1st

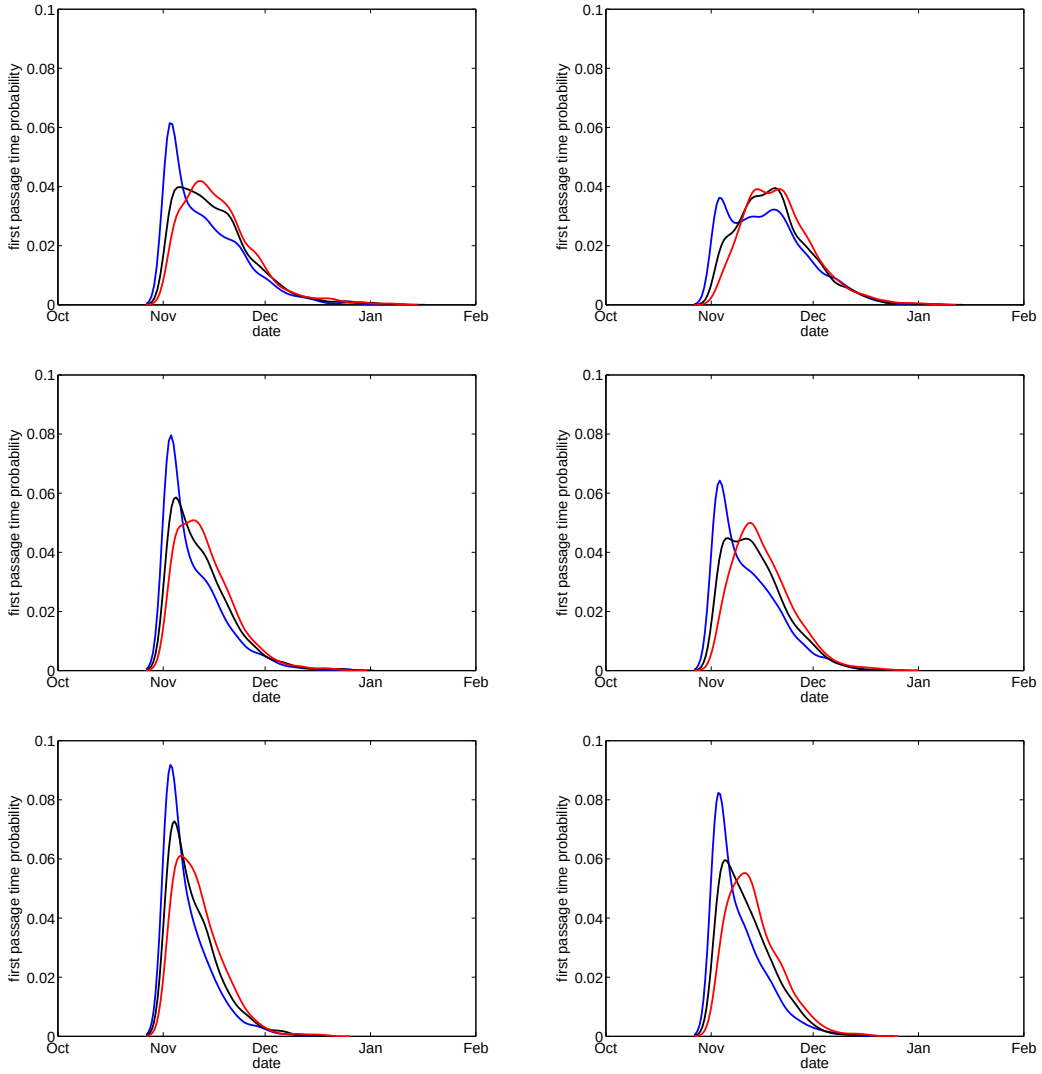
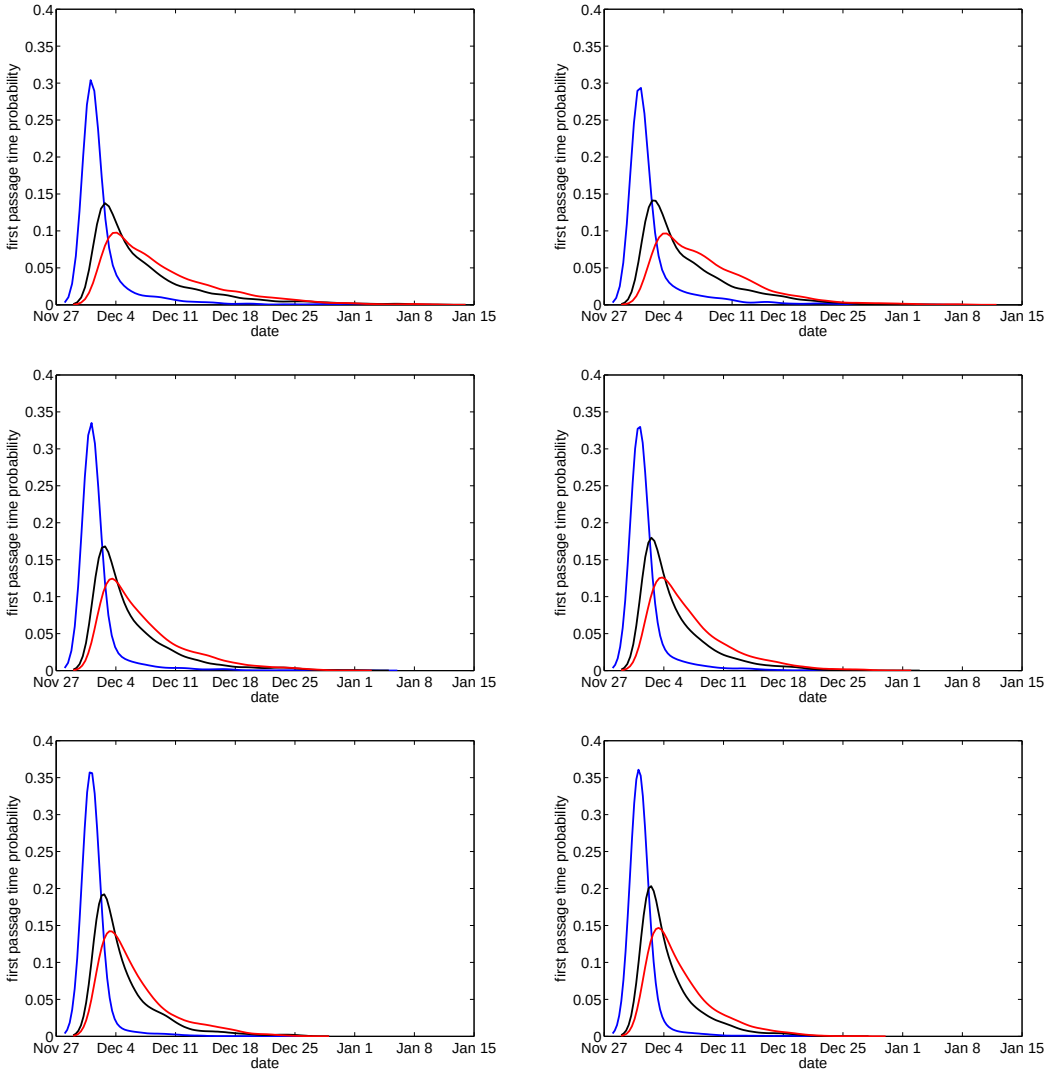


Figure B.22: Initial date: December 1st



B.6 Influence of the initial anomaly decile (improved time series)

In this section, we analysed the influence of the initial anomaly decile on the first passage time distribution to frost for the improved time series. Since here the initial years differ by the observed temperature trend and thus influence the conversion of anomalies back into temperatures, we again chose three different initial years as well as different initial dates for our analysis as also done in the main part of this thesis in Sec. 5.3.2.

The top row was generated starting in 1999, the initial year with the largest observed temperature trend value, the center row uses 1965 as initial year, one whose trend value is closest to the average trend value. The bottom row uses 1941 as initial year, representing the minimum trend value.

In order to estimate the influence of statistical fluctuations due to the finite length of the time series, we also plotted the mean of 1000 bootstrap samples of the time series as a continuous line, as well as error bars representing the width of two standard deviations of the bootstrap calculation (see Sec. 2.2.6 for details).

B.6.1 Distribution location

We first look again at measures of the distribution location. As stated before, while the median is the more robust quantity especially with respect to outliers, it changes abruptly in the case of a bimodal distribution with a strict separation of peaks that lie far apart as is the case for some initial conditions of the first passage time distribution to frost. Indeed, then the median jumps from one peak to the other as soon as the second peak acquires more relative weight than the first.

Since this is mainly the case for initial dates in spring, we chose the mean for those cases. In summer, the two measures are almost indistinguishable. For the sake of clarity, we limited the display to only one of the two throughout this subsection.

The mean is again measured as the average number of days until frost next occurs, while the median is measured in the number of days until first frost has occurred in 50% of all cases.

Figure B.23: Initial date: February 14$^{\text{th}}$ (left) and April 1$^{\text{st}}$ (right)

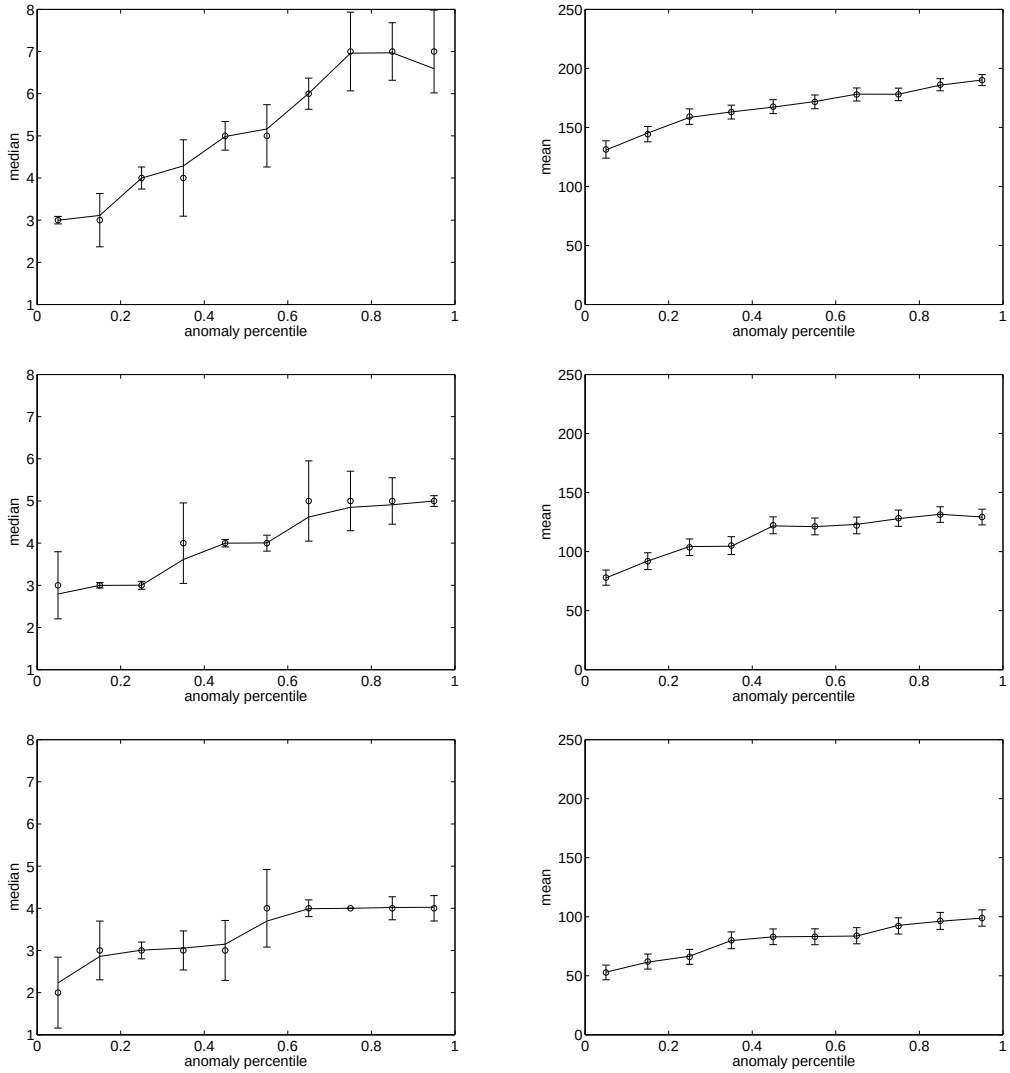


Figure B.24: Initial date: May 15$^{\text{th}}$ (left) and July 1$^{\text{st}}$ (right)

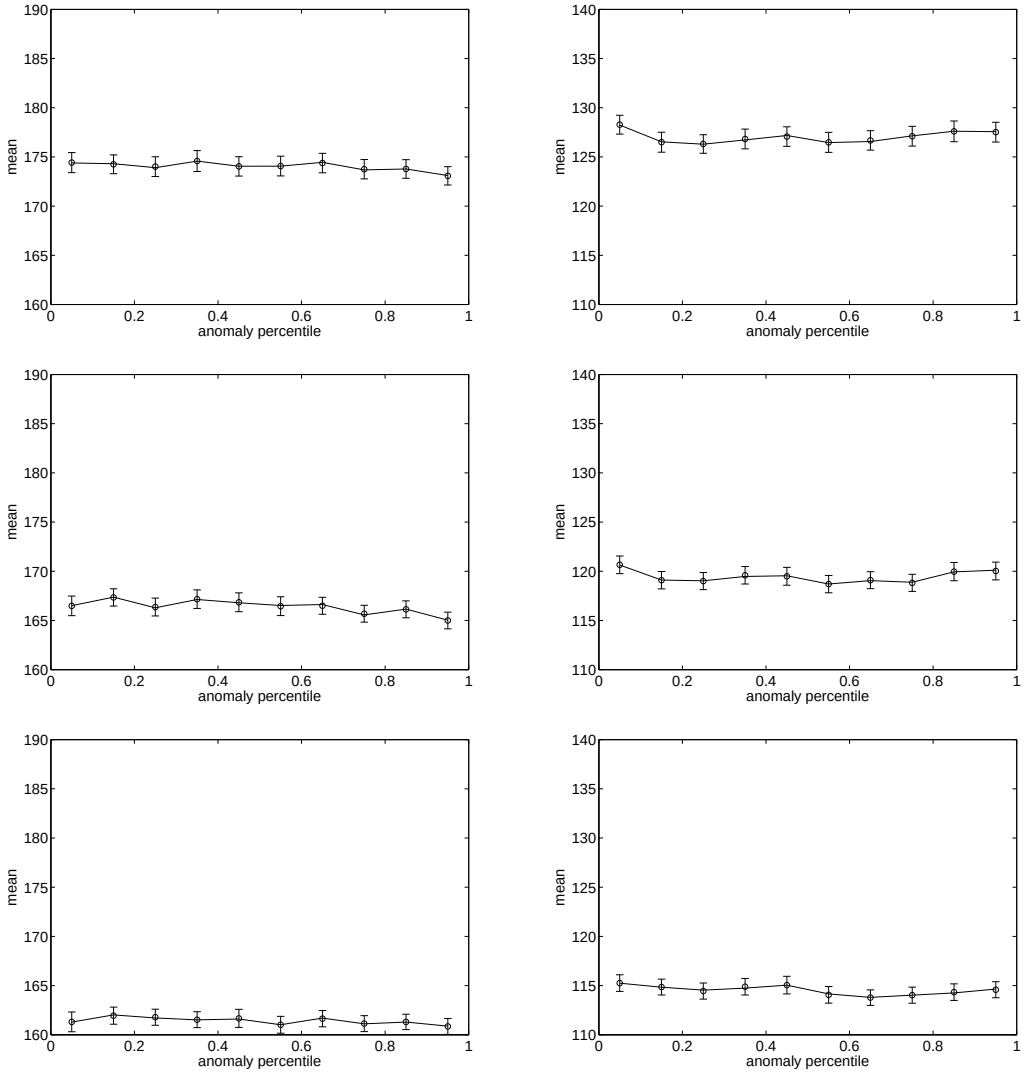


Figure B.25: Initial date: September 15$^{\text{th}}$ (left) and October 1$^{\text{st}}$ (right)

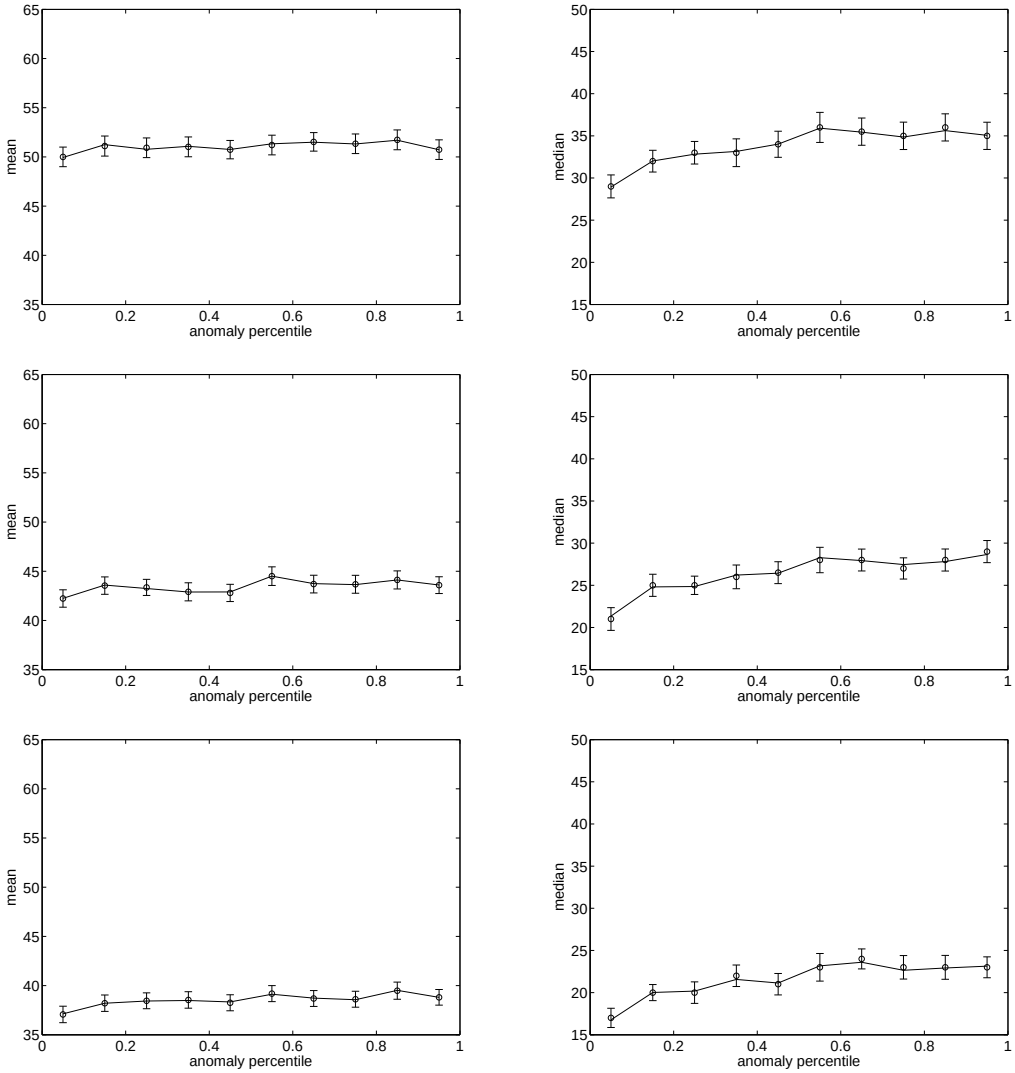
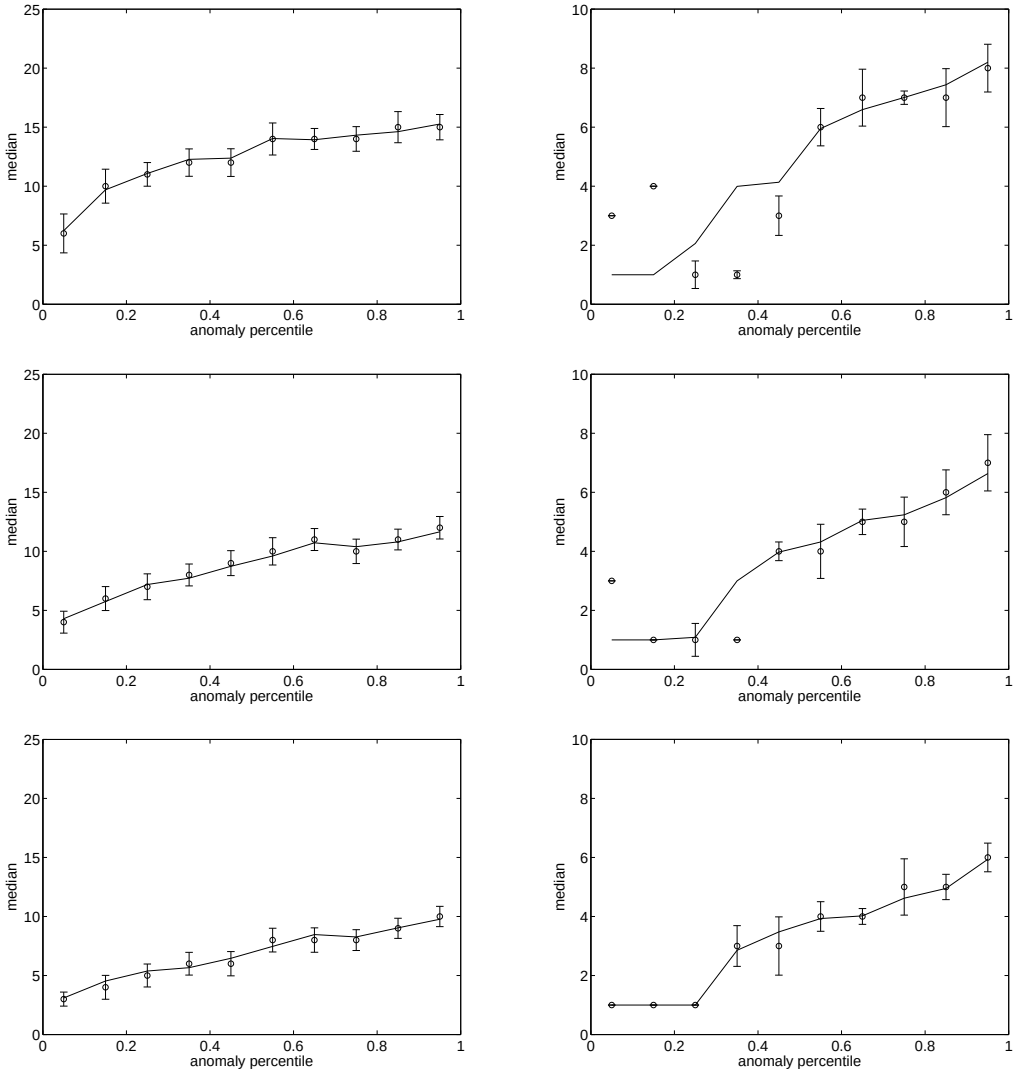


Figure B.26: Initial date: November 1st (left) and December 1st (right)



B.6.2 Distribution spread

We next turn to measures of the distribution spread. The interquartile range is the more robust quantity especially with respect to outliers, but it changes abruptly in the case of a bimodal distribution with a strict separation of peaks. Indeed, then it jumps from describing the spread of the larger peak to describing the distance between the peaks as soon as the second peak contains more than a quarter of all data points. Since this is mainly the case for initial dates in spring, we chose the standard deviation for those cases and the interquartile range in winter when the number of data points for the estimation is small. Both the std and the IQR are measured in units of days.

Figure B.27: Initial date: February 14^{th} (left) and April 1^{st} (right)

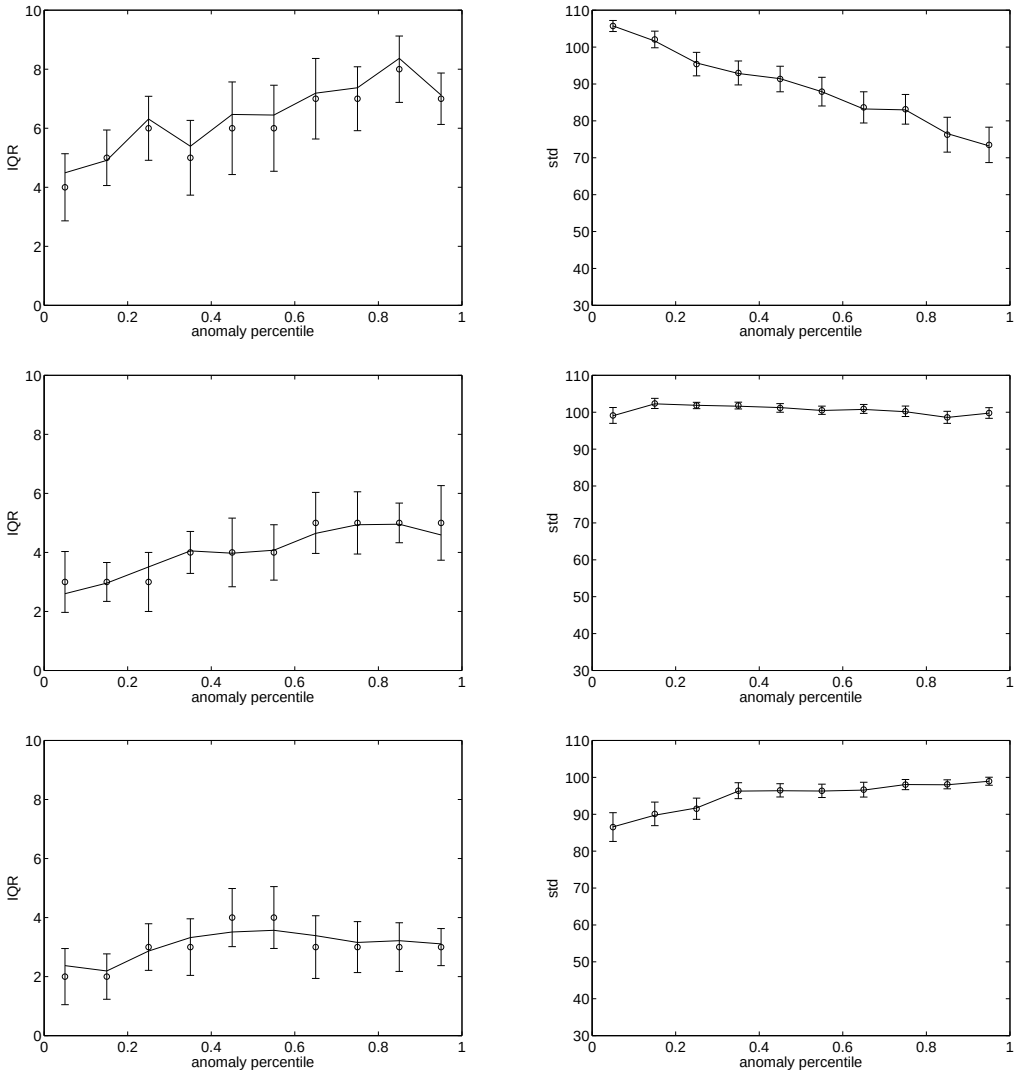


Figure B.28: Initial date: May 15$^{\text{th}}$ (left) and July 1$^{\text{st}}$ (right)

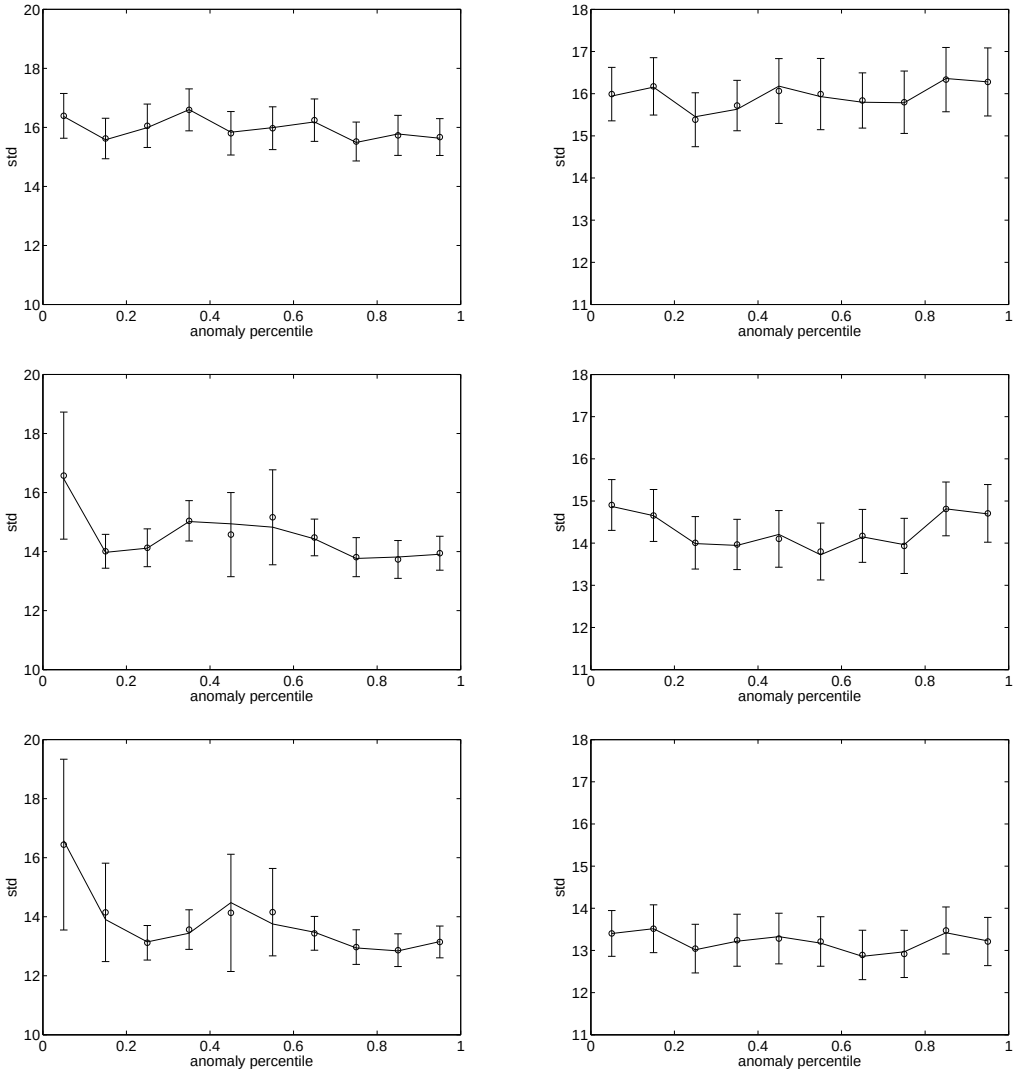


Figure B.29: Initial date: September 15$^{\text{th}}$ (left) and October 1$^{\text{st}}$ (right)

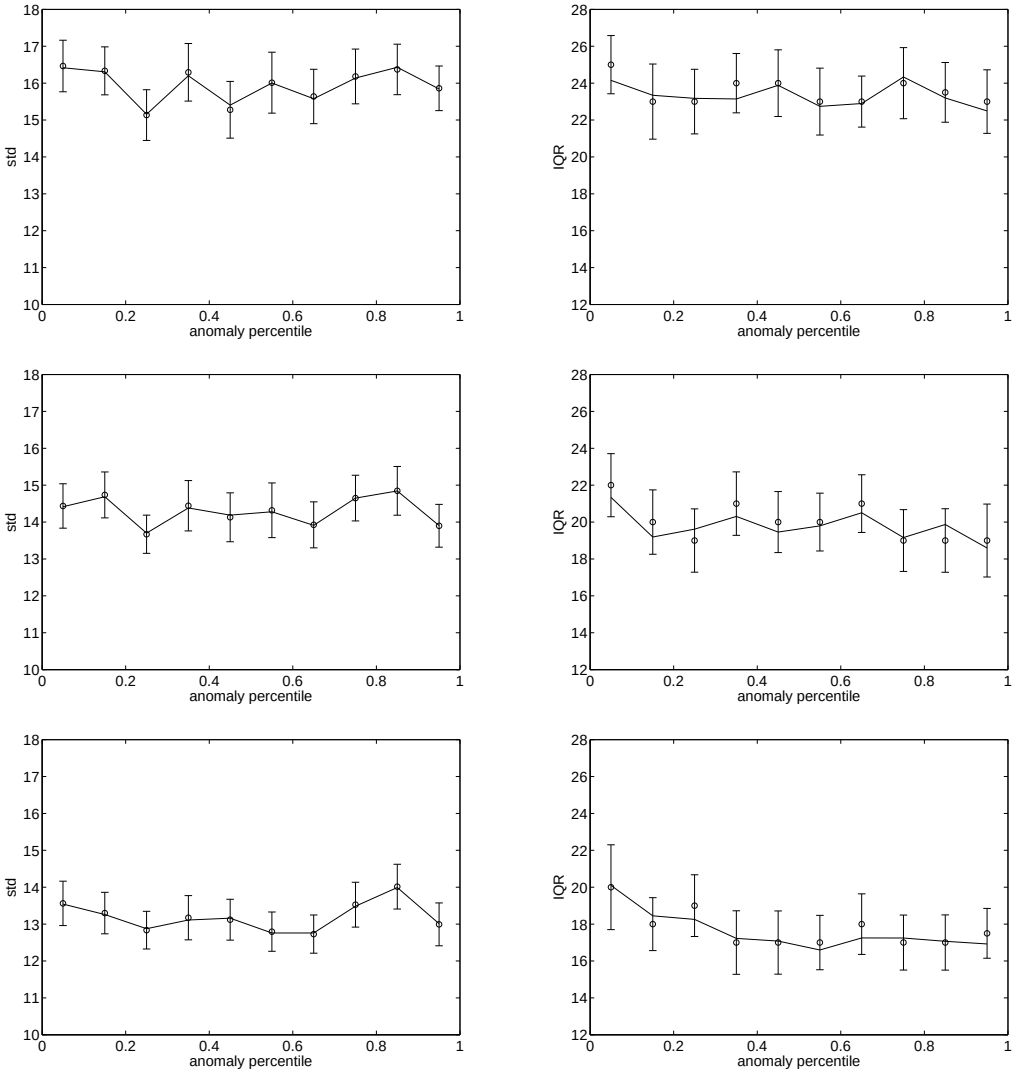
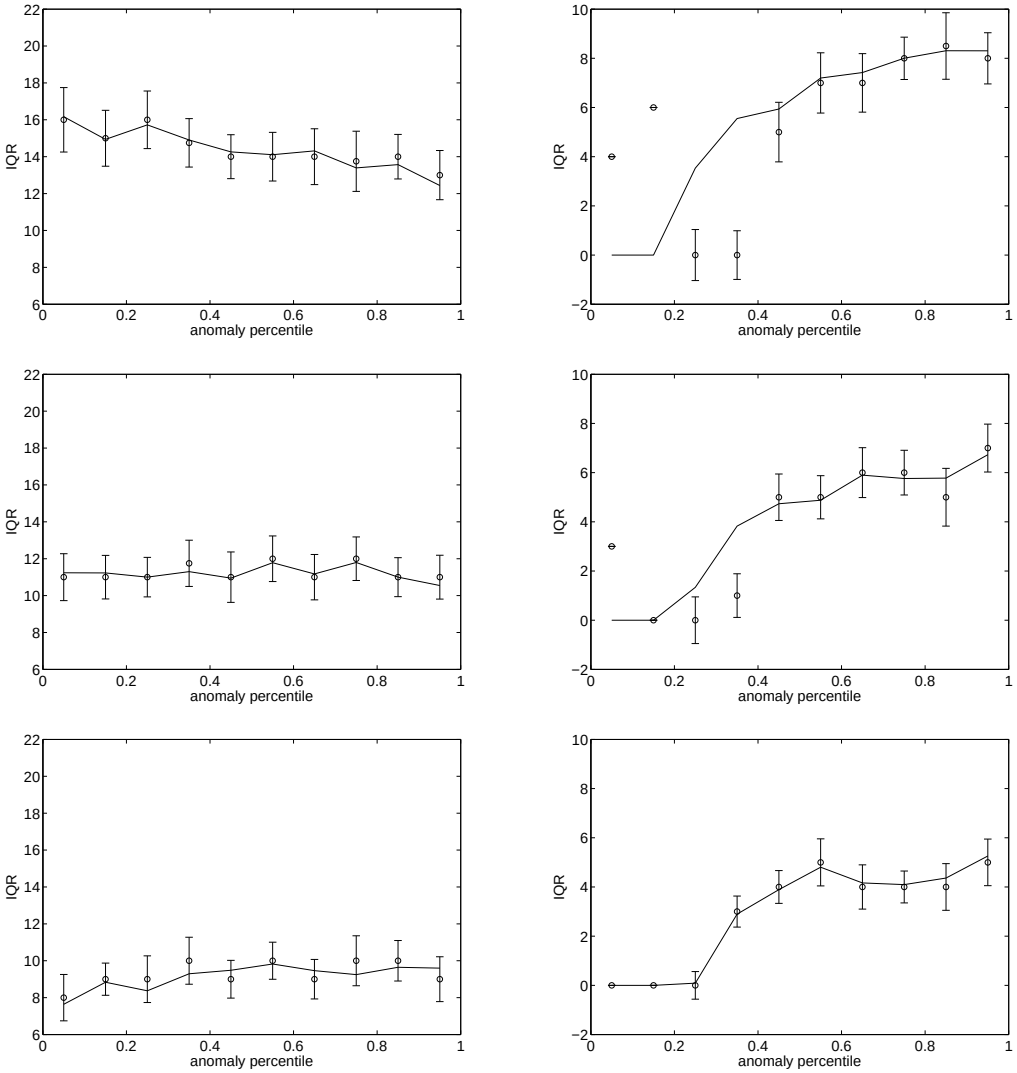


Figure B.30: Initial date: November 1st (left) and December 1st (right)



B.6.3 Date of maximum first frost probability

The third interesting summary measure of the first passage time probability distribution to frost is the date of maximum first frost probability. In the case of strong bimodality with peaks in spring and late autumn, two peaks were determined. Note that since estimating the location of the peak(s) was done using a kernel density estimate for the probability distribution, the peak location is not necessarily on integer days.

Figure B.31: Initial date: February 14$^{\text{th}}$ (left) with zoom into the earlier peak (right)

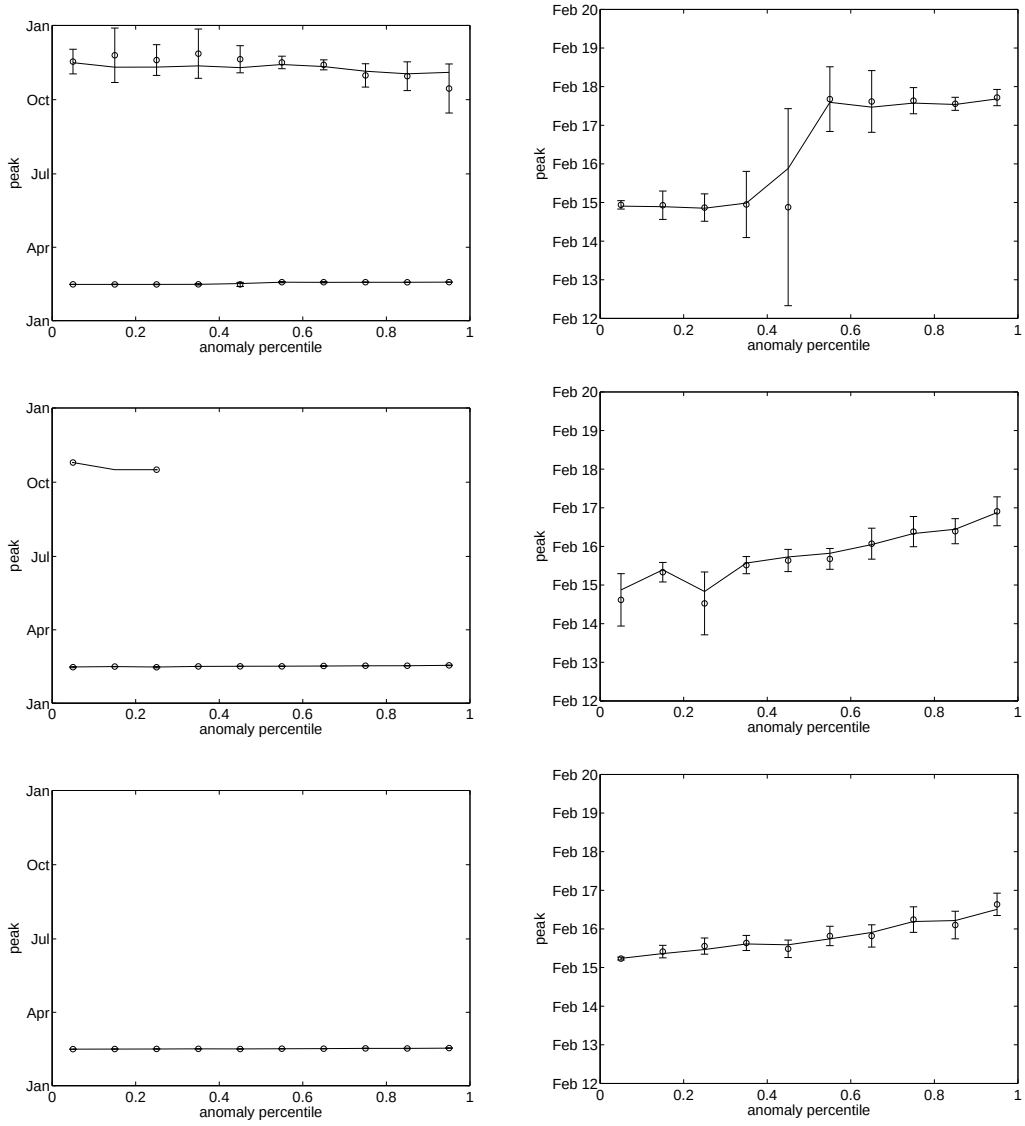


Figure B.32: Initial date: April 1st (left) with zoom into the earlier peak (right)

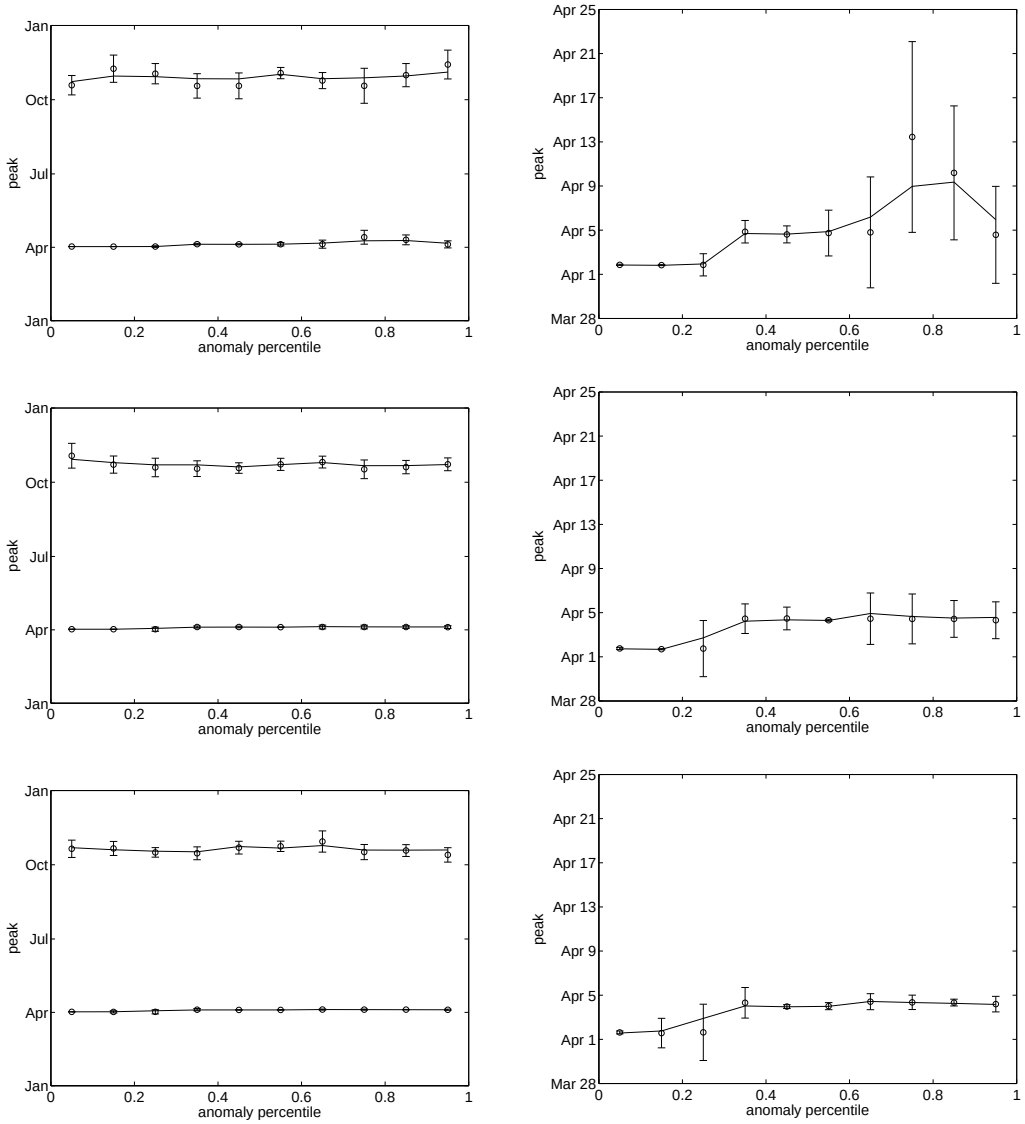


Figure B.33: Initial date: April 1st - zoom into the later peak (left) and July 1st (right)

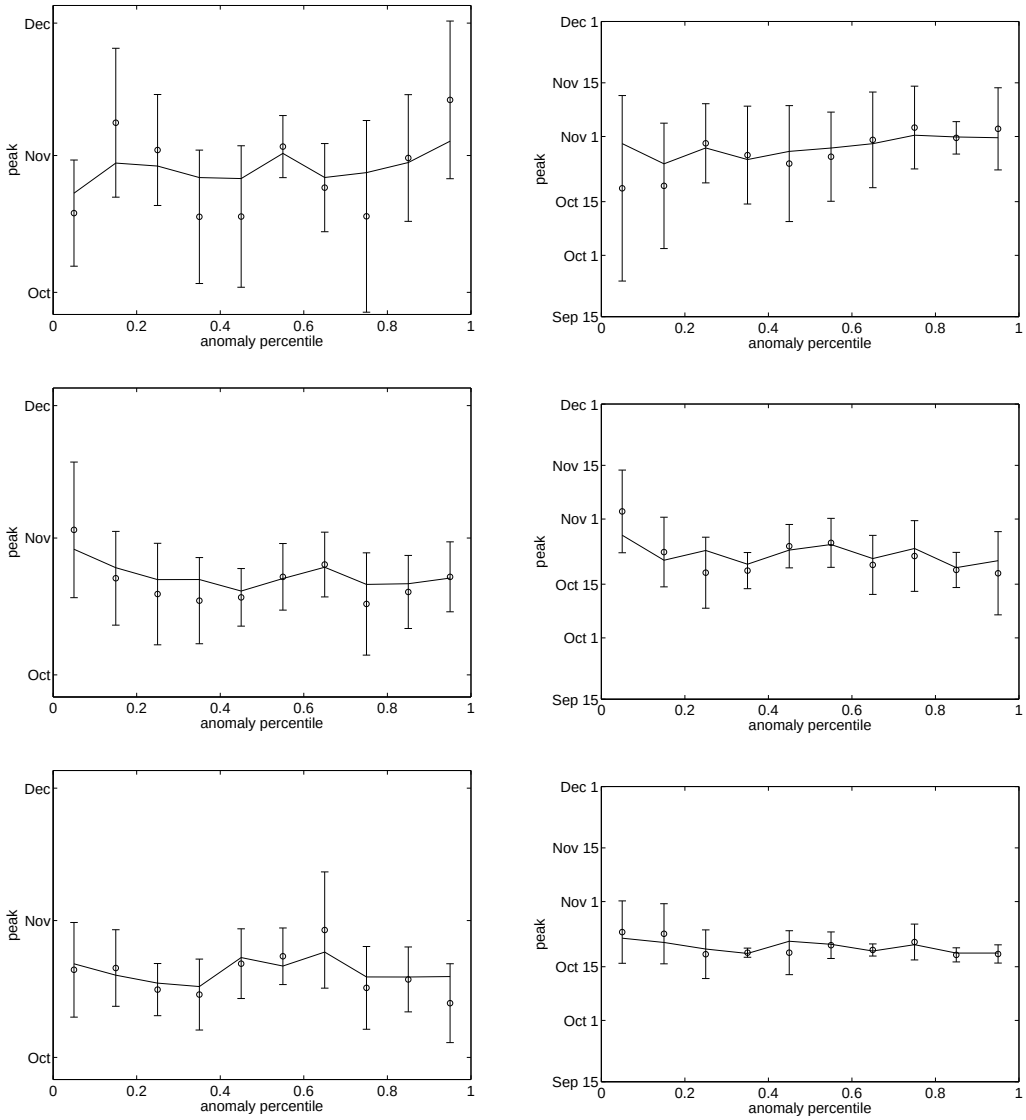


Figure B.34: Initial date: May 15^{th} (left) with zoom into the later peak (right)

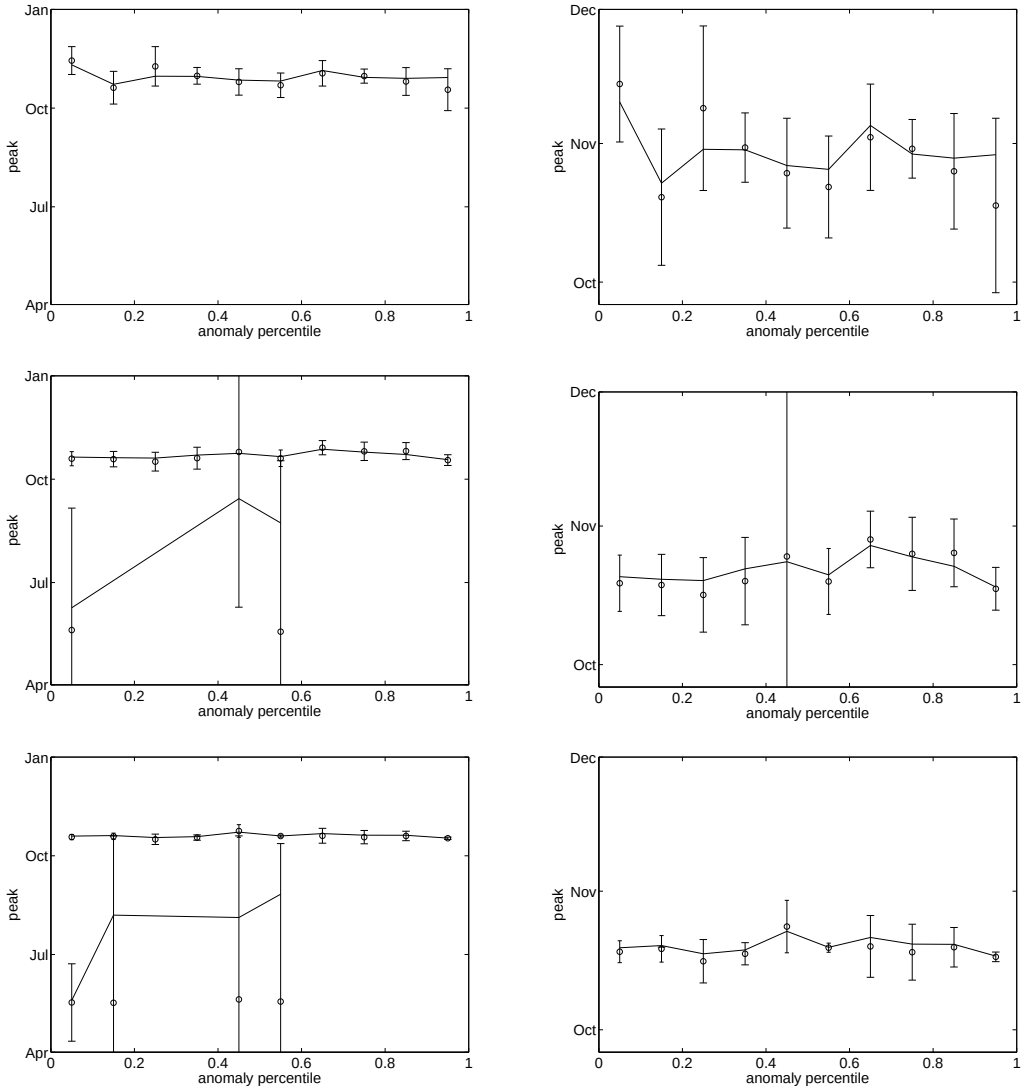


Figure B.35: Initial date: September 15$^{\text{th}}$ (left) and October 1$^{\text{st}}$ (right)

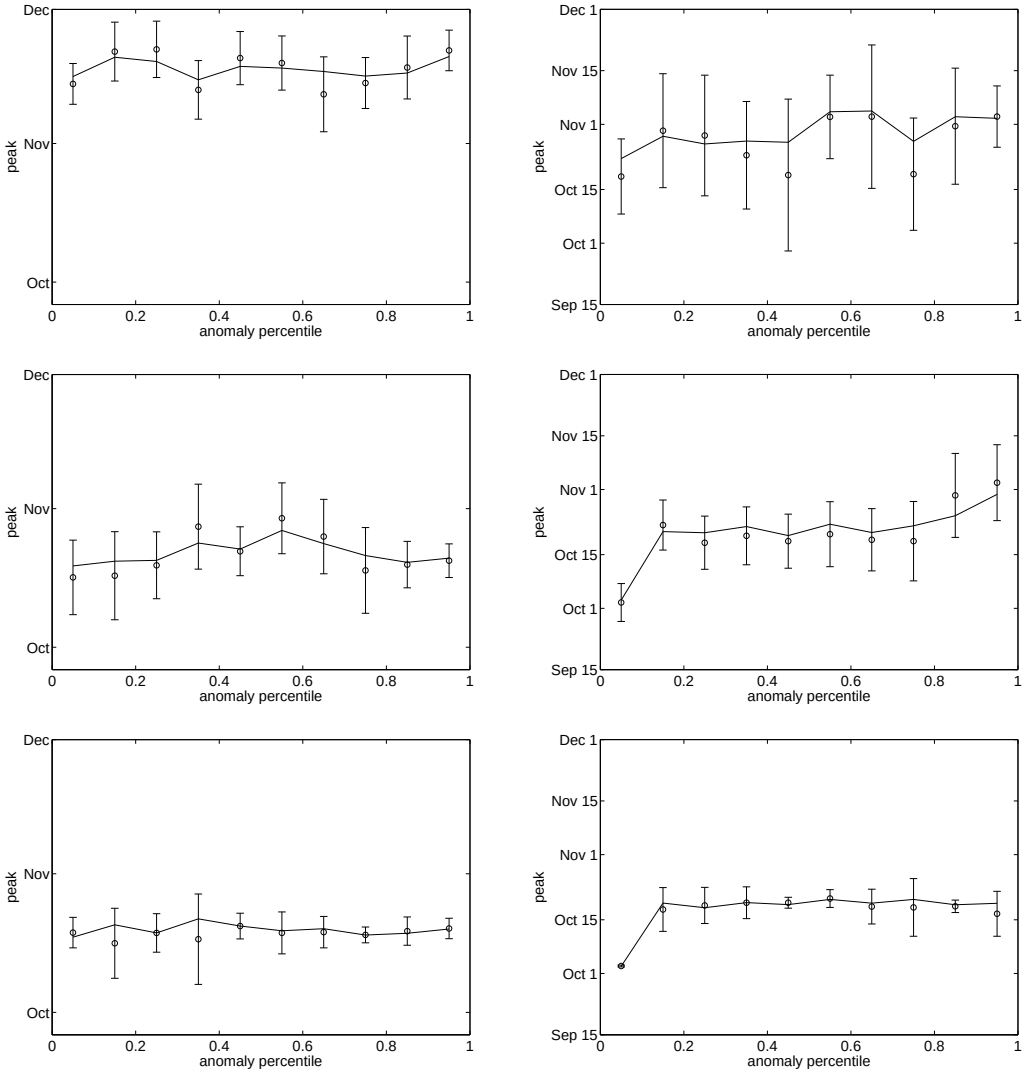
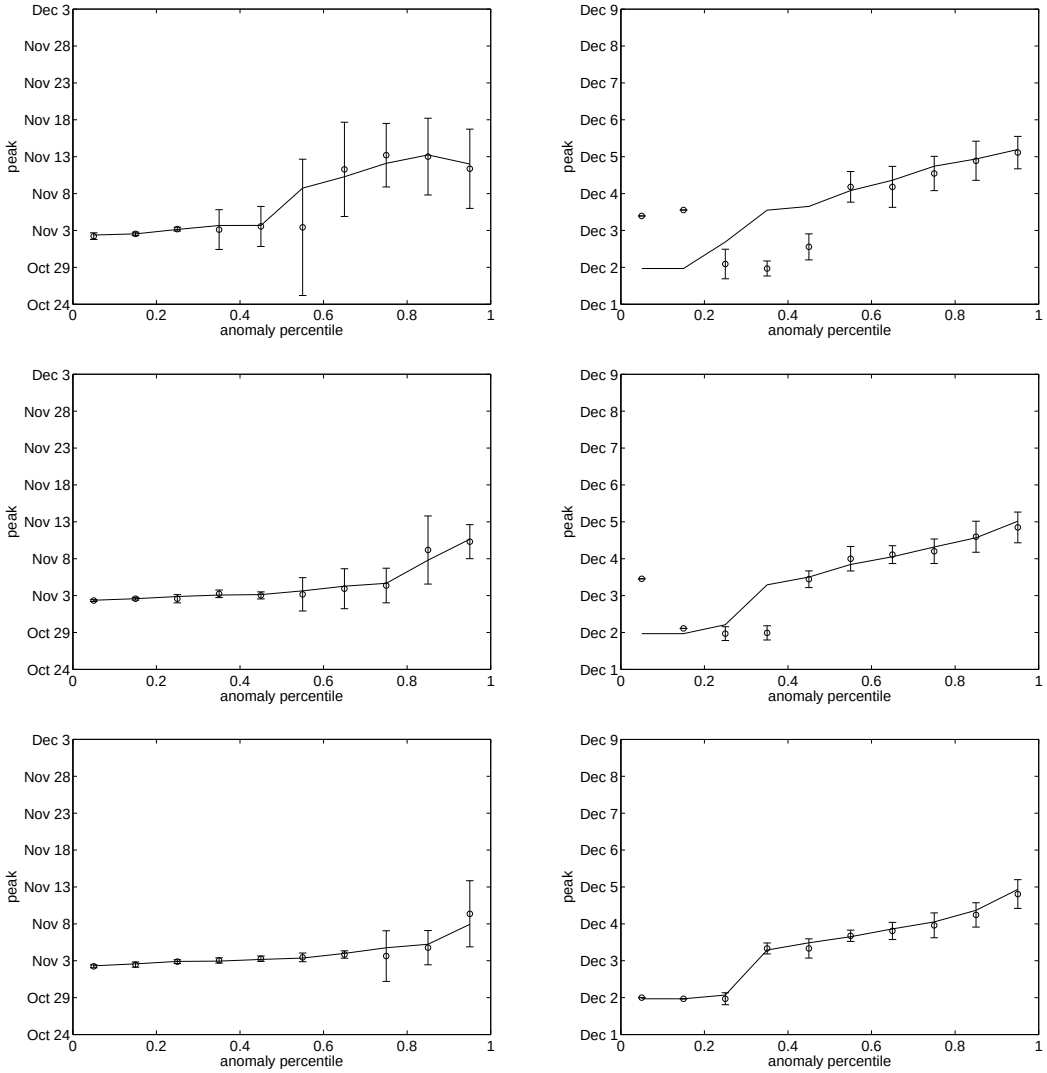


Figure B.36: Initial date: November 1st (left) and December 1st (right)



Bibliography

[1] Deutscher Wetterdienst (DWD). Climate Data for Germany in Standard Format. http://www.dwd.de/, 2011.

[2] T. N. Palmer, A. Alessandri, U. Andersen, P. Cantelaube, M. Davey, P. Délécluse, M. Déqué, E. Diez, F. J. Doblas-Reyes, H. Feddersen, R. Graham, S. Gualdi, J.-F. Guérémy, R. Hagedorn, M. Hoshen, N. Keenlyside, M. Latif, A. Lazar, E. Maisonnave, V. Marletto, A. P. Morse, B. Orfila, P. Rogel, J.-M. Terres, and M. C. Thomson. Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER). *Bull. Amer. Met. Soc.*, 85(6):853–872, 2004.

[3] E. N. Lorenz. Atmospheric Predictability as Revealed by Naturally Occurring Analogues. *J. Atmos. Sci.*, 26(4):636–646, 1969.

[4] G. K. Vallis. Mechanisms of Climate Variability from Years to Decades. In T. Palmer and P. Williams, editors, *Stochastic Physics and Climate Modelling*, pages 1–34. Cambridge University Press, 2010.

[5] M. Déqué. Continuous Variables. In I. T. Jolliffe and D. B. Stephenson, editors, *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, pages 97–119. Wiley, Chichester, 2003.

[6] K. H. Oberthier. Wetternews. http://www.wetter.net/wetternews/wetter-fruehherbst-oder-spaetsommer-das-auf-und-ab-geht-weiter-5094.html, 7 August 2011.

[7] R. Pielke Jr. and R. E. Carbone. Weather Impacts, Forecasts, and Policy: An Integrated Perspective. *Bull. Amer. Met. Soc.*, 83(3):393–403, 2002.

[8] D. S. Wilks and A. H. Murphy. A Decision-Analytic Study of the Joint Value of Seasonal Precipitation and Temperature Forecasts in a Choice-of-Crop Problem. *Atmosphere-Ocean*, 24(4):353–368, 1986.

[9] R. Blench. Seasonal Climatic Forecasting: Who can use it and how should it be disseminated? *Natural Resource Perpectives*, 47, 1999.

[10] J. W. Hansen. Integrating Seasonal Climate Prediction and Agricultural Models for Insights into Agricultural Practice. *Phil. Trans. R. Soc. B*, 360(1463):2037–2047, 2005.

[11] J. W. Mjelde, S. T. Sonka, B. L. Dixon, and P. J. Lamb. Valuing Forecast Characteristics in a Dynamic Agricultural Production System. *Amer. J. Agricult. Econ.*, 70(3):674–684, 1988.

[12] P. Calanca, D. Bolius, A. P. Weigel, and M. A. Liniger. Application of Long-Range Weather Forecasts to Agricultural Decision Problems in Europe. *J. Agricult. Sci.*, 149 (1):15–22, 2011.

[13] M. V. K. Sivakumar. Climate Prediction and Agriculture: Current Status and Future Challenges. *Clim. Res.*, 33(1):3–17, 2006.

[14] P. C. McIntosh, M. J. Pook, J. S. Risbey, S. N. Lisson, and M. Rebbeck. Seasonal Climate Forecasts for Agriculture: Towards Better Understanding and Value. *Field Crops Research*, 104(1-3):130–138, 2007.

[15] H. Meinke and R. C. Stone. Seasonal and Inter-Annual Climate Forecasting: The New Tool for Increasing Preparedness to Climate Variability and Change in Agricultural Planning and Operations. *Clim. Change*, 70(1-2):221–253, 2005.

[16] L. Goddard, S. J. Mason, S. E. Zebiak, C. F. Ropelewski, R. Basher, and M. A. Cane. Current Approaches to Seasonal-to-Interannual Climate Predictions. *Int. J. Climatol.*, 21 (9):1111–1152, 2001.

[17] L. Zeng. Weather Derivatives and Weather Insurance: Concept, Application, and Analysis. *Bull. Amer. Met. Soc.*, 81(9):2075–2082, 2000.

[18] R. J. Murnane, M. Crowe, A. Eustis, S. Howard, J. Koepsell, T. Leffler, and R. Livezey. The Weather Risk Management Industry's Climate Forecast and Data Needs. *Bull. Amer. Met. Soc.*, 83(8):1193–1198, 2002.

[19] J. Cherry, H. Cullen, M. Visbeck, A. Small, and C. Uvo. Impacts of the North Atlantic Oscillation on Scandinavian Hydropower Production and Energy Markets. *Water Res. Management*, 19(6):673–691, 2005.

[20] G. R. McGregor, M. Cox, Y. Cui, Z. Cui, M. K. Davey, R. F. Graham, and A. Brookshaw. Winter-Season Climate Prediction for the U.K. Health Sector. *J. Appl. Meteor. Climatol.*, 45(12):1782–1792, 2006.

[21] M. C. Thomson, T. Palmer, A. P. Morse, M. Cresswell, and S. J. Connor. Forecasting Disease Risk with Seasonal Climate Predictions. *The Lancet*, 355(9214):1559–1560, 2000.

[22] B. Kirtman and A. Pirani. WCRP Position Paper on Seasonal Prediction. WCRP Informal Report No.3/2008, ICPO Publication No.127, World Climate Research Programme, 2008.

[23] P. Mailier. Can We Really Trust Long-Range Weather Forecasts? http://space.fmi.fi/Verification2009/presentations/WEDNESDAY/O8.3_Maillier.pdf, June 2009.

[24] A. G. Barnston, S. Li, S. J. Mason, D. G. DeWitt, L. Goddard, and X. Gong. Verification of the First 11 Years of IRI's Seasonal Climate Forecasts. *J. Appl. Meteor. Climatol.*, 49 (3):493–520, 2010.

[25] M. Patarcic and C. Brankovic. Skill of 2-m Temperature Seasonal Forecasts over Europe in ECMWF and RegCM Models. *Mon. Wea. Rev.*, 140(4):1326–1346, 2012.

[26] A. P. Weigel, D. Baggenstos, and M. A. Liniger. Probabilistic Verification of Monthly Temperature Forecasts. *Mon. Wea. Rev.*, 136(12):5162–5182, 2008.

[27] M. E. Shongwe, C. A. T. Ferro, C. A. S. Coelho, and G. J. van Oldenborgh. Predictability of Cold Spring Seasons in Europe. *Mon. Wea. Rev.*, 135(12):4185–4201, 2007.

[28] World Meteorological Organisation (WMO). Standard Verification System (SVS) for Long-Range Forecasts (LRF). http://www.wmo.int/pages/prog/www/DPS/SVS-for-LRF.html, 2000.

[29] T. N. Stockdale, O. Alves, G. Boer, M. Deque, Y. Ding, A. Kumar, K. Kumar, W. Landman, S. Mason, P. Nobre, A. Scaife, O. Tomoaki, and W.-T. Yun. Understanding and Predicting Seasonal to Interannual Climate Variability - The Producer Perspective. Technical report, WMO World Climate Conference 3, 2009.

[30] W. A. Müller, C. Appenzeller, and M. Latif. NAO und Vorhersagbarkeit. *promet*, 34(3): 130–137, 2008.

[31] J. F. Eichner, E. Koscielny-Bunde, A. Bunde, S. Havlin, and H.-J. Schellnhuber. Power-Law Persistence and Trends in the Atmosphere: A Detailed Study of Long Temperature Records. *Phys. Rev. E*, 68(4):046133, 2003.

[32] A. Király, I. Bartos, and I. M. Jánosi. Correlation Properties of Daily Temperature Anomalies over Land. *Tellus*, 58A(5):593–600, 2006.

[33] D. Vyushin, I. Zhidkov, S. Havlin, A. Bunde, and S. Brenner. Volcanic Forcing Improves Atmosphere-Ocean Coupled General Circulation Model Scaling Performance. *Geophys. Res. Lett.*, 31(10):L10206, 2004.

[34] E. Koscielny-Bunde, H. E. Roman, A. Bunde, S. Havlin, and H.-J. Schellnhuber. Long-Range Power-Law Correlations in Local Daily Temperature Fluctuations. *Phil. Mag. B*, 77(5):1331–1340, 1998.

[35] A. Shabbar and A. G. Barnston. Skill of Seasonal Climate Forecasts in Canada Using Canonical Correlation Analysis. *Mon. Wea. Rev.*, 124(10):2370–2385, 1996.

[36] D. Lavers, L. Luo, and E. F. Wood. A Multiple Model Assessment of Seasonal Climate Forecast Skill for Applications. *Geophys. Res. Lett.*, 36(23):L23711, 2009.

[37] B. Abraham and J. Ledolter. *Statistical Methods for Forecasting*. Wiley Series in Probability and Mathematical Statistics. Wiley, New York, 1983.

[38] H. von Storch and F. W. Zwiers. *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge, 2001.

[39] G. J. van Oldenborgh, M. A. Balmaseda, L. Ferranti, T. N. Stockdale, and D. L. T. Anderson. Did the ECMWF Seasonal Forecast Model Outperform Statistical ENSO Forecast Models over the Last 15 Years? *J. Climate*, 18(16):3240–3249, 2005.

[40] A. J. Majda, C. Franzke, and B. Khouider. An Applied Mathematics Perspective on Stochastic Modelling for Climate. In T. Palmer and P. Williams, editors, *Stochastic Physics and Climate Modelling*, pages 73–104. Cambridge University Press, 2010.

[41] I. Horenko, R. Klein, S. Dolaptchiev, and C. Schütte. Automated Generation of Reduced Stochastic Weather Models I: Simultaneous Dimension and Model Reduction for Time Series Analysis. *Multiscale Model. Simul.*, 6(4):1125–1145, 2008.

[42] P. Sura. Noise-Induced Transitions in a Barotropic β-Plane Channel. *J. Atmos. Sci.*, 59 (1):97–110, 2002.

[43] P. Sura, M. Newman, C. Penland, and P. Sardeshmukh. Multiplicative Noise and Non-Gaussianity: A Paradigm for Atmospheric Regimes? *J. Atmos. Sci.*, 62(5):1391–1409, 2005.

[44] A. J. Majda, C. L. Franzke, A. Fischer, and D. T. Crommelin. Distinct Metastable Atmospheric Regimes Despite Nearly Gaussian Statistics: A Paradigm Model. *Proc. Natl. Acad. Sci. USA*, 103(22):8309–8314, 2006.

[45] D. Entekhabi, I. Rodriguez-Iturbe, and R. L. Bras. Variability in Large-Scale Water Balance with Land Surface-Atmosphere Interaction. *J. Climate*, 5(8):798–813, 1992.

[46] A. K. Garber, N. R. Moloney, and H. Kantz. Hopping Over a Heat Barrier. *Phys. Rev. E*, 83(3):031134, 2011.

[47] C. Gardiner. *Stochastic Methods. A Handbook for the Natural and Social Sciences.* Springer Series in Synergetics. Springer, 4th edition, 2009.

[48] H. Risken. *The Fokker-Planck Equation. Methods of Solution and Applications.* Springer Series in Synergetics. Springer, 2nd edition, 1996.

[49] D. S. Wilks. *Statistical Methods in the Atmospheric Sciences*, volume 91 of *International Geophysics Series.* Academic Press, Burlington, 2d edition, 2006.

[50] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control.* Prentice-Hall International, New Jersey, 3rd edition, 1994.

[51] P. D. Welch. The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms. *IEEE Trans. Audio and Electroacoustics*, AU-15(2):70–73, 1967.

[52] H. Akaike. A New Look at the Statistical Model Identification. *IEEE Trans. Auto. Control*, AC-19(6):716–723, 1974.

[53] E. J. Hannan and B. G. Quinn. The Determination of the Order of an Autoregression. *J. Roy. Stat. Soc. B*, 41(2):190–195, 1979.

[54] G. Schwarz. Estimating the Dimension of a Model. *Ann. Statist.*, 6(2):461–464, 1978.

[55] D. Sheskin. *Handbook of Parametric and Nonparametric Statistical Procedures.* CRC Press, Boca Raton, 2nd edition, 1997.

[56] G. M. Ljung and G. E. P. Box. On a Measure of Lack of Fit in Time Series Models. *Biometrika*, 65(2):297–303, 1978.

[57] G. E. P. Box and D. A. Pierce. Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. *J. Amer. Stat. Assoc.*, 65 (332):1509–1526, 1970.

[58] P. M. Hawkins. Testing a Sequence of Observations for a Shift in Location. *J. Amer. Stat. Assoc.*, 72(357):180–186, 1977.

[59] H. Alexandersson. A Homogeneity Test Applied to Precipitation Data. *J. Climatol.*, 6 (6):661–675, 1986.

[60] J. B. Wijngaard, A. M. G. Klein Tank, and G. P. Können. Homogeneity of 20th Century European Daily Temperature and Precipitation Series. *Int. J. Climatol.*, 23(6):679–692, 2003.

[61] T. A. Buishand. Some Methods for Testing the Homogeneity of Rainfall Records. *J. Hydr.*, 58(1):11–27, 1982.

[62] A. N. Pettitt. A Non-Parametric Approach to the Change-Point Problem. *J. Roy. Stat. Soc. C*, 28(2):126–135, 1979.

[63] D. Rybski and J. Neumann. A Review on the Pettitt Test. In J. P. Kropp and H.-J. Schellnhuber, editors, *In Extremis: Disruptive Events and Trends in Climate and Hydrology*, pages 203–213. Springer, Berlin, 2011.

[64] J. von Neumann, R. H. Kent, H. R. Bellinson, and B. I. Hart. The Mean Square Successive Difference. *Ann. Math. Stat.*, 12(2):153–162, 1941.

[65] J. von Neumann. Distribution of the Ratio of the Mean Square Successive Difference to the Variance. *Ann. Math. Stat.*, 12(4):367–395, 1941.

[66] C. M. Jarque and A. K. Bera. A Test for Normality of Observations and Regression Residuals. *Internat. Stat. Rev.*, 55(2):163–172, 1987.

[67] P. Deb and M. Sefton. The Distribution of a Lagrange Multiplier Test of Normality. *Economics Lett.*, 51(2):123–130, 1996.

[68] C. L. Wood and M. M. Altavela. Large-Sample Results for Kolmogorov-Smirnov Statistics for Discrete Distributions. *Biometrika*, 65(1):235–239, 1978.

[69] F. J. Massey Jr. The Kolmogorov-Smirnov Test for Goodness of Fit. *J. Amer. Stat. Assoc.*, 46(253):68–78, 1951.

[70] J. E. Walsh. Bounded Probability Properties of Kolmogorov-Smirnov and Similar Statistics for Discrete Data. *Ann. Inst. Stat. Math.*, 15(1):153 – 158, 1963.

[71] H. Levene. Robust Tests for Equality of Variances. In I. Olkin and et al., editors, *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pages 278–292. Stanford University Press, 1960.

[72] M. B. Brown and A. B. Forsythe. Robust Tests for the Equality of Variances. *J. Amer. Stat. Assoc.*, 69(346):364–367, 1974.

[73] B. Efron and R. Tibshirani. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. *Statistical Science*, 1(1):54–77, 1986.

[74] S. J. Mason. Understanding Forecast Verification Statistics. *Meteor. Appl.*, 15(1):31–40, 2008.

[75] I. T. Jolliffe. Uncertainty and Inference for Verification Measures. *Wea. Forecasting*, 22 (3):637–650, 2007.

[76] Joint Working Group on Forecast Verification Research. Forecast Verification: Issues, Methods and FAQ. http://www.cawcr.gov.au/projects/verification/, September 2012.

[77] F. W. Zwiers. The Effect of Serial Correlation on Statistical Inferences Made with Resampling Procedures. *J. Climate*, 3(12):1452–1461, 1990.

[78] A. H. Murphy. What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Wea. Forecasting*, 8(2):281–293, 1993.

[79] I. T. Jolliffe and D. B. Stephenson. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, chapter 1: Introduction, pages 1–12. Wiley, Chichester, 2003.

[80] R. E. Livezey. The Evaluation of Forecasts. In H. von Storch and A. Navarra, editors, *Analysis of Climate Variability. Applications of Statistical Techniques*, pages 179–198. Springer, Berlin, 1999.

[81] A. H. Murphy, B. G. Brown, and Y.-S. Chen. Diagnostic Verification of Temperature Forecasts. *Wea. Forecasting*, 4(4):485–501, 1989.

[82] G. W. Brier. Verification of Forecasts Expressed in Terms of Probability. *Mon. Wea. Rev.*, 78(1):1–3, 1950.

[83] J. Bröcker and L. A. Smith. Scoring Probabilistic Forecasts: The Importance of Being Proper. *Wea. Forecasting*, 22(2):382–388, 2007.

[84] I. T. Jolliffe. The Impenetrable Hedge: A Note on Propriety, Equitability and Consistency. *Meteor. Appl.*, 15(1):25–29, 2008.

[85] H. R. Stanski, L. J. Wilson, and W. R. Burrows. Survey of Common Verification Methods in Meteorology. World Weather Watch Tech. Report No. 8, WMO/TD No.358, WMO, Geneva, 1989.

[86] Z. Toth, O. Talagrand, G. Candille, and Y. Zhu. Probability and Ensemble Forecasts. In I. T. Jolliffe and D. B. Stephenson, editors, *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, pages 137–163. Wiley, Chichester, 2003.

[87] R. E. Livezey. Categorical Events. In I. T. Jolliffe and D. B. Stephenson, editors, *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, pages 77–96. Wiley, Chichester, 2003.

[88] B. Casati, L. J. Wilson, D. B. Stephenson, P. Nurmi, A. Ghelli, M. Pocernich, U. Damrath, E. E. Ebert, B. G. Brown, and S. Mason. Forecast Verification: Current Status and Future Directions. *Meteor. Appl.*, 15(1):3–18, 2008.

[89] A. H. Murphy. A New Vector Partition of the Probability Score. *J. Appl. Meteor.*, 12(4): 595–600, 1973.

[90] S. J. Mason. On Using "Climatology" as a Reference Strategy in the Brier and Ranked Probability Skill Scores. *Mon. Wea. Rev.*, 132(7):1891–1895, 2004.

[91] E. S. Epstein. A Scoring System for Probability Forecasts of Ranked Categories. *J. Appl. Meteor.*, 8(6):985–987, 1969.

[92] W. Drosdowsky and H. Zhang. Verification of Spatial Fields. In I. T. Jolliffe and D. B. Stephenson, editors, *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, pages 121–136. Wiley, Chichester, 2003.

[93] European Centre for Medium-Range Weather Forecasts (ECMWF). Measure of skill - the anomaly correlation coefficient. http://www.ecmwf.int/products/forecasts/guide/Measure_of_skill_the_anomaly_correlation_coefficient.html, 2012.

[94] J. W. Hurrell, Y. Kushnir, G. Ottersen, and M. Visbeck. An Overview of the North Atlantic Oscillation. In *The North Atlantic Oscillation. Climatic Significance and Environmental Impact*, volume 134 of *Geophys. Monogr. Ser.*, pages 1–35. American Geophysical Union, Washington D.C., 2003.

[95] J. Li and J. X. L. Wang. A New North Atlantic Oscillation Index and Its Variability. *Adv. Atmos. Sci.*, 20(5):661–676, 2003.

[96] B. Tinz. Die Nordatlantische Oszillation und ihr Einfluss auf die europäischen Lufttemperaturen. Klimastatusbericht 2002: 32-41, DWD, 2002.

[97] A. Hense and R. Glowienka-Hense. Auswirkungen der Nordatlantischen Oszillation. *promet*, 34(3):89–94, 2008.

[98] D. B. Stephenson. NAO thematic web site. http://www1.secam.ex.ac.uk/cat/NAO, November 2011.

[99] U. Ulbrich, G. C. Leckebusch, H. Paeth, and J. G. Pinto. Veränderungen der NAO im anthropogen beeinflussten Klima. *promet*, 34(3):138–142, 2008.

[100] D. B. Stephenson, H. Wanner, S. Brönnimann, and J. Luterbacher. The History of Scientific Research on the North Atlantic Oscillation. In J. W. Hurrell and et al, editors, *The North Atlantic Oscillation. Climatic Significance and Environmental Impact*, volume 134 of *Geophys. Monogr. Ser.*, pages 37–50. American Geophysical Union, Washington D.C., 2003.

[101] H. Wanner, S. Brönnimann, C. Casty, D. Gyalistras, J. Luterbacher, C. Schmutz, D. B. Stephenson, and E. Xoplaki. North Atlantic Oscillation - Concepts and Studies. *Surveys in Geophys.*, 22(4):321–382, 2001.

[102] D. Chen and C. Hellström. The Influence of the North Atlantic Oscillation on the Regional Temperature Variability in Sweden: Spatial and Temporal Variations. *Tellus*, 51A(4): 505–516, 1999.

[103] H. van Loon and J. C. Rogers. The Seesaw in Winter Temperatures between Greenland and Northern Europe. Part I: General Description. *Mon. Wea. Rev.*, 106(3):296–310, 1978.

[104] T. Spangehl and C. C. Raible. Variationen der NAO auf Basis von langen Zeitreihen, Datenrekonstruktionen und Simulationen der letzten 500 Jahre. *promet*, 34(3):101–107, 2008.

[105] H. Malberg and G. Bökens. Die Winter- und Sommertemperaturen in Berlin seit 1929 und ihr Zusammenhang mit der Nordatlantischen Oszillation (NAO). *Meteorol. Zeitschrift, N.F.*, 6(5):230–234, 1997.

[106] C. Franzke, R. Blender, K. Fraedrich, and F. Lunkeit. Dynamische Antriebsmechanismen der NAO. *promet*, 34(3):108–112, 2008.

[107] G. C. Leckebusch, A. Kapla, H. Mächel, J. G. Pinto, and M. Reyers. Indizes der Nordatlantischen und Arktischen Oszillation. *promet*, 34(3):95–100, 2008.

[108] P. D. Jones, T. J. Osborn, and K. R. Briffa. Pressure-Based Measures of the North Atlantic Oscillation (NAO): A Comparison and an Assessment of Changes in the Strength of the NAO and in its Influence on Surface Climate Parameters. In J. W. Hurrell and et al, editors, *The North Atlantic Oscillation. Climatic Significance and Environmental Impact*, volume 134 of *Geophys. Monogr. Ser.*, pages 51–62. American Geophysical Union, Washington D.C., 2003.

[109] P. D. Jones, T. Jonsson, and D. Wheeler. Extension to the North Atlantic Oscillation Using Early Instrumental Pressure Observations from Gibraltar and South-West Iceland. *Int. J. Climatol.*, 17(13):1433–1450, 1997.

[110] J. W. Hurrell. Influence of Variations in Extratropical Wintertime Teleconnections on Northern Hemisphere Temperature. *Geophys. Res. Lett.*, 23(6):665–668, 1996.

[111] J.W. Hurrell. Decadal Trends in the North Atlantic Oscillation: Regional Temperatures and Precipitation. *Science*, 269(5224):676–679, 1995.

[112] J. W. Hurrell and H. van Loon. Decadal Variations in Climate Associated with the North Atlantic Oscillation. *Clim. Change*, 36(3):301–326, 1997.

[113] D. Pozo-Vázquez, M. J. Esteban-Parra, F. S. Rodrigo, and Y. Castro-Díez. A Study of NAO Variability and its Possible Non-Linear Influences on European Surface Temperature. *Clim. Dyn.*, 17(9):701–715, 2001.

[114] P. D. Jones. The Instrumental Data Record: Its Accuracy and Use in Attempts to Identify the CO2 Signal. In H. von Storch and A. Navarra, editors, *Analysis of Climate Variability. Applications of Statistical Techniques*, pages 53–76. Springer, Berlin, 1999.

[115] World Meteorological Organisation (WMO). Guide to Meteorological Instruments and Methods of Observation. Technical Report WMO-No. 8, World Meteorological Organization, 2008.

[116] D. J. Shea, S. J. Worley, I. R. Stern, and T. J. Hoar. An Introduction to Atmospheric and Oceanographic Data. NCAR Technical Note NCAR/TN-404+IA, National Center for Atmospheric Research, Boulder, Colorado, 1994.

[117] A. Bunde and S. Havlin. Power-law Persistence in the Atmosphere and in the Oceans. *Physica A*, 314(1):15–24, 2002.

[118] D. Lettenmaier. Stochastic Modeling of Precipitation with Applications to Climate Model Downscaling. In H. von Storch and A. Navarra, editors, *Analysis of Climate Variability. Applications of Statistical Techniques*, pages 199–214. Springer, Berlin, 1999.

[119] J. Perelló, M. Gutiérrez-Roig, and J. Masoliver. Scaling Properties and Universality of First-Passage-Time Probabilities in Financial Markets. *Phys. Rev. E*, 84(6):066110, 2011.

[120] J. Masoliver and J. Perelló. First-Passage and Escape Problems in the Feller Process. *Phys. Rev. E*, 86(4):041116, 2012.

[121] R. W. Katz and A. H. Murphy. Assessing the Value of Frost Forecasts to Orchardists: A Dynamic Decision-Making Approach. *J. Appl. Meteor.*, 21(4):518–531, 1982.

[122] T. R. Stewart, R. W. Katz, and A. H. Murphy. Value of Weather Information: A Descriptive Study of the Fruit-Frost Problem. *Bull. Amer. Met. Soc.*, 65(2):126–137, 1984.

[123] W. May, D. J. Shea, and R. A. Madden. The Annual Variation of Surface Temperatures over the World. NCAR Technical Note NCAR/TN-372+STR, National Center for Atmospheric Research, Boulder, Colorado, 1992.

[124] R. O. Weber and P. Talkner. Spectra and Correlations of Climate Data from Days to Decades. *J. Geophys. Res.*, 106(D17):20131–20144, 2001.

[125] K. E. Trenberth, P. D. Jones, P. Ambenje, R. Bojariu, D. Easterling, A. Klein Tank, D. Parker, F. Rahimzadeh, J. A. Renwick, M. Rusticucci, B. Soden, and P. Zhai. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, chapter Appendix 3B: Observations: Surface and Atmospheric Climate Change. Cambridge University Press, Cambridge, 2007.

[126] A. Moberg, P. D. Jones, M. Barriendos, H. Bergström, D. Camuffo, C. Cocheo, T. D. Davies, G. Demarée, J. Martin-Vide, M. Maugeri, R. Rodriguez, and T. Verhoeve. Day-to-day Temperature Variability Trends in 160- to 275-year-long European Instrumental Records. *J. Geophys. Res.*, 105(D18):22849–22868, 2000.

[127] T. R. Karl and C. N. Williams Jr. An Approach to Adjusting Climatological Time Series for Discontinuous Inhomogeneities. *J. Climate Appl. Meteor.*, 26(12):1744–1763, 1987.

[128] P. J. Brockwell and R. A. Davis. *Introduction to Time Series and Forecasting*. Springer texts in Statistics. Springer, New York, 1996.

[129] K. E. Trenberth, P. D. Jones, P. Ambenje, R. Bojariu, D. Easterling, A. Klein Tank, D. Parker, F. Rahimzadeh, J. A. Renwick, M. Rusticucci, B. Soden, and P. Zhai. *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, chapter 3: Observations: Surface and Atmospheric Climate Change. Cambridge University Press, Cambridge, 2007.

[130] T. R. Karl, P. D. Jones, R. W. Knight, G. Kukla, N. Plummer, V. Razuvayev, K. P. Gallo, J. Lindseay, R. J. Charlson, and T. C. Peterson. Asymmetric Trends of Daily Maximum and Minimum Temperature. *Bull. Amer. Met. Soc.*, 74(6):1007–1023, 1993.

[131] S. Lennartz and A. Bunde. Trend Evaluation in Records with Long-Term Memory: Application to Global Warming. *Geophys. Res. Lett.*, 36(16):L16706, 2009.

[132] P. D. Jones. Hemispheric Surface Air Temperature Variations: A Reanalysis and an Update to 1993. *J. Climate*, 7(11):1794–1802, 1994.

[133] D. Rybski, A. Bunde, and H. von Storch. Long-Term Memory in 1000-year Simulated Temperature Records. *J. Geophys. Res.*, 113(D2):02106, 2008.

[134] R. Caballero, S. Jewson, and A. Brix. Long Memory in Surface Air Temperature: Detection, Modeling, and Application to Weather Derivative Valuation. *Clim. Res.*, 21(2): 127–140, 2002.

[135] E. Koscielny-Bunde, A. Bunde, S. Havlin, H. E. Roman, Y. Goldreich, and H.-J. Schellnhuber. Indication of a Universal Persistence Law Governing Atmospheric Variability. *Phys. Rev. Lett.*, 81(3):729–732, 1998.

[136] S. Siegert, J. Bröcker, and H. Kantz. Skill of data based predictions versus dynamical models - case study on extreme temperature anomalies. *to be published in AGU Monograph on Extreme Events edited by M. Ghil, M. Chavez, and J. Urrutia.*

[137] European Centre for Medium-Range Weather Forecasts (ECMWF). Annual Report. http://www.ecmwf.int, 2010.

[138] T. M. Hamill and J. Juras. Measuring Forecast Skill: Is It Real Skill or Is It the Varying Climatology? *Q. J. R. Meteor. Soc.*, 132(621C):2905–2923, 2006.

[139] J. Hansen, M. Sato, R. Ruedy, K. Lo, D.W. Lea, and M. Medina-Elizade. Global Temperature Change. *Proc. Natl. Acad. Sci. USA*, 103(39):14288–14293, 2006.

[140] Climate Research Unit (CRU) of the University of East Anglia. North Atlantic Oscillation data (including Tim Osborn's update). http://www.cru.uea.ac.uk/cru/data/nao/, 2011.

[141] T. M. Hamill, J. S. Whitaker, and S. L. Mullen. Reforecasts: An Important Dataset for Improving Weather Predictions. *Bull. Amer. Met. Soc.*, 87(1):33–46, 2006.

[142] V. C. Slonosky, P. D. Jones, and T. D. Davies. Atmospheric Circulation and Surface Temperature in Europe from the 18th Century to 1995. *Int. J. Climatol.*, 21(1):63–75, 2001.

[143] A. Vecchio and V. Carbone. Amplitude-Frequency Fluctuations of the Seasonal Cycle, Temperature Anomalies, and Long-Range Persistence of Climate Records. *Phys. Rev. E*, 82(6):066101, 2010.

[144] E. S. Epstein. Determining the Optimum Number of Harmonics to Represent Normals Based on Multiyear Data. *J. Climate*, 4(10):1047–1051, 1991.

Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.

Dresden, den 15.11.2012

Die vorliegende Dissertation wurde unter der wissenschaftlichen Betreuung von Prof. Dr. Holger Kantz am Max-Planck-Institut für Physik komplexer Systeme in Dresden angefertigt.