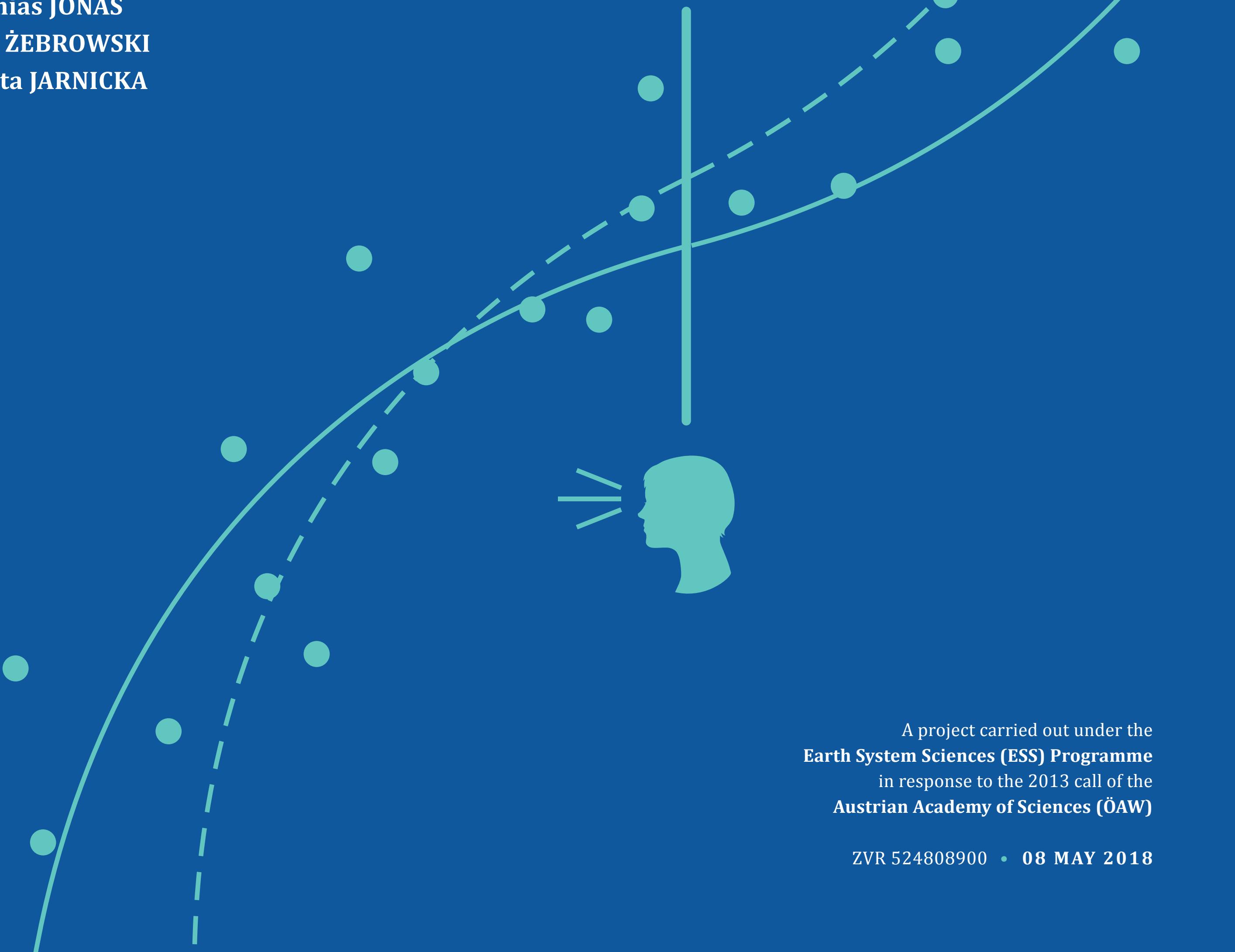
International Institute for
Applied Systems Analysis
IIASA www.iiasa.ac.at

Towards Handling Uncertainty in Prognostic Scenarios: Advanced Learning from the Past

ADVANCED SYSTEMS ANALYSIS (ASA) PROGRAM

PROGRAM DIRECTOR
Elena ROVENSKAYA

AUTHORS
Matthias JONAS
Piotr ŻEBROWSKI
Jolanta JARNICKA



A project carried out under the
Earth System Sciences (ESS) Programme
in response to the 2013 call of the
Austrian Academy of Sciences (ÖAW)

ZVR 524808900 • **08 MAY 2018**

Towards Handling Uncertainty in Prognostic Scenarios:

Advanced Learning from the Past

A project carried out under the

Earth System Sciences (ESS) Programme

in response to the 2013 call of the

Austrian Academy of Sciences (ÖAW)

Online ISBN: 978-3-7001-8290-0
DOI: 10.1553/ESS-PrognosticUncertainty

08 February 2018

International Institute for Applied Systems Analysis (IIASA)
Schlossplatz 1, A-2361 Laxenburg, Austria

Advanced Systems Analysis (ASA) Program
Program Director: Elena ROVENSKAYA
Project Leader: Matthias JONAS
Involved Scientists: Matthias JONAS, Piotr ŻEBROWSKI, Jolanta JARNICKA

Foreword

Here we report on the project

> *Toward Handling Uncertainty in Prognostic Scenarios: Advanced Learning from the Past.*

Per request, our report combines two parts: Part I by Żebrowski, Jonas & Jarnicka (ZJJ hereafter) and Part II by Jonas & Żebrowski (JZ hereafter). The two parts appeared as individual Working Papers of the International Institute for Applied Systems Analysis (IIASA) before (ZJJ: http://pure.iiasa.ac.at/14834/; JZ: http://pure.iiasa.ac.at/14833/). Part I constitutes the main report summarizing the outcome of a one-year project (bearing the same title) under the Earth System Sciences (ESS) Research Program of the Austrian Academy of Sciences (ÖAW); while Part II follows an approach with the focus on systems with memory, typical in Earth system sciences, to determine their explainable outreach. The approach taken by JZ complements the approach taken by ZJJ. Initially, an early draft of Part II was meant to come as attachment or supplementary material to Part I, but it matured in the meantime considerably—which is why it is presented as a self-standing, associated but independent, part to this ÖAW report.

In combining the two IIASA reports, their different formats had not been harmonized. Please note: Working Papers on work of the International Institute for Applied Systems Analysis receive only limited review. Views or opinions expressed herein do not necessarily represent those of the institute, its National Member Organizations, or other organizations supporting the work.

Acknowledgements

The authors are grateful to the Earth System Sciences (ESS) Research Program of the Austrian Academy of Sciences (ÖAW) for financing this research.

About the Authors

Matthias Jonas is a senior research scholar with the IIASA Advanced Systems Analysis (ASA) Program. His interests are in environmental science, and in the development of systems analytical models and tools to address issues of global, universal and regional change, including surprises, and their potential implications for decision and policymakers.

E-mail: jonas@iiasa.ac.at

Piotr Żebrowski joined the IIASA Advanced Systems Analysis (ASA) Program as a research assistant in February 2015. His current research focus is on diagnostic uncertainty of greenhouse gas inventories, uncertainty propagation in climate models and on retrospective learning.

E-mail: zebrowsk@iiasa.ac.at

Jolanta Jarnicka is a researcher in the Systems Research Institute of the Polish Academy of Sciences. Her specialty is probability and statistics, in particular nonparametric statistical methods, data analysis, and mathematical modeling.

E-mail: jarnicka@gmail.com;

PART I. Towards Handling Uncertainty in Prognostic Scenarios: Advanced Learning from the Past

Piotr Żebrowski[1], Matthias Jonas[1], Jolanta Jarnicka[2]

[1] IIASA, Advanced Systems Analysis Program

[2] Systems Research Institute of the Polish Academy of Sciences

Abstract

In this report we introduce the paradigm of learning from the past which is realized in a controlled prognostic context. It is a data-driven exploratory approach to assessing the limits to credibility of any expectations about the system's future behavior which are based on a time series of a historical observations of the analyzed system. This horizon of the credible expectations is derived as the length of explainable outreach of the data, that is, the spatio-temporal extent which, in lieu of the knowledge contained in the historical observations, we are justified in believing contains the system's future observations. Explainable outreach is of practical interest to stakeholders since it allows them to assess the credibility of scenarios produced by models of the analyzed system. It also indicates the scale of measures required to overcome the system's inertia. In this report we propose a method of learning in a controlled prognostic context which is based on a polynomial regression technique. A polynomial regression model is used to understand the system's dynamics, revealed by the sample of historical observations, while the explainable outreach is constructed around the extrapolated regression function. The proposed learning method was tested on various sets of synthetic data in order to identify its strengths and weaknesses, and formulate guidelines for its practical application. We also demonstrate how it can be used in context of earth system sciences by using it to derive the explainable outreach of historical anthropogenic CO_2 emissions and atmospheric CO_2 concentrations. We conclude that the most robust method of building the explainable outreach is based on linear regression. However, the explainable outreach of the analyzed datasets (representing credible expectations based on extrapolation of the linear trend) is rather short.

Keywords: Learning, explainable outreach, uncertainty, greenhouse gas emissions

Contents

1. Introduction

1.1. Scientific context of the project

The problem of uncertainty and horizons of credibility[1] of predictions of future behavior of Earth's climate system has attracted a growing interest as a consequence of the increasing demand for incorporating information about future climate into planning and decision making (e.g., IPCC 2007: FAQ 1.2, FAQ 8.1; NSF 2012; IPCC 2013: Box 11.1; Otto et al. 2015). Numerous scientific institutions, including IIASA, use a variety of complex integrated assessment models to generate a great number of prognostic scenarios in order to identify policy options and effectiveness of different measures for mitigating climate change. Modelers make huge efforts to ensure the credibility of their scenarios and gauge their uncertainty; for example, by carrying out sensitivity tests or inter-model comparisons under standardized conditions. In particular, multi-model-scenario exercises are becoming increasingly popular (e.g., Meinshausen *et al.* 2009). Nevertheless, such efforts are not entirely convincing, and judging the credibility of climate model projections remains a notorious and unresolved issue (cf. Otto et al. 2015).

In contrast to these model-related issues we propose an alternative, data-driven perspective looking at the limits to how our current understanding of the Earth system can be used to predict its future behavior. We seek to assess these limits by answering the following questions:

(1) *Given the data reflecting a system and their diagnostic uncertainty can we deduce the* ***explainable outreach***[2] *of these data, which expresses our understanding of the prevailing patterns of the system's behavior and their typical duration?*

and

(2) *Can the explainable outreach be used for assessing the limits of credibility of predictions?*

In order to answer these questions, we develop and apply a new (to our knowledge) exploratory method, which we call **learning in a controlled prognostic context**[3] (or prognostic learning **(PL)** for simplicity). Its main idea is **to learn about the nature of the analyzed system from its past**: we use a part of the historical observations of the system to understand its basic dynamic and formulate our expectations about its future evolution (expressed as the explainable outreach) and then test these expectations against the remaining part of the sample. This way of testing the limits of our understanding of the system based on partial and uncertain knowledge (carried by a finite set of (possibly imprecise[4]) observations) may inform us about the likely time horizon within which our

[1] Credibility of predictions is understood as our expectations (predictions) of its performance (Otto et al. 2015)

[2] The region – both in terms of time horizon and the range of plausible future values – within which we may have justifiable belief based on the past system's behaviour, that it will contain future trajectory of the process' evolution.

[3] Description of the method together with explanation of its name is provided in Chapter 2.

[4] We assume that the data are accurate (i.e., no systematic bias of the system's observations).

expectations about its future evolution may be considered plausible, in lieu of the available historical data. Therefore, the proposed method belongs to the realm of **data analysis, NOT modeling.** The difference between learning in a controlled prognostic context and modeling is explained by Figure 1.)

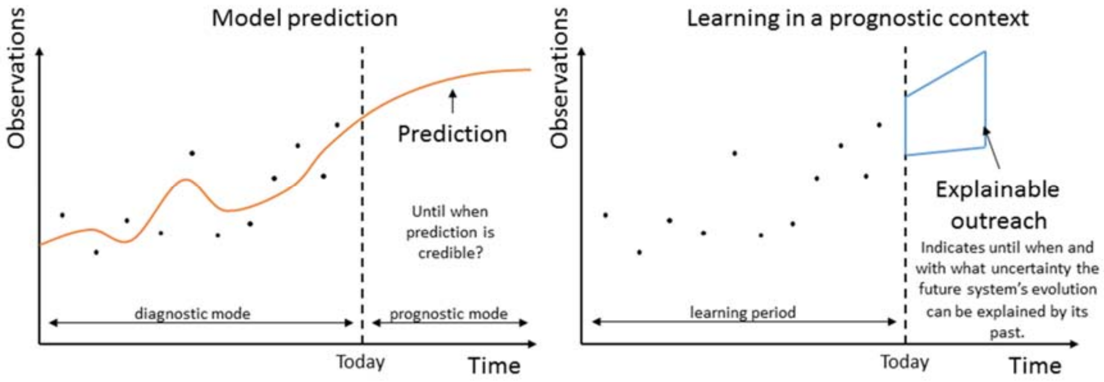


Figure 1. Model prediction vs. learning in a prognostic context. Left panel: Model prediction. A model is calibrated against historical data (diagnostic mode) before making a prediction, for example by extrapolating the historical trend into the future or generating a scenario pathway (prognostic mode). Modelers typically do not (or cannot) indicate until when a model prediction is in accordance with the systems past (i.e. is credible). Right panel: Learning in a prognostic context. Given the historical data the system's dynamics can be understood and the data's explainable outreach be constructed. The explainable outreach specifies both spatial and temporal extent beyond which we no longer can explain our system in accordance with its past. The purpose of deriving explainable outreach directly from the data is to indicate limits of predictability of the model which we built to reflect the underlying system.

1.2. Motivation: problems with judging the credibility of predictions

Credibility of predictions is one of the central problems of statistical modeling. A variety of well-established statistical methods—such as regression models and machine learning techniques (Hastie *et al*. 2009, Murphy 2012) or time series analysis techniques (Brockwell & Davis 2002)—aim at predicting responses of the analyzed system in as yet unobserved states[5]. Predictions are typically expressed in terms of a regression function or, more generally, as conditional expected value of the system's response given the value of the explanatory state variable. Quality of predictions, usually understood as expected prediction error, can be controlled[6], provided that the state in which we wish to make a prediction lies **within** the range of the data sample on which the analysis is based. However, analogous error control is formally unavailable for predictions of the system's responses in states lying **beyond** the range of the data sample (i.e., in conditions which may be significantly different to those of the historical observations).

Similar problems also haunt the modeling community. Their common and apparently unavoidable practice is to extrapolate the current understanding of the system (e.g., discovered trends or relationships) **beyond** the range of historical data sample in order to predict its future behavior, possibly in yet unobserved states. For example, this approach

[5] That is, in conditions not covered by the available data (out-of-sample predictions).

[6] The upper bands for probability of large prediction errors are available and depend on the complexity of the statistical model and the length of the data sample.

was employed in a study by Meinshausen *et al*. 2009 aiming to predict the level of global warming in the future, when greenhouse gas concentrations in the atmosphere will be at higher levels than any time in (recent) history. However, making such predictions by extrapolating the observed trends beyond the range of the sample is problematic. Unless one assumes that the observed process is in some sense stationary (which may be too strong an assumption, e.g., in presence of varying exogenous forcing) one loses control over the quality of predictions, whose errors may rapidly increase the further away from the sample of historical observations one moves. Typically, modelers try to assess credibility of predictions by either (1) providing uncertainty ranges for the predictions[7]; (2) using sensitivity analyses[8]; or (3) exploring the range of possible futures using selected scenario pathways (in particular in the case of computationally expensive models). Unfortunately, these methods are not entirely convincing due to a certain degree of arbitrariness in their application (e.g., assumed distributions of parameters underlying Monte Carlo methods or the choice of storylines for scenario pathways). More importantly, they **do no indicate the time horizon within which model predictions remain in accordance with the system's past[9].**

The paradigm of learning in a controlled prognostic context offers at least a partial solution to these problems. It is a data analysis method designed to control the growing uncertainty of our expectations about the system's evolution in the immediate future. Moreover, this approach may provide a model-independent indicator of the time range within which the projections of a model may be judged credible in lieu of past system behavior.

1.3. Objectives and scope of the report

The objectives of this report are the following: (1) to introduce the generic paradigm of the **learning in a controlled prognostic context** allowing us to assess the **explainable outreach**, that is, the region—specified in terms of time horizon and the range of plausible future values (uncertainty)—which we can be justified in believing (based on historical observations) will contain future trajectory of the system's evolution; (2) to propose a way (based on regression techniques) of implementing the prognostic learning (PL) paradigm; and (3) to demonstrate its usefulness in analysis of the real data samples relevant to understanding the Earth's climate system (e.g., anthropogenic CO_2 emissions and atmospheric CO_2 concentrations).

The paradigm of learning in the controlled prognostic context is applicable to: (1) univariate regression—problems in which one is interested in the form of dependence of one quantity characterizing a system (the response variable) on another quantity (the independent variable) which represents the state of the system or its forcing; and (2) analysis of time series—in which case time is treated as the independent variable.

[7] Assuming suitable probability distributions for values of exogenous parameters of the model they may be derived analytically or by means of Monte Carlo simulations.

[8] In this case possible correlations between exogenous parameters of the model are typically ignored. Changes in model responses are usually analysed by varying values of one of the parameters while keeping the rest constant.

[9] By "remaining in accordance with the system's past" we mean that predicted future trajectory of the system's evolution exhibits behavior similar to this observed in the past, such as the level of "system's inertia" or the type of dynamics. Note that this is weaker notion than stationarity of the process.

In this report we restrict ourselves to analysis of time series data only. The reason for that is two-fold. Firstly, in the context of time series "predicting beyond the range of sample" means "forecasting or predicting the future" which facilitates understanding of the idea of explainable outreach. Secondly, a time series perspective is relevant both in the context of prognostic modeling and in the context of understanding the relevant Earth systems processes (such as the abovementioned CO_2 emissions or CO_2 concentrations). Hence, from now on (unless stated otherwise), all considered data samples will be assumed to consist of pairs (t, x_t), where x_t denotes the value of the observable describing the system of interest which was recorded at time t. We will call this observable a system's state variable[10].

1.4. Structure of the report

In Chapter 2 we introduce the concept of learning in a controlled prognostic context. There we give a definition of the explainable outreach of the data, which is a central notion of the proposed methodology. Next, we formulate a generic procedure for learning in a controlled prognostic context and discuss how it should be applied and how to interpret its results. We conclude Chapter 2 by comparing the proposed approach to standard time series analysis.

In Chapter 3 we propose a way of implementing the generic procedure of learning in a controlled prognostic context. Namely, we show how it can be operationalized by using polynomial regression. We discuss how to define the shape of the explainable outreach and how to determine its length. We summarize Chapter 3 with a formulation of the regression-based procedure of prognostic learning.

The next two chapters are devoted to analysis of the performance of the proposed method. In Chapter 4 we present insights from the experiments on various synthetic datasets. The purpose of these experiments is to identify the strengths and weaknesses of the proposed method and to formulate guidelines for its application in real-life data analysis. In Chapter 5 we test these insights by applying the method to determine the explainable outreach of the time series representing anthropogenic CO_2 emissions and atmospheric CO_2 concentrations.

We conclude this report with a summary and outlook for future research followed by an appendix in which we present yet another way of implementing the prognostic learning method—this time based on non-parametric regression techniques. We also demonstrate the potential of this variant of prognostic learning method by applying it to the abovementioned real-life time series.

2. Learning in a controlled prognostic context

In this chapter we present the notion of learning in a controlled prognostic context (prognostic learning, PL). Broadly speaking, the purpose of this method is to indicate both the typical length of time intervals over which the trends observed in the historical

[10] or simply state variable

data sample persist, and the level of uncertainty in estimating and extrapolating these trends.

PL can be classified as a method of exploratory data analysis. Its aim is not to find a formal statistical model which can be used for testing a hypothesis about the historical data sample and making predictions for the future. Instead, the PL method offers a semi-formal first-order description of the system's dynamics and its "inertia"[11] exhibited by the system over the period in which the data sample was collected. This "inertia" is a critical factor in determining the limits to credibility of predictions about the system's behavior[12].

As such, the PL method informs us solely about the system's behavior in the past. However, in this report we demonstrate that it is also useful in context of expressing expectations about its immediate future. The rationale for this approach is provided by the observation that patterns in the system's behavior in the relatively recent past are also likely to occur in the nearby future. Therefore, the findings of the PL method, which, in essence, concerns only the past of the system, can also be informative about its near future. Note that the requirement for this line of thinking to be valid is just that the nature of the system itself or its external forcing do not change too rapidly over time. This is considerably weaker requirement than stationarity of the system usually assumed by the formal statistical modeling methods[13].

It is also important to note that the PL method is data-driven (i.e., is based only on the sample of historical observations) which implies also that it adopts a conservative view of the system. Namely, it cannot anticipate systemic surprises and behaviors which had not occurred in the period over which the sample of historical observations was collected.

2.1. Generic notion of the explainable outreach of the data

The core idea of the PL approach is to **deduce directly from the data** their **explainable outreach (EO)**, that is, the spatial and temporal extent beyond which using knowledge about its past can no longer explain the system's behavior. The EO is characterized by four key attributes: (i) the time it begins; (ii) the diagnostic uncertainty of the state variable describing the system in this initial moment (defining the initial opening of EO); (iii) the increase of prognostic uncertainty in time; and (iv) the temporal extent (quantifying the time in the future beyond which the system's behavior can no longer be shown to be in accordance with its past behavior).

Explainable outreach can be seen as a region in cartesian product space of time and the domain to which the values of the observations belong (e.g., real numbers). The shape of

[11] Understood as a system's memory—a typical period within which the system does not undergo a significant change of its dynamics (e.g., average time horizon within which system exhibits linear dynamics with constant slope).

[12] For example, if a system has undergone sudden and unexpected changes of its dynamics in the past it has a low "inertia". In this case any long term prediction of the future system's behaviour is not very credible.

[13] Some sort of stationarity is required by statistical models applied for making predictions of the future system's behaviour. That way they avoid the question of the credibility of such predictions—their uncertainty may be growing in time but, due to stationarity, the dynamics of the system does not change in any limited time horizon. In contrast, the PL method aims to identify the time horizon within which the system's behaviour is sufficiently well described—thus assumptions are significantly weaker. Cf. Table 2 for further discussion.

this region is determined by our understanding of the system (for example, the form of trend function used to describe system's dynamics). Its spatial boundaries are given by uncertainties related to the projection of our understanding of the system into the future (e.g., prediction bands[14] centered on an extrapolated trend), while its temporal extent is characterized by the time this projection starts and the time horizon within which the uncertainty region covers the trajectory of the system.

Obviously, different hypotheses about the type of trend the system follows will result in different EOs. Some of them may be very long and wide (if the system's behavior is described robustly but very imprecisely) or short and narrow (if our understanding of the system is quite precise but only locally correct). A long and narrow EO is most preferable.

Comparison of different EOs derived for the same sample may be facilitated by a score assigning a numeric value to the combination of EO attributes (i) – (iv). For example, one could use the following

$$\text{Score of EO} = \frac{\text{Length of temporal extent of EO}}{\text{Width of EO at its end}}$$

This score increases as the length of EO increases or its width decreases. An EO with a higher score is preferable.

2.2. Prognostic learning procedure

Note that an EO as defined above expresses our expectations about the consequent system's behavior from a certain fixed moment in time. Because of data variability and possible imprecision in our understanding of the system, an EO starting at another time may have a different shape and length. Therefore, to gain some understanding of a system's behavior it is insufficient to look at just one EO. One should rather derive this understanding from a sequence of consecutive EOs resulting from a learning procedure.

Below we provide a generic procedure of learning in a controlled prognostic context given the learning sample $\mathrm{X}_0, \dots, \mathrm{X}_\mathrm{T}$ of observations of the analyzed system collected over the period $[0, \mathrm{T}]$:

1. Choose a suitable set of hypotheses (e.g., a family of regression functions) about the rules governing system behavior and the minimal number k of data points required to select the one which represents the system best.
2. Choose the initial length $\tau = k$ of the subsample $\mathrm{X}_0, \dots, \mathrm{X}_\tau$, which we call the learning block (LB).
3. Choose the hypothesis which reflects the system's behavior best in the LB $\mathrm{X}_0, \dots, \mathrm{X}_\tau$ (e.g., estimate parameters of the regression function) and quantify its uncertainty (e.g., with use of prediction bands).
4. Find the EO starting point τ. To determine the shape of the EO calculate the uncertainty region $R \subset [\tau, \infty) \times \mathbb{R}$ spanned by the prediction of future system

[14] For each moment in time, prediction bands give the range which is expected to contain, with predefined probability (called confidence level), an observation taken at that time. In contrast, confidence bands give a range which we expect to cover the true value of an observation. In this report we prefer to use prediction bands since we want to test our understanding of the system with individual data points.

behavior based on the hypothesis chosen in in point 3 and its uncertainty. To determine the length of the EO project the remainder of the data $\mathrm{X}_{\tau+1}, \dots, \mathrm{X}_{\mathrm{T}}$, which we call the testing block (TB), onto region R and find the largest H such that[15]

$$\forall\, \tau < t \leq \tau + H \ (t, X_t) \in R$$

If $H < T - \tau$ then the length of the EO starting point τ is set to H; otherwise it is set to ∞.

5. If $\tau < T$ then set $\tau = \tau + 1$ and go to step 3; otherwise end the procedure.

The above procedure explains the meaning of the name "learning in a controlled prognostic context": we learn about the patterns of the past system behavior (step 3) and then test this knowledge by applying it in a prognostic mode in the controlled context of the remainder of the data sample (step 4).

Assessment of the temporal extent of the EO, H, from step 4 of the learning procedure requires a discussion. It is either finite (no longer than the historical sample itself) or set to infinity. In the first case, the finite time horizon of the EO indicates limits within which we can predict a system's evolution sufficiently well after time τ by means of the method selected in step 1 to understand the system's dynamics in the LB. In other words, it indicates the limits to credibility of predictions of the system's behavior after time τ, based on our understanding of the system's dynamics given the knowledge carried by the LB $\mathrm{X}_0, \dots, \mathrm{X}_\tau$. On the other hand, an infinite time horizon indicates that we are unable to falsify this understanding of the system's behavior with the TB $\mathrm{X}_{\tau+1}, \dots, \mathrm{X}_{\mathrm{T}}$ (i.e., we have no grounds to reject our hypothesis about the system's nature). There are two possible reasons for such a situation: either our understanding of the system is exceptionally good or the TB is too short to provide evidence against it[16]. This indicates an important constraint of the PL approach (indeed, of any data-driven method), namely that data resources (the length of sample of historical observations) set limits to the level of detail[17] with which we wish to describe analyzed system.

2.3. Applying the prognostic learning procedure and interpretation of its results

Learning in a controlled prognostic context is essentially a model-independent paradigm of exploratory data analysis. By this we mean that it does not presuppose any particular model which reflects our *a priori* knowledge[18] or belief about the analyzed system, and which may be calibrated on the sample of historical observations and then used for making predictions. On the contrary, the PL approach is purely data-driven: we explore a sufficiently broad family of alternative methods of describing the system's behavior

[15] If the hypothesis about the system's behaviour is formulated in terms of a regression model, the requirement that all points between time τ and $\tau + H$ belong to R may be relaxed—only a sufficient portion of these points need fall into R.

[16] Falsifying a good hypothesis may require a very long testing sample. In the extreme (but very unlikely) case, when we perfectly understand our system (i.e., know the process generating data—both in the past and in future) we wouldn't be able to falsify it with use of any test sample of finite length.

[17] Understood as the complexity of the hypothesis about the system's dynamics.

[18] Additional knowledge (e.g., about a particular type of dynamics the system follows) obtained beforehand from some other source than the learning sample $\mathrm{X}_0, \dots, \mathrm{X}_{\mathrm{T}}$.

(e.g., different types of regressions) by running a PL procedure (cf. section 2.2) for each of them and then selecting the one which yields the best EOs.

After completing this task, we obtain a sequence of EOs indexed by their starting points $\tau = k, k+1, \ldots, T$. Technically, this will tell us how credible our predictions based on partial knowledge about the system[19] were over the time interval $[0, T]$. In particular, it provides no confirmed (tested) information about the EO starting at time T, which expresses our expectations about the immediate future of the system. This cannot be done formally without additional and restrictive assumptions (e.g., stationarity of the system), however, such an exercise still may be informative. If the behavior of the EOs over the period $[0, T]$ was regular enough (i.e., EOs have comparable scores, implying similar lengths and widths) and the last τ for which EO has finite length is sufficiently close to T we may attempt to extrapolate the characteristics of (tested) EOs to formulate expectations about likely shape and temporal extent of the (untested) EO starting at time T.

In principle, the results of the PL method give us insight into system's "inertia". Such information may be useful for decision makers trying to influence future behavior of the system (e.g., mitigate global warming by implementing certain policies). First, it indicates likely directions of future system evolution under "business as usual" conditions[20] which is useful reference point for policy making. Second, it indicates the time horizon within which we may have some confidence in quality of predictions based on our understanding of the system. Third, it indicates the strength of the measures needed to overcome the system's inertia and to shift its future evolution towards the desirable path[21].

PL methodology may also be applied to assess scenarios produced by a particular model of the system of interest. If a scenario falls out of the EO before its end, it means that the model predicts a change in the system's dynamics (with respect to its past behavior). If so, then modeler should explain the reason for that, for example, what significant changes the system is expected to undergo under that scenario. If the future trajectory under the "business as usual" scenario falls outside the EO it may indicate an inadequacy of the model to describe the system of interest.

[19] That is, knowledge carried by learning blocks $\mathrm{X}_0, \ldots, \mathrm{X}\tau, \tau < T$.

[20] That is, in a situation where the current dynamics of the process and external forcing / policies / measures will not change.

[21] If the system's trajectory under a scenario corresponding to introduction of a certain policy stays within the EO it indicates that the effectiveness of such a policy remains uncertain within the time horizon of this EO.

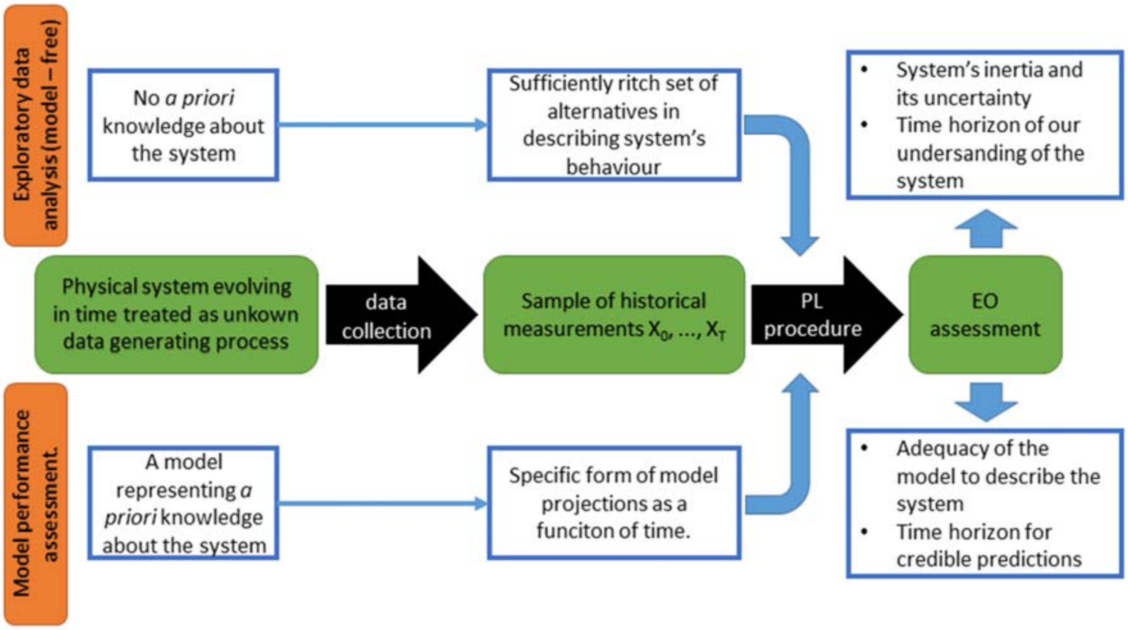

Exploratory data analysis (model – free)
No *a priori* knowledge about the system
Sufficiently ritch set of alternatives in describing system's behaviour
System's inertia and its uncertainty
Time horizon of our undersanding of the system
Physical system evolving in time treated as unkown data generating process
data collection
Sample of historical measurements $X_0, \ldots, X_T$
PL procedure
EO assessment
Model performance assessment.
A model representing *a priori* knowledge about the system
Specific form of model projections as a funciton of time.
Adequacy of the model to describe the system
Time horizon for credible predictions

Figure 2. Two modes of applying the learning in a controlled prognostic context paradigm. In exploratory data analysis mode the selection of the best method to represent system's behavior and construct EO is purely data driven without use of any *a priori* knowledge. EO indicates the inertia of the system and the uncertainty and time horizon of our understanding of the system. In model assessment mode a model-specific form of a trend function is fed into the PL procedure in order to assess model's ability to accurately describe the system and to quantify limits to its predictions (this mode is not considered in this report).

We also speculate that a modification of the PL method may be applied to assess a particular model and its projections even more directly. If it is possible to express the model prediction as a function of time (of a certain form, and dependent on initial conditions and values of exogenous parameters) and calculate a region spanned by the projection and its uncertainty, one can use this function directly in the prognostic learning procedure (see section 2.2). Resulting EOs could then indicate the time horizon within which the model is sufficiently adequate to describe the system's evolution. However, this generic approach would require a model-specific implementation of the PL procedure to be designed. This modification of PL approach has not yet been tested and will not be covered in this report.

2.4. Prognostic learning versus forecasting with use of time series analysis

The PL approach discussed in this report examines time series data. It is, however, quite different from the commonly used time series analysis (TSA) methodology. PL trades only approximate understanding of the behavior of the data itself for ability to indicate the limits to this understanding and generality of the method. TSA, on the other hand, strives for complete understanding of the data generating process and applies this knowledge for making predictions. This approach however does not allow for specifying the limits for predictions[22].

Typically, TSA is based on decomposition of the time series into a deterministic component (functional trend, seasonal component, oscillations) and a stochastic part. The deterministic part can be estimated from the data with use of a broad range of various techniques (such as regressions, curve fitting, smoothing methods, wavelet analysis, etc.) The overarching goal is to estimate the deterministic part so that it fits the data as closely as possible; its extrapolation properties are a lower priority concern. The nature of the stochastic part is inferred from the behavior of residuals (i.e., the part remaining after removing the estimated deterministic component from the data). This is usually done by fitting a suitable time series model (such as ARIMA or GARCH).

Obviously, the estimate of the deterministic component of the time series significantly influences the behavior of residuals and thus the statistical model of the stochastic part. As the latter may be quite complex and difficult to estimate (e.g., due to scarcity of the data resources with respect to the number of parameters in the model), the problem of estimation of the deterministic component is somewhat subordinate to the analysis of residuals. The estimate of the deterministic part is expected to produce residuals for which the statistical model is as simple as possible. The literature of the subject puts much more emphasis on the statistical models of the residuals, typically assuming that the deterministic component of analyzed time series has already been removed with use of some suitable technique (e.g., Brockwell & Davis 2002).

Once the time series is described in terms of deterministic function of time and statistical model of residuals one may use this knowledge for making forecasts. In order to do so, the deterministic trend is extrapolated and the behavior of the stochastic part (i.e., the residuals) is either determined theoretically (e.g., prediction bands obtained under stationarity assumptions) or simulated (using the statistical model of residuals). However, such forecasts should be considered with caution. Technical problems may arise due to an incorrect structure of the model of stochastic part and/or bad extrapolation properties of the function describing the deterministic component (such as instability due to uncertainty in estimated values of function parameters). Some techniques of describing the deterministic part, such as smoothing splines, even rule out the possibility of extrapolation. Moreover, when making forecasts the description of the analyzed time series (i.e. the deterministic function plus statistical model of residuals) are treated as the true process generating data which will never change. As a result, indicator of a time horizon within which the predictions are credible cannot be derived from TSA methodology.

We conclude this section with Table 1 summarizing differences between PL method and TSA.

[22] In fact, TSA approach does not even recognise it as a problem. If our understanding of the system is complete then we are able to predict its behaviour in any time horizon.

Table 1. Prognostic learning versus time series analysis.

	Learning in a controlled prognostic context	**Time series analysis**	
		Deterministic component	**Stochastic component**
Approach	Data-driven exploratory analysis. Emphasis on striking a balance between approximate understanding of the system and ability to indicate the limits to this understanding.	Inferring the data-generating process. Emphasis on the statistical model of the stochastic component, while the estimate of the deterministic component is to yield desired statistical properties of the residuals.	
Assumptions	No systemic surprises (behaviors unobserved in the past will not happen in the future).	Particular form of trend function.	Particular form of the dependence structure / model of residuals. Usually also normality and weak stationarity of residuals is required.
Principle	**Optimization of the EO**. Selecting the type of trend generating the longest and narrowest EO.	Fitting a function minimizing **in-sample error**.	**Estimation from the data values of the model parameters that minimize expected forecast error.**
Measure of performance	Score of the explainable outreach	Typically sum of squared errors or mean squared error	Typically expected mean squared error
Predictions	Data-driven model describes the system only approximately correctly and uncertainty of predictions inevitably grows in time. **The method does not strive for perfect predictions. It aims to understand their limits.**	Within the range of the observed sample the fitted function is interpreted as expected value of observations. Extrapolation of the fitted function beyond the range of sample may be interpreted in the same way but there is no possibility for controlling the error of predictions with use of such extrapolation.	Future behavior of the stochastic component (typically expressed in form of prediction or confidence bands) is derived from the statistical model of residuals either theoretically (usually under assumption of stationarity) or by means of simulations utilizing model structure.
Time horizon within which forecasts are supposed to be reliable	Expected length of the EO based on the assessment of the results of the prognostic learning procedure.	Unknown. Fitted model of the time series (i.e., estimated deterministic component and statistical model of the stochastic part) is treated as the true data generating process and as such universally correct.	
Sources of uncertainty	(1) Diagnostic uncertainty (measurements errors) reflected by initial opening of the EO; and (2) prognostic uncertainty which grows into the future reflected by the shape of the EO	(1) Uncertainty in the form of the function describing deterministic component; and (2) uncertainty in the parameter estimates.	(1) Uncertainty in estimate of deterministic component defining residuals; (2) uncertainty of structure of model of residuals; and (3) uncertainty of estimates of model parameters.

3. Regression – based construction of the explainable outreach

In this chapter we propose a practical method of implementing the generic paradigm of learning in a controlled prognostic context presented in Chapter 2. Making this generic notion operational requires us to address the following problems:

1. Understanding the behavior of the data from the LB and quantifying the diagnostic uncertainty in order to specify the direction and initial width of the EO.
2. Defining the shape of the EO (i.e., its spatial boundaries).
3. Determining the length of the EO by testing it against the data from the TB.

Below, we propose a solution to these questions which is based on the regression techniques.

Ad 1. The trend in the data is identified by means of a regression function fitted to the points from the LB. For each moment t belonging to the LB, the value of regression function at that moment is interpreted as the expected value of the observation taken at time t. The extrapolation of the regression function defines the main axis around which the EO is constructed. The diagnostic uncertainty is expressed as the standard deviation of residuals (i.e., differences between the regression function and the actual observations) and defines the initial width of the EO.

Ad 2. The shape of the EO (i.e., its upper and lower band) is given by extrapolation of the prediction bands calculated for the regression model fitted to the LB.

Ad 3. Given the shape of the EO, its length is determined by projecting the remainder of the learning sample (i.e., the TB) onto it. The moment the EO ends is defined as the earliest moment for which the position of the testing points with respect to the EO starts to be very unlikely if the regression model fitted over the LB is correct and true also beyond its range.

The details of the proposed solution depend on the specific regression technique to be applied. In the remainder of this section we give these details for the PL procedure based on polynomial regression. In Appendix B we present an alternative PL procedure based on a local linear regression method.

3.1. Analysis of historical patterns in learning phase with use of polynomial regression

Polynomial regression is a widely used parametric technique of data analysis. Its popularity comes from the fact that it is a relatively simple and straightforward generalization of the classic linear regression method, as well as from the flexibility of the family of polynomial regression functions[23]. It is also a popular technique for estimating the deterministic part of a time series (Brockwell & Davies 2002).

In order to approximate the historical trend in the LB we use a model of polynomial regression of order p

[23] Indeed, any continuous trend in the data can be locally approximated with arbitrary precision by a polynomial of sufficiently high order.

$$x(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \cdots + \alpha_p t^p + \varepsilon_t$$

where $x(t) = X_t$ is a value of the observation taken at time t and the noise term ε_t is normally distributed with zero mean and standard deviation σ. Moreover, we assume that ε_t, $t = 0, 1, 2, \ldots$, are independent and identically distributed.

Let the LB contain n observations taken in times $t_1, \ldots, t_n$. We estimate the parameters of the regression function

$$\hat{x}(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_p t^p$$

with use of the ordinary least squares (OLS) method (Wolberg 2006: chapt. 2). The uncertainty of the fitted regression function at time t is then given by formula

$$s_x(t) = \sqrt{\frac{\sum_{i=1}^{n}(\hat{x}(t_i) - x(t_i))^2}{n-(p+1)} \sum_{j=1}^{p+1}\sum_{k=1}^{p+1} t^{j+k-2}[C^{-1}]_{j,k}}$$

where $[C^{-1}]_{j,k}$ is the entry at the cross-section of the *j*-th row and *k*-th column in the inverse of matrix

$$C = \left[\sum_{i=1}^{n} t_i{}^{j+k-2}\right]_{\substack{j=1,\ldots,p+1 \\ k=1,\ldots,p+1}}$$

The diagnostic uncertainty over the LB is assumed to be constant and is estimated as a standard deviation of the model residuals

$$s_r = \sqrt{\frac{\sum_{i=1}^{n}(\hat{x}(t_i) - x(t_i))^2}{n-(p+1)}}$$

Upper and lower prediction bands at the confidence level $(1-\alpha)$ for the observations taken at time *t* are then given by the formulas

$$f_{up}(t) = \hat{x}(t) + \mathrm{t}_{n-(p+1),1-\alpha/2}\sqrt{s_x(t)^2 + s_r^2}$$

and

$$f_{low}(t) = \hat{x}(t) - \mathrm{t}_{n-(p+1),1-\alpha/2}\sqrt{s_x(t)^2 + s_r^2}$$

respectively, where $\mathrm{t}_{n-(p+1),1-\alpha/2}$ is $(1-\alpha/2)$ quantile of the t-Student distribution with $n - (p+1)$ degrees of freedom. Note that parameter α regulates the width of the prediction bands (the lower the α the wider the prediction bands). Observe also that distance between prediction bands, that is, $f_{up}(t) - f_{low}(t)$, increase with p-th power of t.

3.2. Construction of the explainable outreach

The EO starts at time $\tau = t_n$, that is, the moment in which the last observation of the LB was taken. The EO is built around the extrapolated polynomial trend that was fitted to the data in the LB, that is around $\hat{x}(t), t \geq \tau$. Its initial width is defined as $f_{up}(\tau) - f_{low}(\tau)$ and is determined by the diagnostic uncertainty s_r. The shape of the EO (i.e., its upper and lower band) are given by functions $f_{up}(t)$ and $f_{low}(t)$ for $t > \tau$, that is, the prediction bands for the regression model extrapolated beyond the LB.

Note that in order to define the initial width and the shape of the EO, only the information about the system's behavior in the LB is needed. However, to determine its temporal extent (time horizon) additional knowledge carried by the remainder of the learning sample (TB) is required. This remaining subsample is used to determine until when our expectations about the future system's evolution after time τ represented by the EO (based only on the knowledge contained by the LB) are in accordance with the actual evolution of the system after that time.

To explain how we determine the moment at which the EO ceases to be in accordance with the actual system's evolution, let us assume that we know the evolution of the analyzed process only up to the moment τ and the m remaining points in the TB (t_1, X_1), …, (t_m, X_m), $t_1 = \tau$, $t_m = T$, are unknown. In addition, let us define an auxiliary sequence of random variables

$$E_k = \begin{cases} 0 \text{ if } X_k \notin [f_{low}(t_k), f_{up}(t_k)] \\ 1 \text{ if } X_k \in [f_{low}(t_k), f_{up}(t_k)] \end{cases}$$

where (t_1, X_1), …, (t_m, X_m) are the yet unknown points from the TB.

Now observe that if the regression model fitted to the LB correctly describes the evolution of the analyzed process then the points from the TB should also follow this model. If that is so, then by definition of the prediction bands at the confidence level $(1-\alpha)$ the probability that the future observation taken at time $t \geq \tau$ will fall into interval $[f_{low}(t), f_{up}(t)]$ is equal to $(1-\alpha)$. Thus $E_k = 1$ with probability $(1-\alpha)$ and $E_k = 0$ with probability α. In other words, all $E_k, k = 1, \dots, m$ follows the Bernoulli distribution with parameter $(1-\alpha)$[24]. Moreover, if the regression model fitted to the LB is also correct for the observations in TB, then these observations are independent. Therefore, all $E_k, k = 1, \dots, m$ are not only identically distributed but also mutually independent. As a consequence, for each $k = 1, \dots, m$, a random variable

$$S_k = \sum_{i=1}^{k} E_i$$

has a binomial distribution $B(k, (1-\alpha))$[25]. S_k may be interpreted as the number of points among the first k points of the TB which falls into the prediction bands.

In order to determine the length of the EO we confront our expectations about the distribution of future observations (based on fitted regression model) with the actual observations from the TB, denoted by (t_1, x_1), …, (t_m, x_m). Let $e_1, \dots, e_m$ be the actual values of the random variables $E_1, \dots, E_m$ and let for each $1 \leq k \leq m$

[24] Random variable X follows the Bernoulli distribution with parameter p if $P(X = 1) = p = 1 - P(X = 0)$. Random variable X is the outcome of a so called Bernoulli trial, i.e. a random experiment with only two possible results: success (coded as 1) which occurs with probability p or failure (coded as 0) which happens with probability $(1-p)$.

[25] Binomial distribution $B(n, p)$ is a distribution of a number of successes in the n independent Bernoulli trials with probability of success p.

$$s_k = \sum_{i=1}^{k} e_i$$

be the actual number of points among the first k points of the TB which fall into the prediction bands. Recall that if our regression model is true, s_k should follow the binomial distribution $B(k, (1 - \alpha))$. This key observation allows us to find the temporal extent of the EO. We set the end of the EO to be the first moment, t_k, for which an actual value of s_k is an unlikely outcome given our understanding of the past of the process (represented by the fitted regression model). The observed value s_k is considered unlikely if the joint probability of all outcomes for which from the first k points of the TB at most s_k of them fall into the prediction bands is less than some suitably selected low threshold p_0. For the sake of consistency, we use $p_0 = \alpha$.

To summarize the above argument we present the algorithm for finding the length of the EO:

1. Select threshold p_0 (e.g., equal to α) and set $k = 1$.
2. Calculate s_k (i.e., the number of points among the first k points of the TB which fall into the prediction bands).
3. Let $F_{k,(1-\alpha)}$ be the cumulative distribution function of the binomial distribution $B(k, (1 - \alpha))$. If $F_{k,(1-\alpha)}(s_k) < p_0$ then we set the end of the EO to the moment t_{k-1}, its length H to $k - 1$ and we stop the algorithm.
4. If $k = m$ (i.e., TB is exhausted) then we cannot determine the end point of the EO. We stop the algorithm and set EO length H to ∞.
5. Set $k = k + 1$ and go to point 2.

3.3. Procedure of prognostic learning based on regression method

Below we provide the procedure for PL based on the regression techniques presented above. It is a method-specific version of the generic PL procedure formulated in Section 2.2.

1. Choose the regression technique (e.g., polynomial regression of certain order) which will be used to understand the data behavior in the LB.
2. Choose the initial length k of the LB $X_0, \dots, X_\tau$, $\tau = k$. (Note that k should be large enough with respect to the complexity of selected type of regression function in order to ensure good estimates of the trend function parameters and to prevent overfitting[26].)
3. Fit the regression model to the LB $X_{\tau-k}, \dots, X_\tau$.
4. Construct the EO starting at time τ following the guidelines presented in Section 3.2 and determine its length H.

[26] That is, a situation in which the flexible trend function is not sufficiently constrained by the short sample of data points and too closely mimics the random layout of the data points. Overfitting has strong negative impact on the quality of model predictions.

5. If $\tau < T$ set $\tau = \tau + 1$ and go to step 3. If not, end the procedure.

Note that in step 3 we ignore a part of the LB $X_0, \dots, X_\tau$ discarding all but last k points. In effect, at each stage of the learning procedure we fit a regression model to the data points falling into a window of fixed length k, which we move along the learning sample in the course of the learning procedure. We call this version of PL method "rolling window". Using a window of fixed length is advantageous in two ways. First, it allows for easier comparison of EOs at different stages of the PL procedure, since the width of each EO is determined not only by the uncertainty of the regression model but also by the number of points used for fitting the model. If this number is fixed, the width of the EO depends only on appropriateness of regression model to grasp the data behavior in corresponding LBs. Second, using only k last points from each LB makes the method more responsive to the local behaviour of the data, acknowledging that the recent data points are more relevant to the direction of the EO than the points from the beginning of the learning sample. Throughout this report the "rolling window" learning procedure will be used[27].

We conclude this chapter by emphasizing that the formulas for the estimates of diagnostic and prognostic uncertainty as well as for the prediction bands defining the shape of the EO given in Section 3.1 are applicable exclusively to polynomial regression. However, the method of constructing the EO described in Section 3.2, and prognostic learning procedure given in Section 3.3, are readily applicable to any type of regression method for which the prediction bands can be calculated and extrapolated beyond the range of the LB. (Note, however, that the assumption of independence of residuals of the fitted regression model must be satisfied). For example, these sections are immediately applicable to the prognostic learning procedure based on non-parametric regression (as demonstrated in the appendix).

4. Assessment of prognostic learning performance in the controlled conditions: Monte Carlo experiments

Before we apply the PL procedure based on polynomial regression (described in the previous chapter) to real-life data we first test its performance under controlled conditions, that is, we conduct Monte Carlo experiments by repetitively running the PL method on synthetic datasets.

Having full knowledge about the true trend in the synthetic data and control over the strength of noise disturbing that trend allows us to clearly identify the strengths and weaknesses of the PL method and the reasons for them. This enables us to draw useful conclusions and to formulate guidelines for applying the PL method in analysis of the real-life data.

By choosing to work with synthetic data we overcome a problem of data scarcity, which often occurs when working with real-life data. A real data sample is often too short to

[27] Another version of the PL method which makes use of the whole learning block at each stage and is as easy to implement as the "rolling window" procedure (in step 3 of the procedure one only needs to fit a model to all points $X_0, \dots, X_\tau$ instead of the last k ones). We call this version "expanding". It is useful when we want to check whether the selected regression model is able to correctly describe the system's dynamics over the whole period covered by the learning sample. This method is also used in the appendix where we employ nonparametric regression techniques to describe the behaviour of the data in the learning block. As these methods use only local information (i.e., regression curve is determined only by the nearby points, not the whole sample) the effect of increasing length of LBs on the EO (especially in its width) is negligible.

support the application of a PL method of higher order[28], whereas a synthetic data sample may be of any desired length. In addition, we can always afford to have an extra sample used exclusively for testing our expectations about the length of the EO. Moreover, we can generate multiple independent data samples following the same fixed deterministic trend and compare the performance of the PL method applied to each of them. This allows us to study the stability of the method. In addition, we can repeatedly compare the predicted and actual lengths of the EO starting at the end of the learning sample in order to test the extent to which we can use the insight given by the PL method about the dynamics of the observed system to inform us about its immediate future.

In the present chapter we describe the method which we use to generate synthetic data samples used for testing the PL method in controlled conditions, the purpose and setup of performed numerical experiments, and their results. We conclude this chapter with some general observations and guidelines of applying the prognostic learning procedure based on the polynomial regression.

4.1. Method of generating the synthetic data

The synthetic data samples are generated in the following way:

1. We choose the length of the sample N. For simplicity we assume that $t_k = k, 1 \leq k \leq N$, where t_k denote the times for which synthetic observations are generated.
2. We choose a suitable trend function f which synthetic data will follow.
3. We choose the strength of the noise with which we disturb the true trend f. This strength is defined by the standard deviation σ of the noise, which we express as a percentage of the width of range of the trend function values[29], for example, $\sigma = 0.01 \times \left(\max_{1 \leq k \leq N} f(t_k) - \min_{1 \leq k \leq N} f(t_k) \right)$.
4. We generate a synthetic sample (t_k, x_k), $1 \leq k \leq N$, by setting $x_k = f(t_k) + \varepsilon_k$, where $\varepsilon_1, \dots, \varepsilon_N$ is a sequence of independent random variables following normal distribution of zero mean and standard deviation σ.

In Section 4.3 we present results of running the PL method on five different synthetic datasets. Two of them follow polynomial trends which belong to the family of regression functions used in the employed regression method. These are: the linear trend and the 4th order polynomial trend. They were selected in order to test the performance of the PL method on trends of low (linear) and high (4th order polynomial) complexity in nearly ideal conditions[30], where polynomial regression may give an unbiased[31] model fit.

[28] Learning block required for good estimation of parameters of higher order polynomial trend may be of comparable length as the whole learning sample leaving too few points for meaningful testing of the explainable outreach

[29] Expressing the strength of noise in relation to the range of the true trend function instead of in absolute terms allows us for easy comparison of different types of synthetic data samples.

[30] In principle, in noiseless conditions it would be possible to determine both past and future behaviour of the data given only relatively few points in the LB.

[31] We say that estimator is unbiased if its expected value is equal to the estimated quantity. In case of regression methods, we say that fitted trend $\hat{f}$ is unbiased estimate of true trend f if $\mathrm{E}\left(\hat{f}(t)\right) = f(t)$ for

The remaining three synthetic datasets do not follow trends of the polynomial type, thus allowing us to test the performance of the PL method in situations where the employed regression technique is not able to reproduce the true trend in the data (i.e., it provides only a biased estimate of the true trend). Moreover, they are intended to mimic the types of behavior often encountered in the real-life data. The considered synthetic samples follow: an exponential trend (an increasing trend whose rate of increase accelerates), a logarithmic trend (increasing but with decreasing slope) and a sinusoidal trend with long period of oscillations, comparable with the length of the sample (to mimic a situation when apparent local trends in the historical data are in fact results of slow, long-term oscillations).

Before we present the actual results of applying the PL method on the abovementioned synthetic data samples, in the following section we describe the setup and details of performed experiments.

4.2. Description of experiments on synthetic data

The numerical experiments we perform for each of the abovementioned types of synthetic data involve multiple Monte Carlo runs of the "rolling window" variant of the polynomial regression based PL procedure. Each of the experiments corresponds to a fixed combination of value of order of the method (i.e., the degree of polynomial used in the regression model), level of noise, and length of the LB.

Objectives of these experiments are two-fold. First, we want to identify situations (i.e., patterns in the local behavior of the data forming the LB and the strength of the noise) in which the proposed method of prognostic learning presents its strengths or performs poorly. Second, we investigate the influence of the order of the PL method, the strength of the noise, and the length of the LB on the performance of the PL method.

In addition to realizing these objectives, we explore the reliability of predictions of future EO lengths both in-sample (i.e., using the actual EO lengths[32] in stages up to the present one in order to predict the EO length in the next stage of the PL procedure) as well as out-of-sample (i.e., using EO lengths calculated for all stages of the PL procedure in order to predict the length of the EO starting at the end of the learning sample on which the PL procedure was run). In both cases predictions are made by fitting the linear function (with use of the OLS method) to all available (finite) values of past EO lengths and then extrapolating it to the future point of interest[33].

Note that in-sample predictions may be compared against the actual EO lengths calculated during the learning procedure. Predictions of EO out-of-sample lengths can be tested in similar way, however, this requires an additional testing sample back-to-back with the

all t within the range (period) of the sample. A fitted regression model is necessarily biased if the true trend does not belong to the family of considered regression functions.

[32] Actual EO length is the length of the EO determined with use of data from the testing block. In contrast, predicted EO length is just our (untested) expectation about the length based on the knowledge of actual lengths of EOs from previous stages of the learning procedure.

[33] This is just one, straightforward but possibly crude way of making such predictions. Application of some more subtle methods (e.g., time series model) may improve reliability of such predictions. This will be tested in future research.

learning sample used in the PL procedure. Obtaining such sample is not a problem for the synthetic data—one can easily generate it.

For a single learning sample and corresponding additional testing sample one can only get one pair of predicted and actual lengths of EO starting at the end of the learning sample. However, both values may be to a large extent random, and having only one such pair is not very informative. Much more information carries their joint distribution. Working with synthetic data allows us to easily obtain an empirical estimate of this joint distribution by means of repetitive Monte Carlo simulations.

Below we describe the procedure that each of the experiments follow:

1. Select the functional trend which the synthetic data sample will follow. Choose the length N of the learning sample and the strength of the noise.
2. Select the order of the PL method and the length of the LB (window) k to be used.
3. Select the number of repetitions of the experiment M.
4. Set the current iteration (Monte Carlo run) number i to 1.
5. Generate the synthetic data sample of length $2N$ (cf. Section 4.1). Use the first N points as a learning sample for PL procedure and the remaining data as the additional testing sample to be used exclusively for determining the actual length of the EO starting at the end of the learning sample.
6. Run the "rolling window" prognostic learning procedure on the learning sample generated in step 5. At each stage of the procedure check the fulfillment of assumptions of the polynomial regression model fitted to the LB and record the score of the EO, its actual length and the predicted EO length for this stage, given the EO lengths for previous stages (cf. Figure 3, left panel).
7. After the PL procedure is complete use the calculated EO lengths (in-sample) to predict the length of the EO starting at the end of learning sample (out-of-sample).
8. In order to test the predicted length of the EO starting at the end of learning sample (cf. step 7) calculate the actual length of the EO starting at the end of this sample. To do so, take the LB consisting of the last k points of the learning sample, fit a regression model to it and extrapolate the prediction bands to determine the shape of the EO. To find its length use the data from the additional testing sample (cf. Figure 3, right panel).
9. If $i < M$ then set $i = i + 1$ and go to step 5. Otherwise end the experiment.

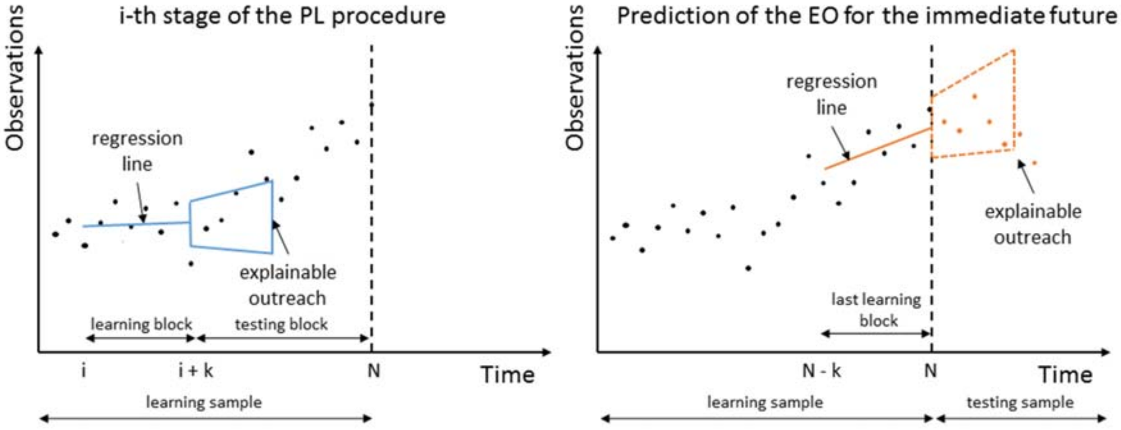


Figure 3. Schematic picture of the Monte Carlo experiment. Left panel: One stage of the prognostic learning procedure with "rolling window" of length k. Regression model is fitted to the data forming a LB $[i, i+1, \dots, i+k]$. Prediction bands for this model define the shape of EO starting at $i+k$. Actual length of the EO is determined with use of the data from the TB. **Right panel:** Determining the actual length of the EO starting at the end of the learning sample (prediction for the immediate future). The direction and shape of the EO is given by the last k points from the learning sample (last LB). Since there are no points left in the testing sample to form a TB, the actual length of the out-of-sample EO is determined with use of the additional testing sample.

With use of the insights gathered by performing the abovementioned experiments we formulate guidelines for selecting the order of the method and length of the LB yielding optimal performance of the PL method. By this we mean:

(1) Satisfactory level of fulfillment of the assumptions of the regression model fitted to each LB.

(2) EOs calculated at different stages of PL method that are as long and narrow as possible (i.e., with high score - cf. Section 2.1). Stable behavior of EO lengths at different stages of the PL procedure is desirable.

(3) Ideally, good reliability of the predictions of EO lengths (both in-sample and out-of-sample).

4.3. Results

In this section we present the results of five sets of Monte Carlo experiments on five different types of synthetic data. This allows us to assess usefulness of the proposed methods of prognostic learning under controlled conditions. In each set of experiments we investigate the influence of: (1) the order of the method, (2) the length of the LB and (3) the level of noise on the performance of prognostic learning, by varying these parameters. Below we present results only for Monte Carlo runs of the PL methods on synthetic data with a low level of noise[34]. For each considered order of method the optimal length of the LB is presented. General conclusions about the marginal influence of each

[34] Results of Monte Carlo runs on data with a higher level of noise are used to formulate general conclusions about the influence of the strength of noise on the PL method.

of the three abovementioned factors on the performance of PL method are presented in Section 4.4.

4.3.1. Data following a linear trend

We begin our analysis of performance of the PL method by testing it in the simplest possible setting, that is, on the synthetic noisy data following a linear trend. This type of trend in the data is easily detected and robustly estimated using the OLS technique, even for relatively short samples. Hence, even the simplest linear regression model fitted to the data in (any) LB not only accurately represents the in-sample data behavior but also correctly grasps the dynamic governing the whole sample. Figure 4 depicts an exemplary synthetic sample following the linear trend which is used in the set of Monte Carlo experiments, the parameters of which are outlined in Table 2.

As one might have expected, the 1st order PL method is able to accurately approximate the true trend in the data, even with use of short LBs of 30 points—see Figure 5. However, ability to correctly estimate the true trend means that for majority of stages of the learning procedure EOs have infinite (undefined) lengths (cf. Figure 6: infinite EO lengths do not appear on the plot, finite lengths occur sporadically). This is due to the fact that an exact description of the true trend in the whole sample (given only information contained in the LB) is, in this case, equivalent to obtaining a precise model of the data generating process, which also holds true beyond the LB. As a consequence, we cannot falsify our understanding of the process based on the data form LB with use of the TB (i.e., part of the learning sample which follows the LB), and thus EO is infinite. Since most of the EOs in-sample are of infinite length we are also unable to formulate expectations about the limits to extrapolating our understanding of the process beyond the learning sample (i.e., the length of EO starting at the end of learning sample).

Table 2. Experiments setup.

True trend formula	$f(t) = 0.1 \times t$
Length of the synthetic data sample	200 points
Length of the learning sample	100 points
Order of PL method	1, 2
Length of the LBs	30, 40
Strength of the noise[35]	0.05
Number of Monte Carlo runs for each parameter combination	40

[35] Expressed as a fraction of the range of the true trend (cf. Section 4.1)

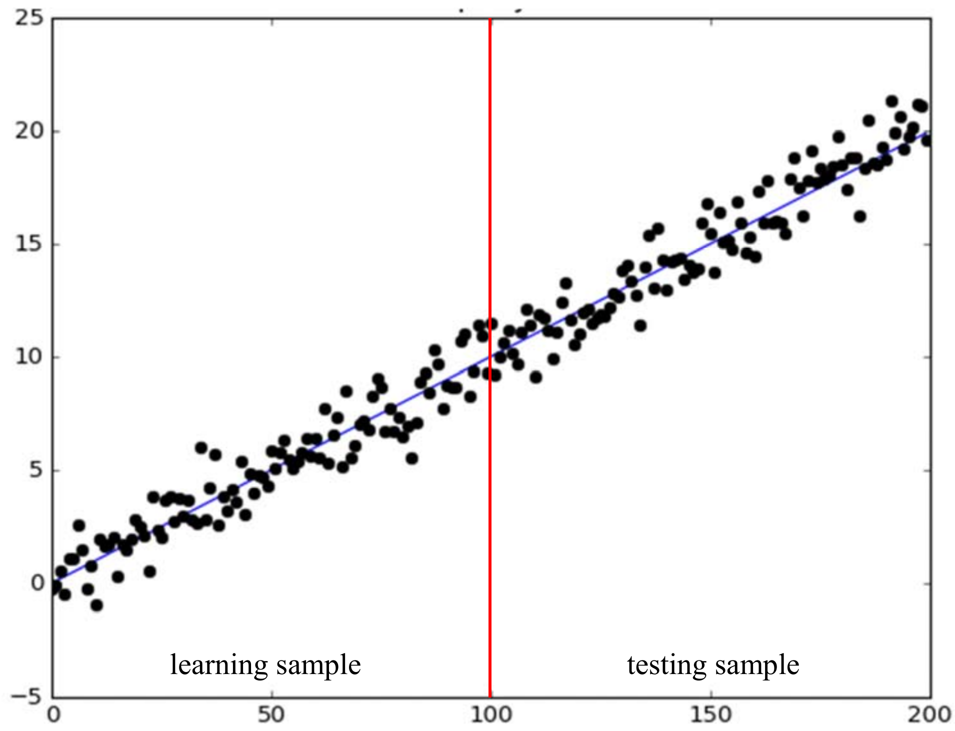


Figure 4. Exemplary data (black dots) following a linear trend $f(t) = 0.1\ \times t$ (blue line). Standard deviation of noise $\sigma = 0.05 \times (\max f - \min f)$.

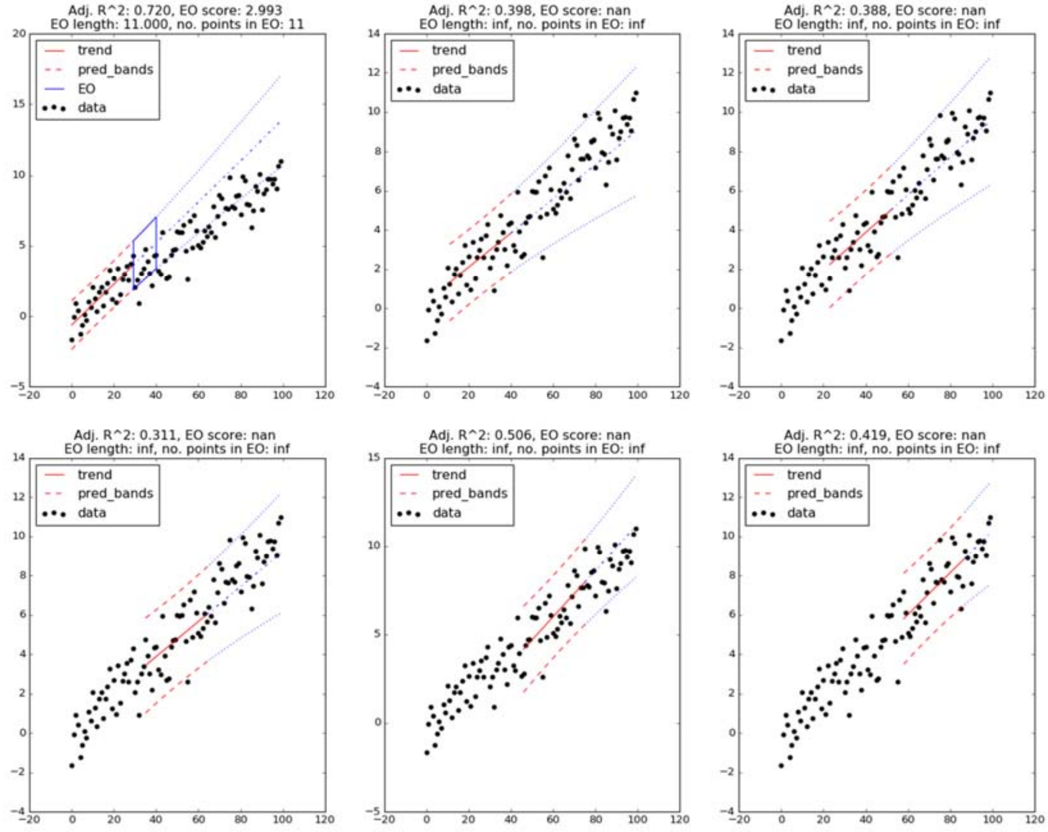


Figure 5. Six exemplary stages of the 1st order PL procedure with LB length of 30 points.

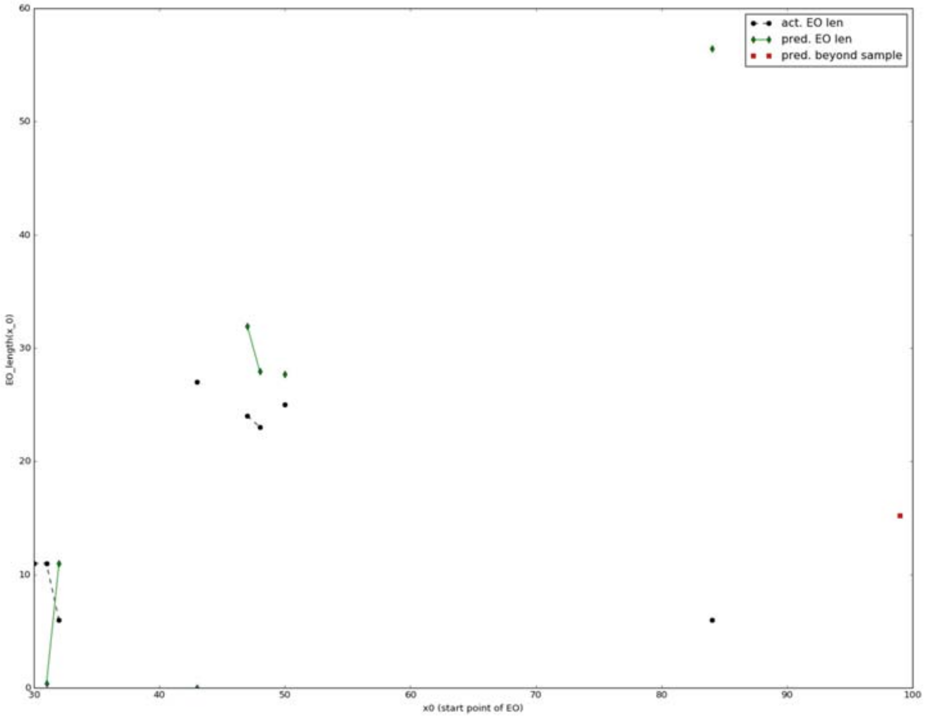


Figure 6. Actual (black dots) and predicted (green diamonds) EO lengths for all stages of the 1st order PL procedure with LB length of 30 points. Correlation between actual and predicted EO lengths is -0.175. The red square marks the predicted length of the EO starting at the end of testing sample. The prediction is based on all finite actual EO lengths calculated in the learning procedure (i.e., all of the black dots). Note that all of the EO lengths (both actual and predicted) are not longer than the length of the LB.

In the case of noisy data following a linear trend, the use of higher order PL methods (using trend functions more complex than the true linear trend) is not advisable. We demonstrate this with the example of 2nd order PL procedure. As one can see on Figure 7, prediction bands for the 2nd order polynomial regression diverge much faster than the analogous prediction bands for linear regression. As a result, the EOs obtained in the process of 2nd order PL procedure mostly have infinite lengths. Moreover, the more flexible 2nd order polynomial model is more visibly susceptible to the influence of noise in the data, and thus producing less certain and robust, often ill-directed projections. Therefore, any EO of finite length obtained with use of the 2nd order method is unreliable as it is most likely ill-directed and overly wide.

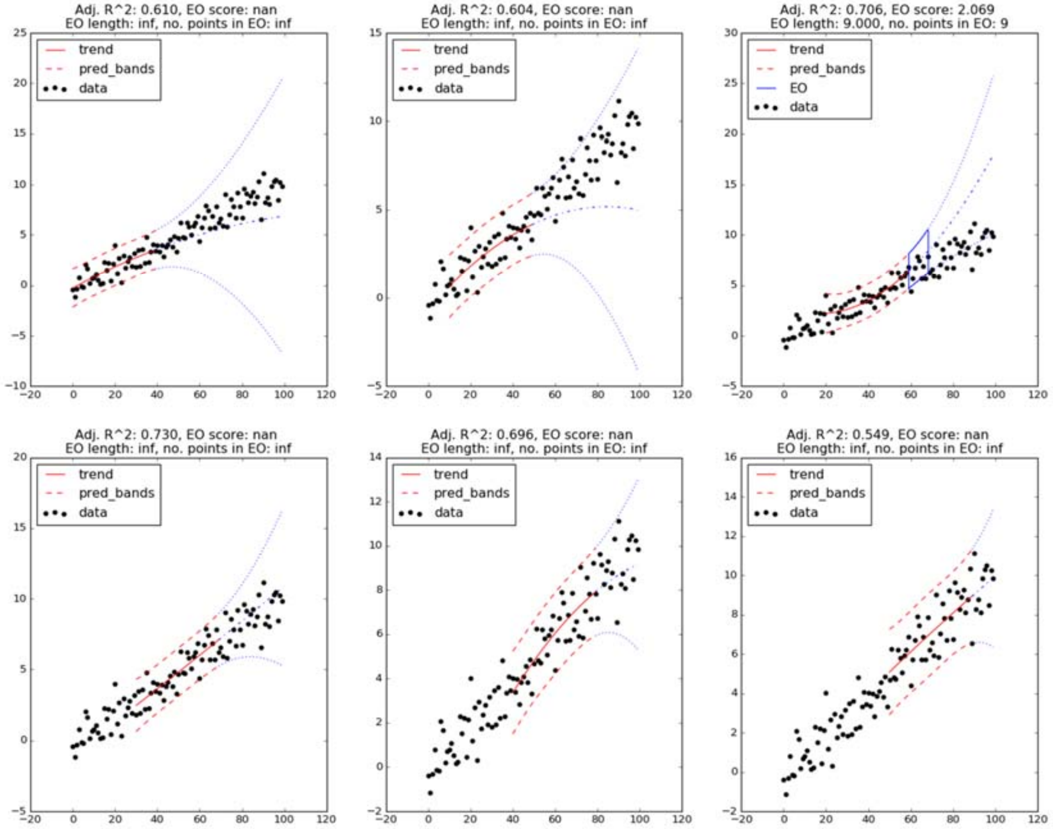


Figure 7. Six exemplary stages of the 2nd order PL procedure with LB length of 40 points.

4.3.2. Data following a 4th order polynomial trend

In the next set of experiments we analyze the performance of prognostic learning method applied to the noisy data following the trend of higher complexity. Method of polynomial regression is in principle able to provide an unbiased estimate of such trend. In Table 3 we gather the parameters of these experiments. Figure 8 shows an exemplary synthetic data sample used in these experiments.

Table 3. Experiments setup. 4th order polynomial trend.

True trend formula	$f(t) = (0.001 \times (t-50))^4 - (0.09 \times (t-50))^3 + (0.5 \times (t-50))^2 - t - 50$
Length of the synthetic data sample	400 points
Length of the learning sample	200 points
Order of PL method	1, 2, 3, 4
Length of the LBs	20, 30, 40, 50, 60
Strength of the noise	0.01, 0.05, 0.1
Number of Monte Carlo runs for each parameter combination	40

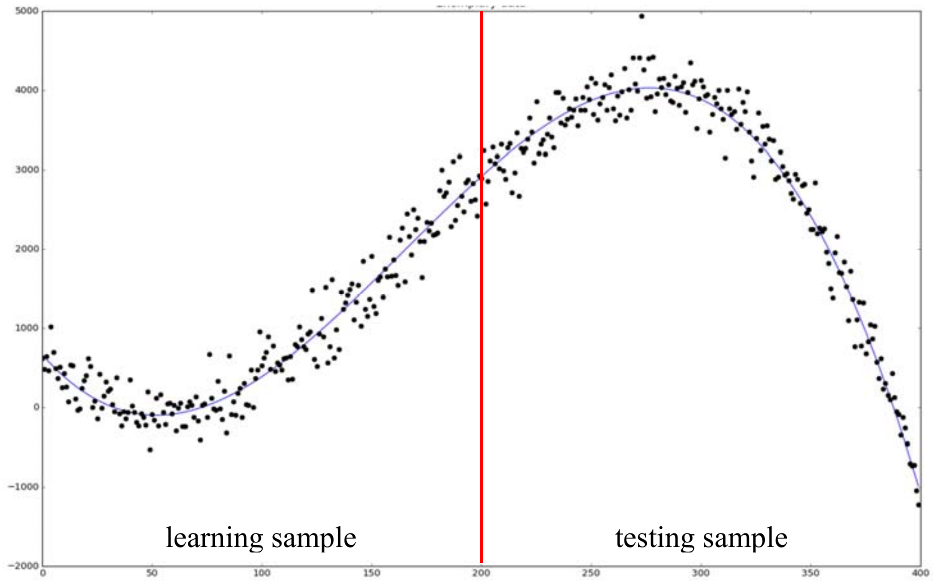


Figure 8. Exemplary data (black dots) following 4th order polynomial trend (blue line) given by the formula $f(t) = (0.001 \times (t-50))^4 - \left(0.09 \times (t-50)\right)^3 + (0.5 \times (t-50))^2 - t - 50$. Standard deviation of the noise $\sigma = 0.05 \times (\max f - \min f)$.

Table 4 presents the results obtained for the synthetic data with a low level of noise[36] (i.e. 0.01 of width of the trend function range). For each order of the PL method the optimal LB length is used.

Table 4. Choices of the LB lengths for different orders of the PL method yielding the best results of experiments on data following a 4th order polynomial trend.

Method order	LB length	Noise level	Regression assumptions	EO Scores	EO lengths	Correlation: actual vs. predicted EO lengths (in sample)	Actual EO lengths (out-of-sample)	Predicted EO lengths (out-of-sample)	Correlation: actual vs. predicted EO lengths (out-of-sample)
1	40	0.01	Ok	0.01 -0.08	Slightly increasing Average: 15 (1 – 30)	0.54	Mode 25 [6 – 37]	Mode 18 [12 – 33]	0.2 (finite EO length in 40 out of 40 runs)
2	50	0.01	Ok	0.03 – 0.08	Oscillating decreasing 130 to 0	0.63	Flat Mode below 50 [0-180]	Left skew Mode 0 [0 – 40]	0.09 (finite EO length in 38 out of 40 runs)
3	40	0.01	acceptable (possible autocorrelation of residuals)	Up to 0.03, mostly undefined	Oscillating [2 – 10] few outliers up to 18	-0.05	[3 – 14]	[0 – 10]	0.09 (finite EO length in 7 out of 40 runs)

[36] For stronger noises the performance of the PL method deteriorates, which to certain extent may be compensated by increasing the length of the learning block.

4	50	0.01	Ok	Up to 0.02, mostly undefined	Oscillating [1 – 15] outlier at 48	-0.09	[1–19], mostly below 6	[0 – 19], mostly below 5	-0.39 (finite EO length in 10 out of 40 runs)

Surprisingly, the best performance is achieved for the variant of PL method which employs a 1st order regression over short LBs (just 40 points). Figure 9 illustrates six exemplary stages of such PL procedure. This optimal combination of the order of method and the length of LB yields relatively stable behavior of the EO lengths with oscillations that are not too strong around a slightly increasing trend (cf. Figure 10). The ranges of the actual and predicted lengths of the EO starting at the end of learning sample are in good agreement, although the correlation between these lengths is weak (see Figure 11). Notice also that all EO lengths are not longer than the LB.

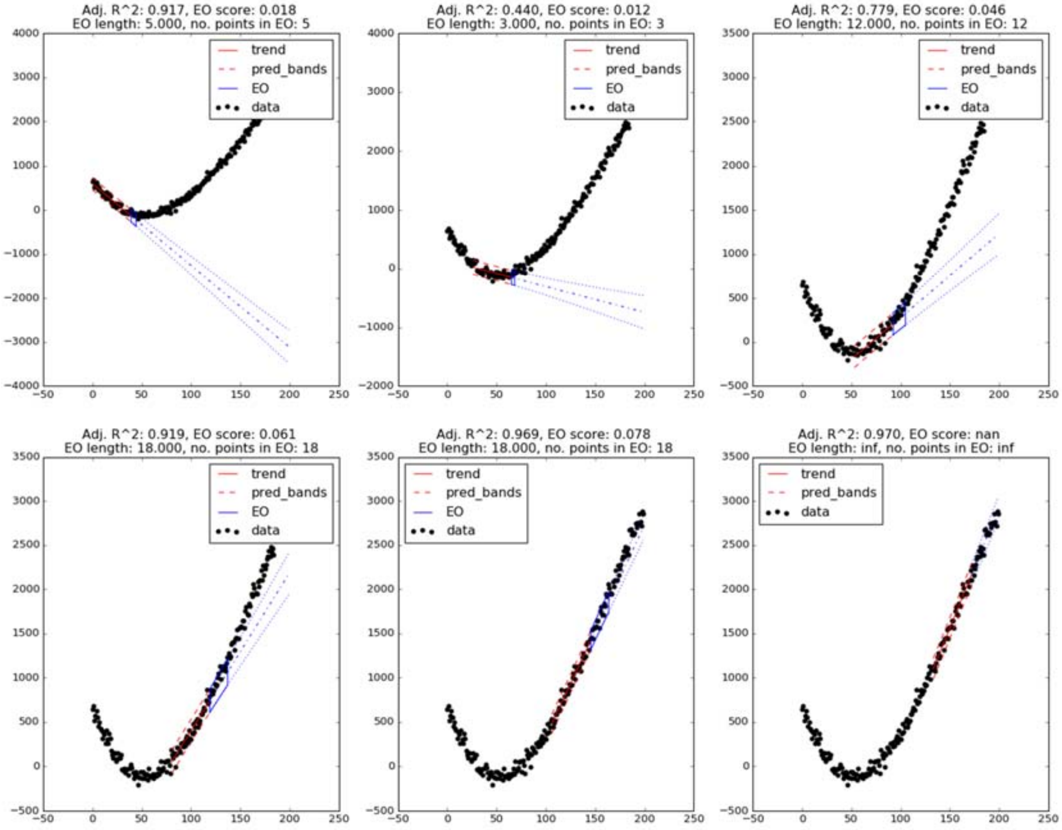


Figure 9. Six exemplary stages of the 1st order PL procedure with LB length of 40 points. In regions where the curvature of the true trend is significant, the linear model does not fit well to the data in the LB and the actual lengths of the EO are low. In regions where the true trend has approximately constant slope the PL method performs well.

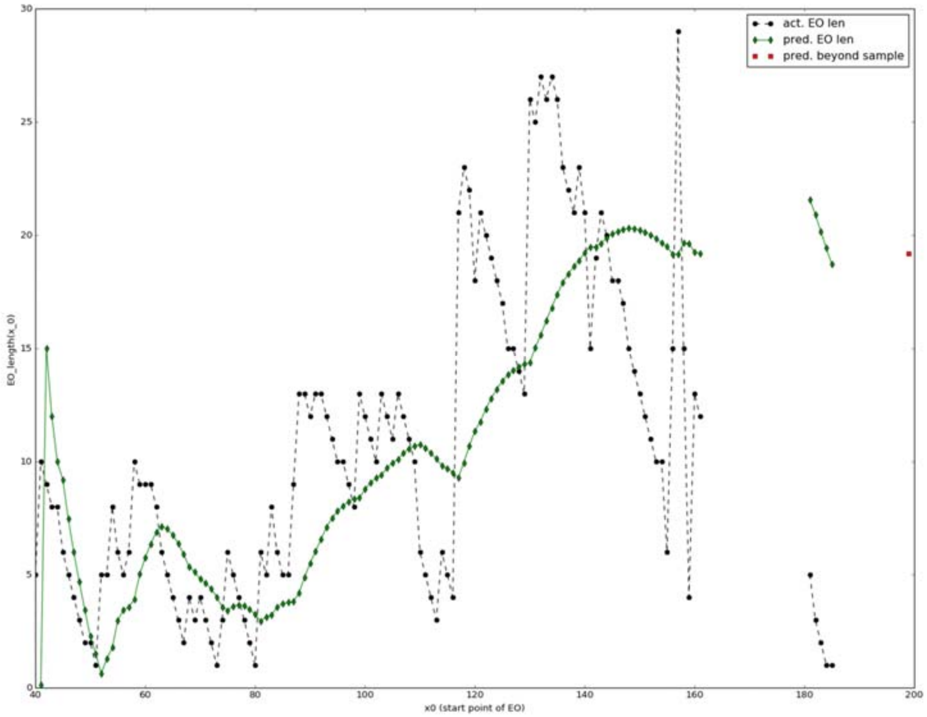


Figure 10. Actual (black dots) and predicted (green diamonds) EO lengths for all stages of the 1st order PL procedure with LB length of 40 points. Correlation between actual and predicted EO lengths is 0.537. The red square marks the predicted length of the EO starting at the end of testing sample. Prediction is based on all finite actual EO lengths calculated in the learning procedure (i.e,. all of the black dots). Note that all of the EO lengths (both actual and predicted) are no longer than the length of the LB.

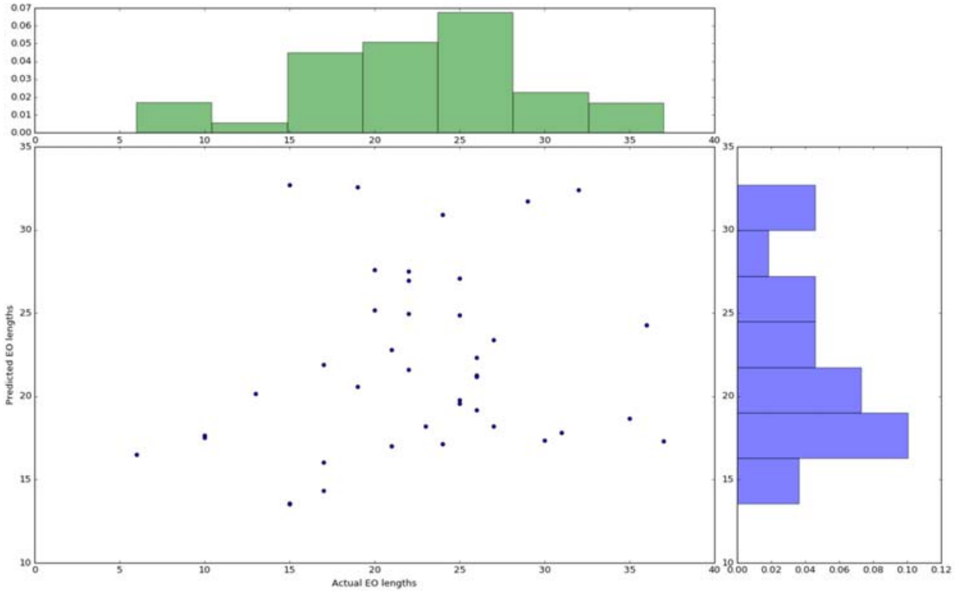


Figure 11. Estimate of joint distribution of actual and predicted lengths of the EO starting at the end of the learning sample. Each of 40 points on the scatter plot represents the result of one Monte Carlo run resulting in finite actual EO length. The total number of Monte Carlo runs is 40. Histograms approximate marginal distributions of actual EO lengths (green) and predicted EO lengths (blue). Their correlation is 0.195.

Equally surprising is a relatively poor performance of the 4th order PL method. Fourth order polynomial trends fitted to learning blocks of length 50 describe the behavior of the data better than linear trends. However, extrapolations using 4th order polynomial regression functions to predict the future behavior of the data are highly uncertain. This is caused by their high flexibility, which within the LB is forced to minimize distance from the data points, but beyond it, when it is unconstrained, it may strongly deviate from the actual trend. This high uncertainty is represented by the fast divergence of the prediction bands. As a result, for most of the stages of the PL procedure we cannot determine the length of the EO because the extremely wide prediction bands cover all points in the TB (cf. Figures 12 and 13). This phenomenon also has a strong impact on both predicted and actual lengths of the EO starting at the end of the learning sample. Although ranges of the actual and predicted lengths are in very good agreement, there are only a few cases in which these lengths are finite, undermining the meaningfulness of the results of Monte Carlo experiments (cf. Figure 14).

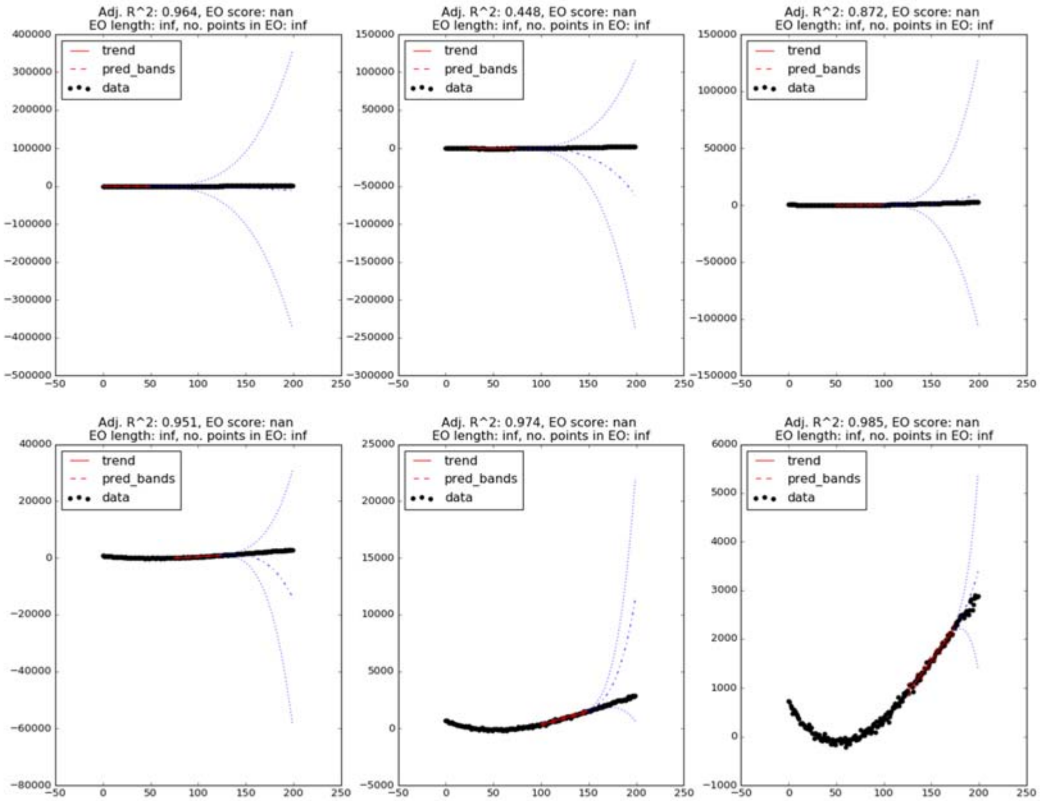


Figure 12. Six exemplary stages of the 4th order PL procedure with LB length of 50 points. Note that often extrapolated trend deviates substantially from the actual data in the testing sample. High uncertainty of these predictions is exhibited by quickly diverging prediction bands.

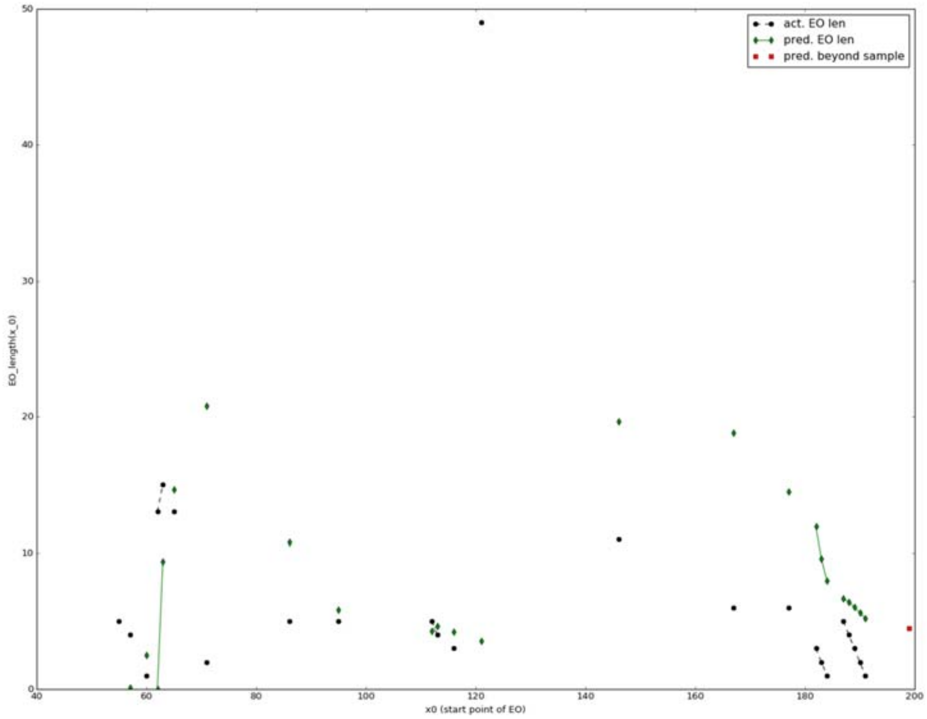


Figure 13. Actual (black dots) and predicted (green diamonds) EO lengths for all stages of the 4th order PL procedure with LB length of 50 points. Correlation between actual and predicted EO lengths is -0.086. The red square marks the predicted length of the EO starting at the end of testing sample. Prediction is based on all finite actual EO lengths calculated in the learning procedure (i.e., all of the black dots). Note that all of the EO lengths (both actual and predicted) are not longer than the length of the LB.

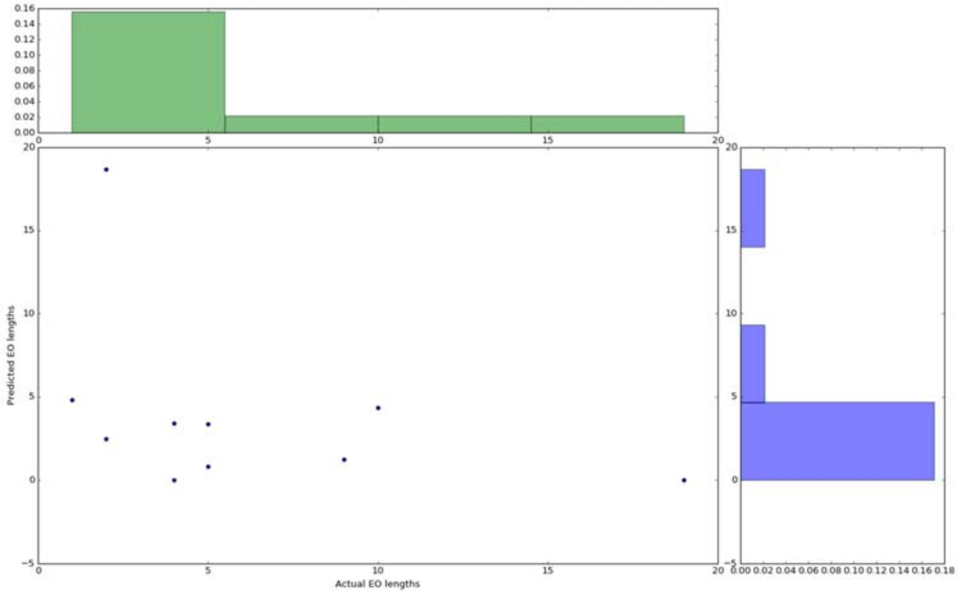


Figure 14. Estimate of joint distribution of actual and predicted lengths of the EO starting at the end of the learning sample. Each of 10 points on the scatter plot represents the result of one Monte Carlo run resulting in a finite actual EO length. The total number of Monte Carlo runs is 40. The histograms approximate marginal distributions of actual EO lengths (green) and predicted EO lengths (blue). Their correlation is -0.387.

4.3.3. Data following exponential trend

In this set of experiments we analyze the performance of the PL method applied to the noisy data following a commonly occurring type of trend not belonging to the family of polynomials. Although it is not possible to model the data following exponential trend with any polynomial in the long run, it is possible to achieve a satisfactory local approximation with the use of a polynomial function of sufficiently high order. Hence, a PL method describing the local[37] behavior of the data with a polynomial regression model is also expected to be applicable in this case. In Table 5 we gather the parameters of Monte Carlo experiments on synthetic exponential data. Figure 15 shows an exemplary synthetic data sample used in these experiments.

Table 5. Experiments setup. Exponential trend.

True trend formula	$f(t) = \exp(0.01\ \times (t + 100))$
Length of the synthetic data sample	400 points
Length of the learning sample	200 points
Order of PL method	1, 2, 3,
Length of the LBs	20, 30, 40, 50
Strength of the noise[38]	0.001, 0.005, 0.01
Number of Monte Carlo runs for each parameter combination	50

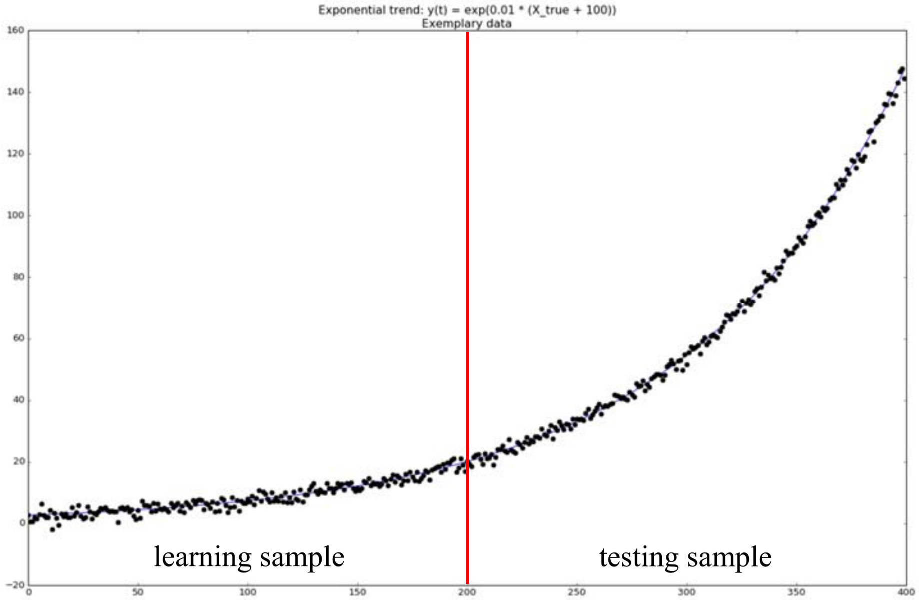


Figure 15. Exemplary data (black dots) following exponential trend (blue line) given by formula $f(t) = \exp(0.01\ \times (t + 100))$. Standard deviation of noise $\sigma = 0.01 \times (\max f - \min f)$.

[37] i.e. only within relatively short learning block

[38] Expressed as the fraction of trend function range width – cf. Section 4.1.

Table 6 gathers the results obtained for the synthetic data with low level of noise[39] (i.e., 0.001 of width of the trend function range). For each order of the PL method the optimal LB length is used.

Table 6. Choices of the LB lengths for different orders of the PL method yielding the best results of experiments for synthetic data following an exponential trend.

Method order	LB length	Noise level	Regression assumptions	EO Scores	EO lengths	Correlation: actual vs. predicted EO lengths (in sample)	Actual EO lengths (out-of-sample)	Predicted EO lengths (out-of-sample)	Correlation: actual vs. predicted EO lengths (out-of-sample)
1	40	0.001	Ok (possible autocorrelation of residuals)	Oscillating, gradually decreasing [42 to 2]	Oscillating, decreasing [30 to 1]	0.75	Flat [1 – 10]	Flat [0 – 5]	-0.03 (finite EO length in 50 out of 50 runs)
2	40	0.001	Ok	Oscillating below 20, mostly undefined	Oscillating, slight decrease [35 to 1], few outliers up to 80	0.34	Flat [0 – 200]	Left skew [0 – 30] Mode 0	0.27 (finite EO length in 50 out of 50 runs)
3	50	0.001	Ok	Oscillating below 11.2, mostly undefined	Decreasing [20 to 3] Outliers up to 75	0.02	Left skew [4 – 80] Majority below 20	[0 – 8]	0.26 (finite EO length in 10 out of 50 runs)

The best performance is achieved for the 1st order PL method using short LBs (of just 40 points). Six exemplary stages of such PL procedure are visualized in Figure 16. For the initial stages of the PL procedure, EOs are relatively long (because of small initial changes in the slope of the exponential trend), but become shorter over the course of the procedure (as the increase in exponential trend accelerates) – cf. Figure 17. The ranges of the actual and predicted lengths of the EO starting at the end of the learning sample are comparable (see Figure 18). The range of values of predicted EO lengths is narrower than the range of actual EO lengths, which means that expected EO length is likely to underestimate the actual EO length. However, they are virtually uncorrelated.

[39] For stronger levels of noise the performance of the PL method deteriorates, which to certain extent may be compensated by increasing the length of the learning block.

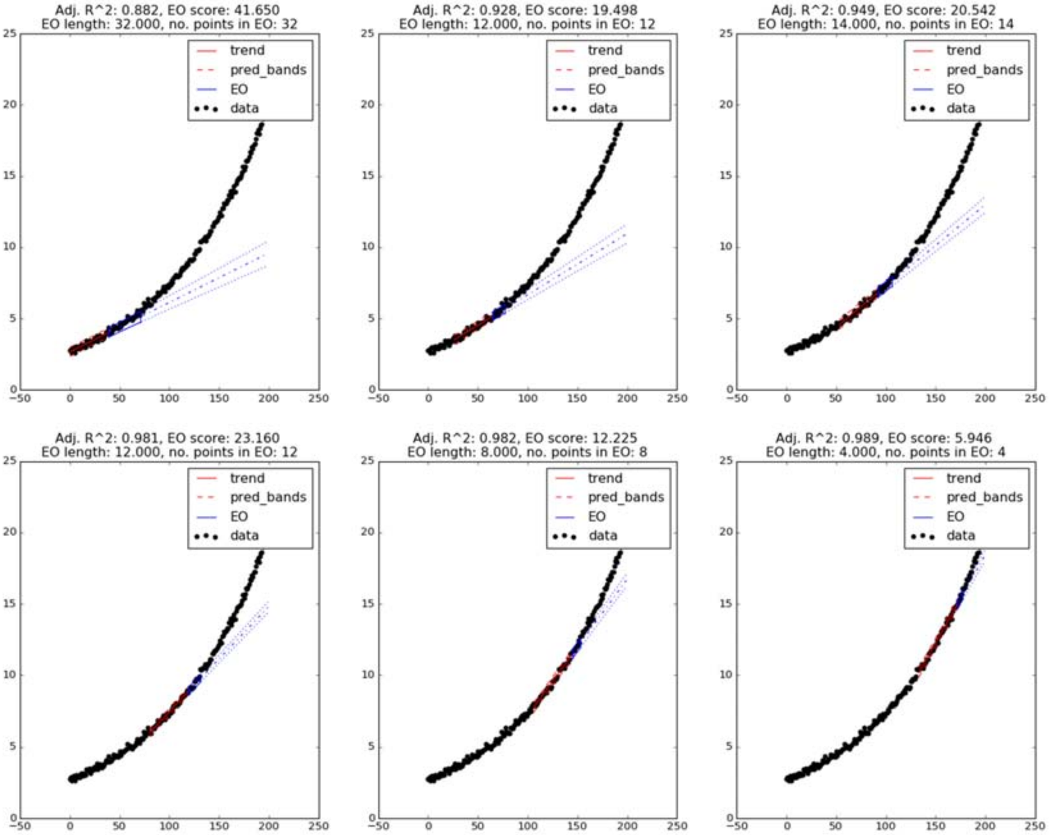


Figure 16. Six exemplary stages of the 1st order PL procedure with LB length of 40 points. For the initial stages of the PL procedure the lengths of the EO are comparable with the length of the LB. This is due to a slow initial increase of the exponential trend. As this increase begins to accelerate in later stages the EO lengths get shorter.

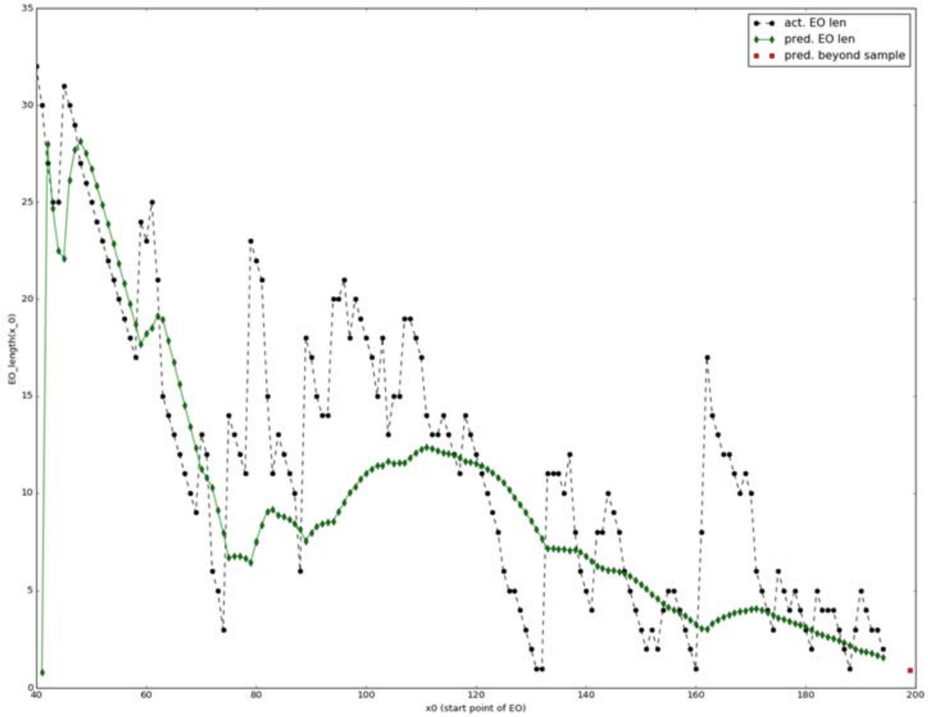


Figure 17. Actual (black dots) and predicted (green diamonds) EO lengths for all stages of the 1st order PL procedure with a LB length of 40 points. Correlation between the actual and predicted EO lengths is 0.746. The red square marks the predicted length of the EO starting at the end of testing sample. Prediction is based on all finite actual EO lengths calculated in the learning procedure (i.e., all the black dots). Note that all of the EO lengths (both actual and predicted) are not longer than the length of the LB.

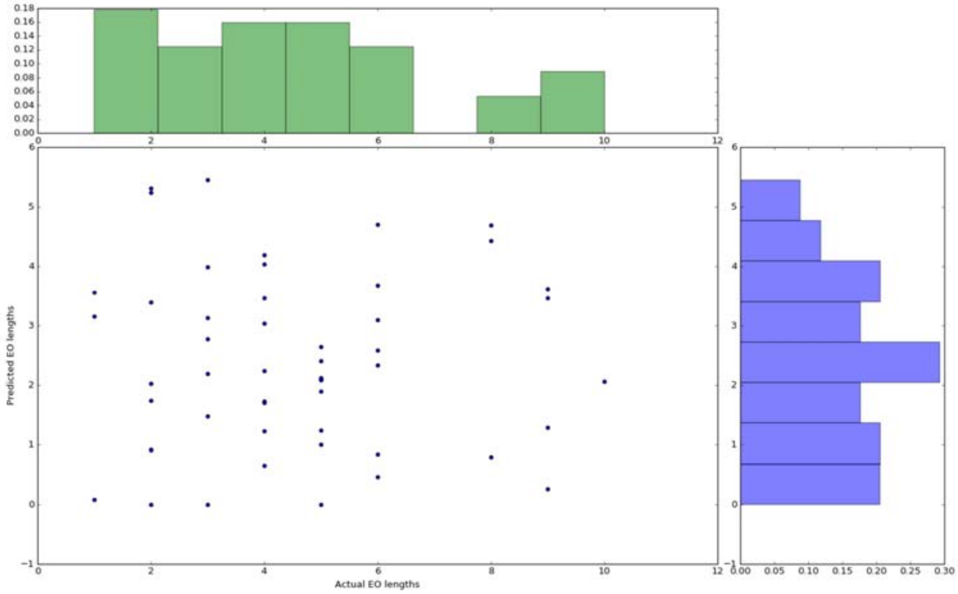


Figure 18. Estimate of joint distribution of actual and predicted lengths of the EO starting at the end of the learning sample. Each of 50 points on the scatter plot represents the result of one Monte Carlo run resulting in finite actual EO length. Total number of Monte Carlo runs is 50. The histograms approximate the marginal distributions of actual EO lengths (green) and predicted EO lengths (blue). Their correlation is - 0.032.

Higher order polynomials are much better at approximating the exponential trend, yet the performance of higher order PL methods is worse than for the one based on linear regression. We discuss this using the example of the 2nd order polynomial method. Fitted quadratic trends extrapolated beyond the corresponding LBs always increase slower than true exponential trend (yet quicker than linear trends). However, prediction bands are usually wide enough to cover all the data points in the TB. As a result, for most of the stages of the PL procedure we cannot determine the length of the EO (cf. Figures 19 and 20). The distribution of the predicted lengths of EO starting at the end of learning sample is strongly skewed to the left and has much narrower support than the relatively flat distribution of the actual EO lengths at the end of the learning sample (see Figure 21). Thus, the predicted EO length is likely to heavily underestimate the actual length of the EO, while the correlation of these two is weak.

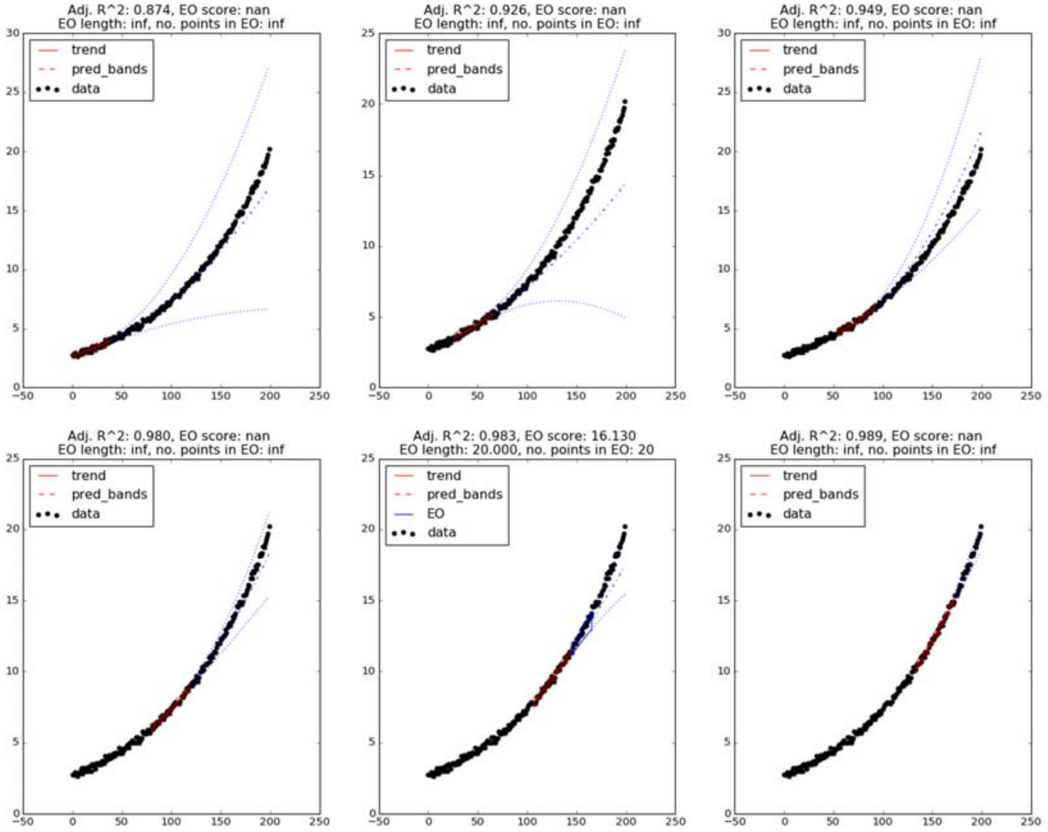


Figure 19. Six exemplary stages of the 2nd order PL procedure with a LB length of 50 points.

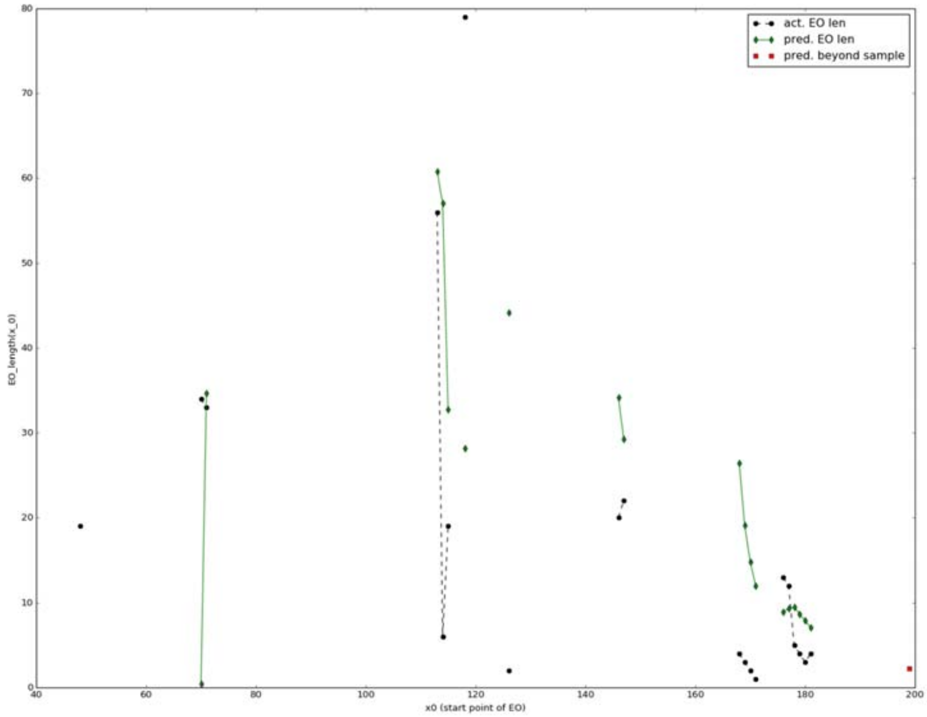


Figure 20. Actual (black dots) and predicted (green diamonds) EO lengths for all stages of the 2nd order PL procedure with a LB length of 50 points. Correlation between the actual and predicted EO lengths is 0.335. The red square marks the predicted length of the EO starting at the end of testing sample. The prediction is based on all finite actual EO lengths calculated in the learning procedure (i.e,. all of the black dots). Note that majority of the EO lengths (both actual and predicted) are not longer than the length of the LB.

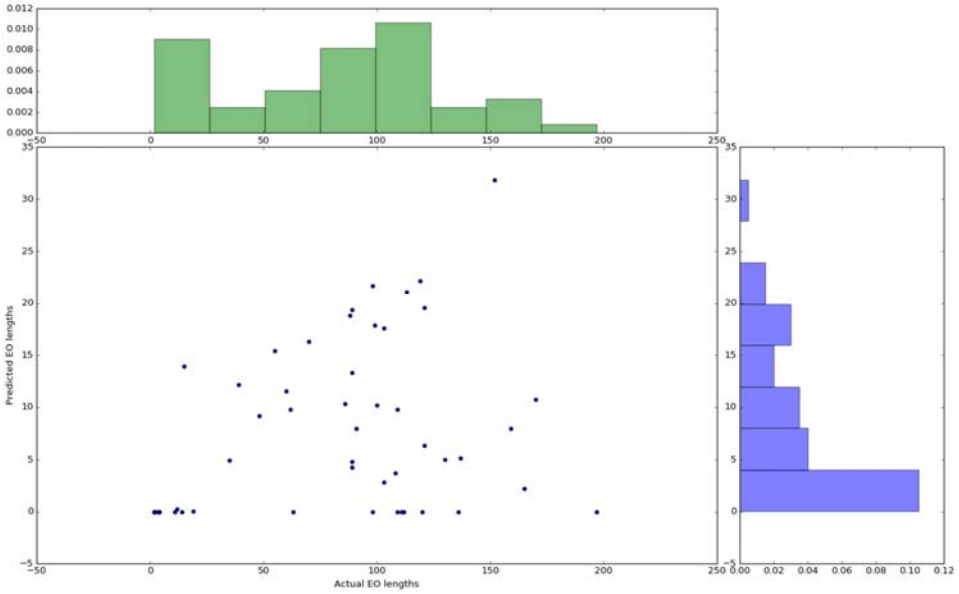


Figure 21. Estimate of a joint distribution of actual and predicted lengths of the EO starting at the end of the learning sample. Each of 50 points on the scatter plot represents the result of one Monte Carlo run resulting in finite actual EO length. The total number of Monte Carlo runs is 50. The histograms

approximate marginal distributions of actual EO lengths (green) and predicted EO lengths (blue). Their correlation is 0.286.

4.3.4. Data following logarithmic trend

Now we examine the performance of the PL method on the synthetic data following an increasing but decelerating trend—exemplified by a logarithmic trend. This trend, often encountered in real-life data, cannot be approximated well by any polynomial in the long run, however, a satisfactory local (i.e., for a relatively short subsample) agreement may be achieved. This is the rationale for applying the PL method to such type of data. In Table 7 we gather the parameters of the Monte Carlo experiments on synthetic logarithmic data. Figure 22 shows an exemplary synthetic data sample used in these experiments.

Table 7. Experiments setup. Logarithmic trend.

True trend formula	$f(t) = \log(0.05\ \times (t + 50))$
Length of the synthetic data sample	400 points
Length of the learning sample	200 points
Order of PL method	1, 2, 3,
Length of the LBs	20, 30, 40, 50
Strength of the noise[40]	0.01, 0.025, 0.05
Number of Monte Carlo runs for each parameter combination	50

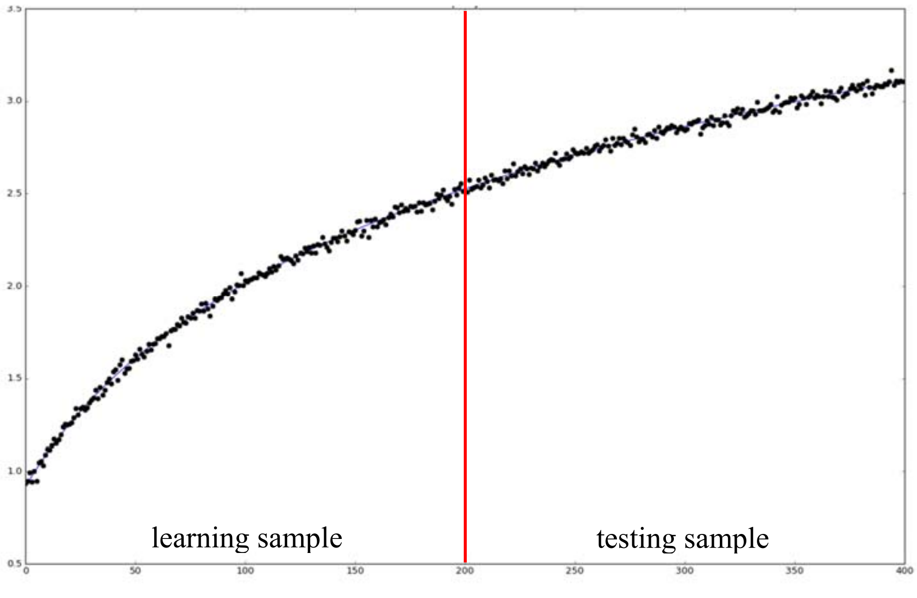


Figure 22. Exemplary data (black dots) following a logarithmic trend (blue line) given by the formula $f(t) = \log(0.05\ \times (t + 50))$. Standard deviation of noise $\sigma = 0.01 \times (\max f - \min f)$.

[40] Expressed as the fraction of trend function range width – cf. Section 4.1.

Table 8 summarizes the results obtained for the synthetic data with a low level of noise[41] (i.e., 0.01 of the width of the trend function range). For each order of the PL method the optimal LB length is used.

Table 8. Choices of the LB lengths for different orders of the PL method yielding the best results of experiments on synthetic data following a logarithmic trend.

Method order	LB length	Noise level	Regression assumptions	EO Scores	EO lengths	Correlation: actual vs. predicted EO lengths (in sample)	Actual EO lengths (out-of-sample)	Predicted EO lengths (out-of-sample)	Correlation: actual vs. predicted EO lengths (out-of-sample)
1	50	0.01	Ok	Oscillating, below 405, often undefined	Oscillating, max increasing to 40	0.62	[0 – 110] Mode 40	[15 – 50] Mode 30	0.14 (finite EO length in 50 out of 50 runs)
2	50	0.01	Ok	Oscillating [20 – 160], mostly undefined	Oscillating, decreasing [120 to 1]	0.63	[3 – 26]	Left skew [0 – 23] Mode 0	0.66 (finite EO length in 7 out of 50 runs)
3	50	0.01	Ok (occasionally autocorrelation of residuals)	Oscillating [10 – 67], mostly undefined	Oscillating below 15, diminishing outliers (max 30)	0.5	[3 – 26]	[1 – 11]	-0.26 (finite EO length in 7 out of 50 runs)

As in previous sets of experiments, the best performance is achieved with the 1st order PL method—this time using slightly longer LBs of 50 points. Six exemplary stages of this PL procedure are shown on Figure 23. EOs calculated for the initial stages of the PL procedure are short because of the sharply decelerating trend at the beginning of learning sample. The slower rate of decrease of slope of the logarithmic trend in the further part of the learning sample results in longer EOs for later stages (cf. Figure 24). Note also that the range of all (finite) lengths of EOs in-sample is narrower than the LB. This is also the case for predicted lengths of the EO starting at the end of learning sample (see Figure 25). However, the actual lengths of EOs starting at the end of learning sample are significantly longer, while the correlation between the actual and predicted lengths is weak.

[41] For greater levels of noise the performance of the PL method deteriorates, which to certain extent may be compensated by increasing the length of the learning block.

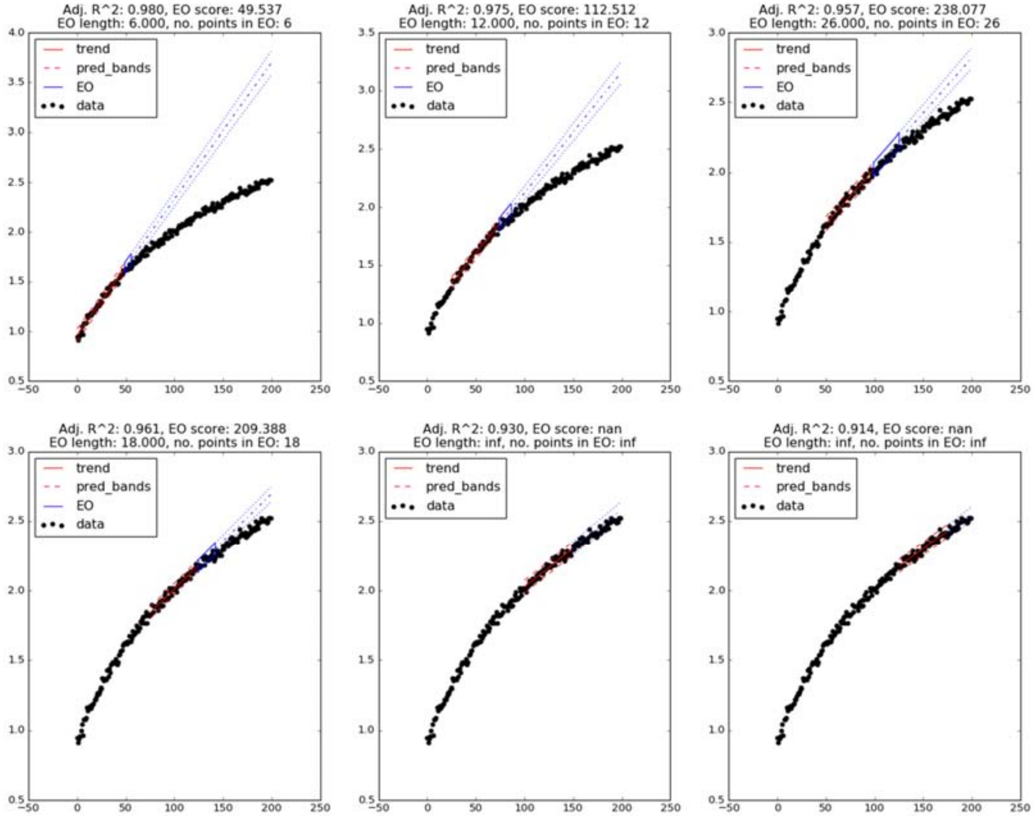


Figure 23. Six exemplary stages of the 1st order PL procedure with a LB length of 50 points. For the initial stages of the PL procedure the lengths of the EO are short because of an initially sharp decrease in the slope of the logarithmic trend. As this decrease begins to decelerate in later stages the EOs get longer.

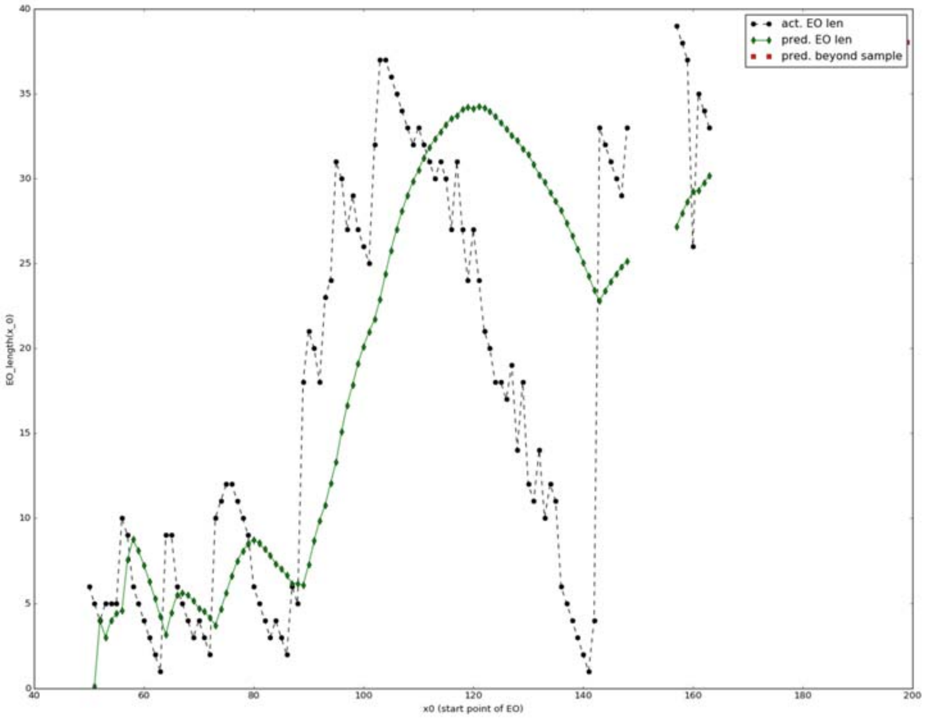


Figure 24. Actual (black dots) and predicted (green diamonds) EO lengths for all stages of the 1st order PL procedure with LB length of 50 points. Correlation between actual and predicted EO lengths is 0.619. The red square marks the predicted length of the EO starting at the end of testing sample. The prediction is based on all finite actual EO lengths calculated in the learning procedure (i.e,. all of the black dots). Note that all of the EO lengths (both actual and predicted) are not longer than the length of the LB.

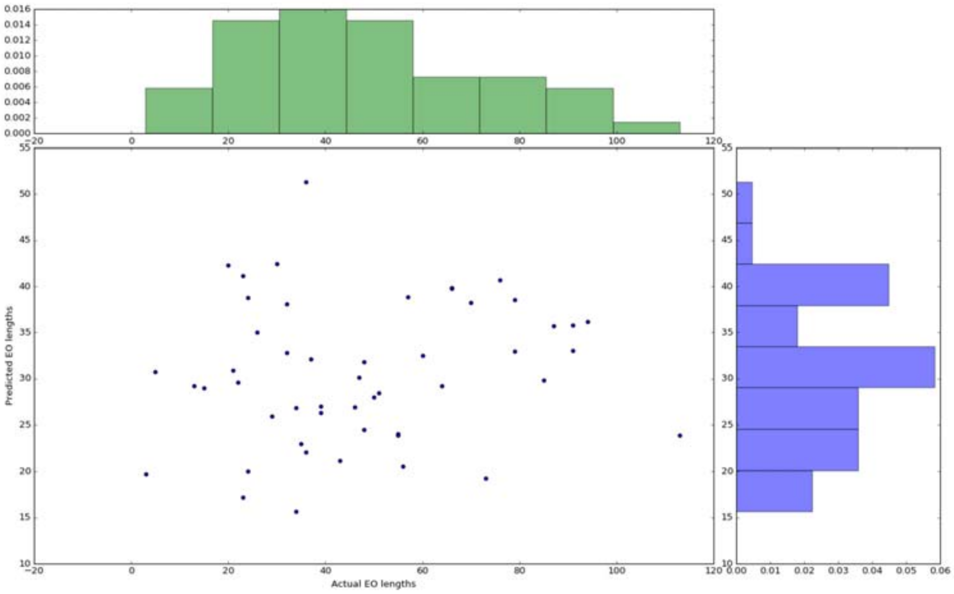


Figure 25. Estimate of the joint distribution of actual and predicted lengths of the EO starting at the end of the learning sample. Each of the 50 points on the scatter plot represents the result of one Monte Carlo run resulting in finite actual EO length. The total number of Monte Carlo runs is 50. The histograms approximate marginal distributions of actual EO lengths (green) and predicted EO lengths (blue). Their correlation is 0.144.

PL methods based on higher order polynomial regressions perform worse than the 1st order method when applied to data following a logarithmic trend (or a similar shape). We discuss this using the example of 2nd order polynomial method. The deviations from the testing data of fitted quadratic trends extrapolated beyond the corresponding LBs increase faster than the analogous deviations of the extrapolated linear trends. In addition, there is often strong misdirection of extrapolated higher order trends, and their prediction bands diverge much faster than those of linear models—see Figure 26. As a result, for the majority of the PL procedure the EOs have an infinite (undefined) length (cf. Figure 27). In addition, the actual length of the EO starting at the end of the learning sample is infinite for the most of the Monte Carlo runs—making any analysis of the joint behavior of predicted and actual lengths of the EO out-of-sample virtually impossible (cf. Figure 28).

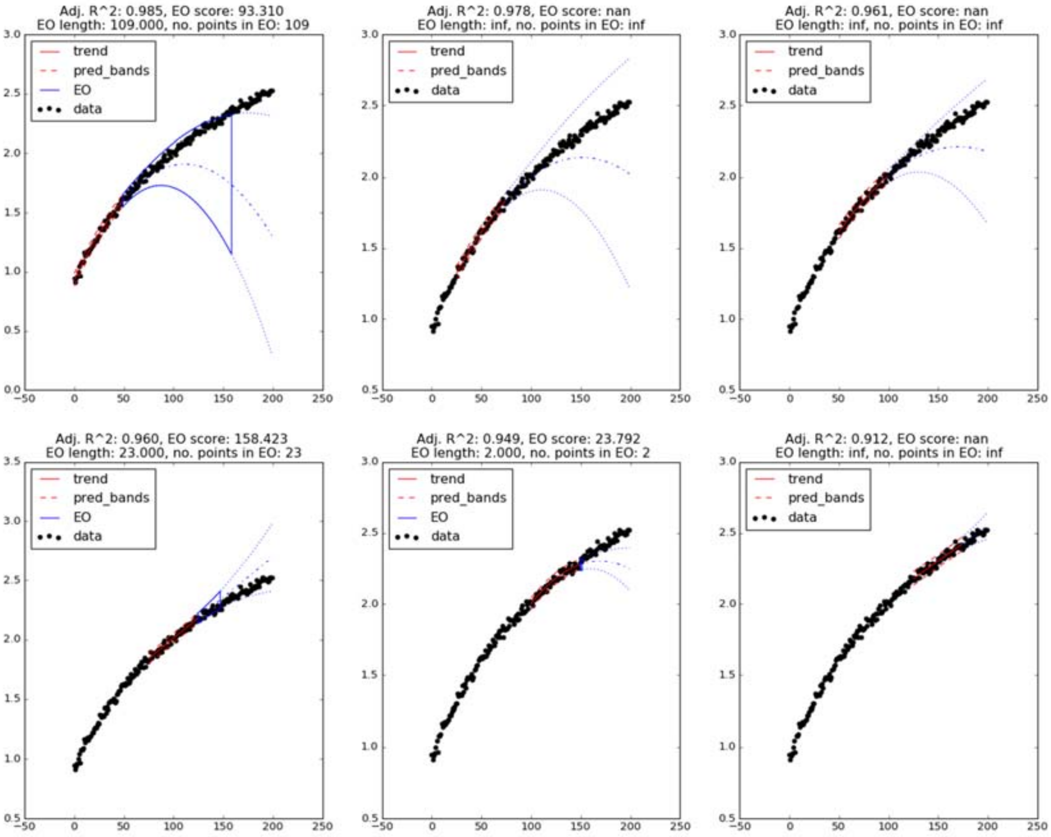


Figure 26. Six exemplary stages of the 2nd order PL procedure with a LB length of 50 points.

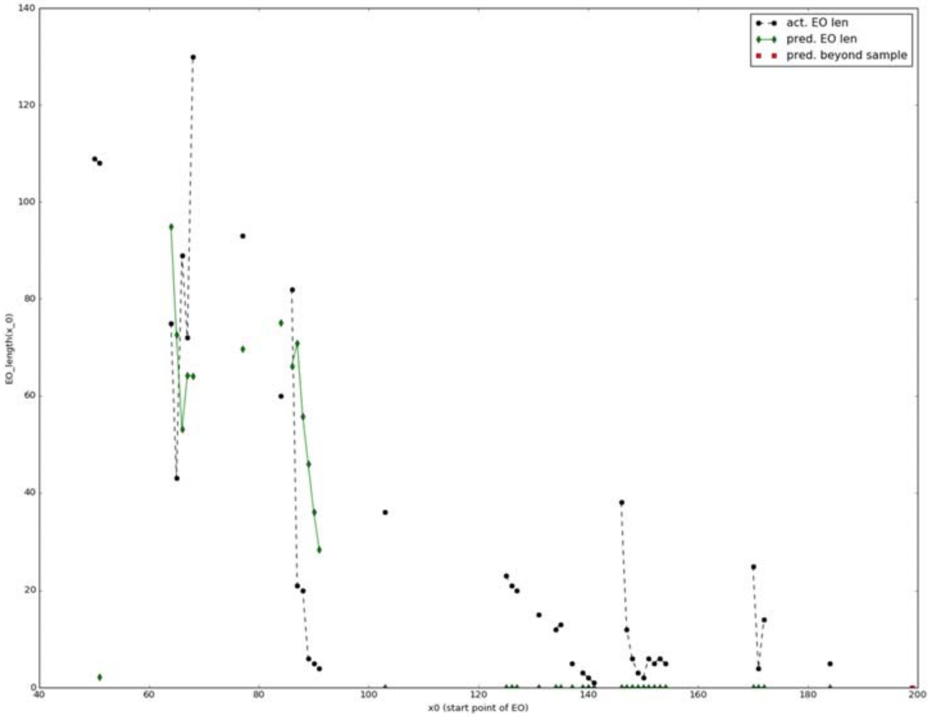


Figure 27. Actual (black dots) and predicted (green diamonds) EO lengths for all stages of the 2nd order PL procedure with LB length of 50 points. Correlation between actual and predicted EO lengths is 0.628. The red square marks the predicted length of the EO starting at the end of testing sample. The prediction is based on all finite actual EO lengths calculated in the learning procedure (i.e., all of the black dots).

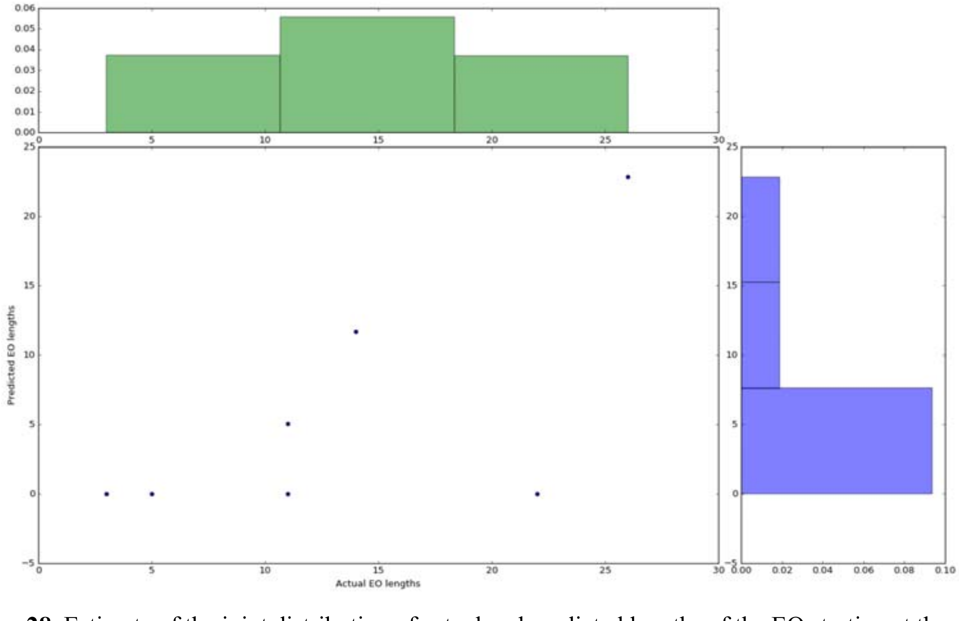


Figure 28. Estimate of the joint distribution of actual and predicted lengths of the EO starting at the end of the learning sample. Each of seven points on the scatter plot represents the result of one Monte Carlo run resulting in a finite actual EO length. The total number of Monte Carlo runs is 50. The histograms approximate marginal distributions of actual EO lengths (green) and predicted EO lengths (blue). Their correlation is 0.664.

4.3.5. Data following periodic trend

In the last set of experiments we investigate the usefulness of the PL method for analysis of data following a sinusoidal trend over a period comparable to the length of learning sample. Within short time intervals (i.e., comparable in length to the LB) such data may appear to follow a clear non-periodic trend, which may be locally approximated by a polynomial. By applying the PL method based on polynomial regression we want to understand the limits of such local approximations. Table 9 outlines the setup of the Monte Carlo experiments on synthetic data following a periodic trend. Figure 29 exhibits an exemplary synthetic data sample used in these experiments.

Table 9. Experiments setup. Exponential trend.

True trend formula	$f(t) = \sin(0.018\ \times (t - 100))$
Length of the synthetic data sample	400 points
Length of the learning sample	200 points
Order of PL method	1, 2, 3,
Length of the LBs	20, 30, 40, 50
Strength of the noise[42]	0.01, 0.05, 0.1
Number of Monte Carlo runs for each parameter combination	50

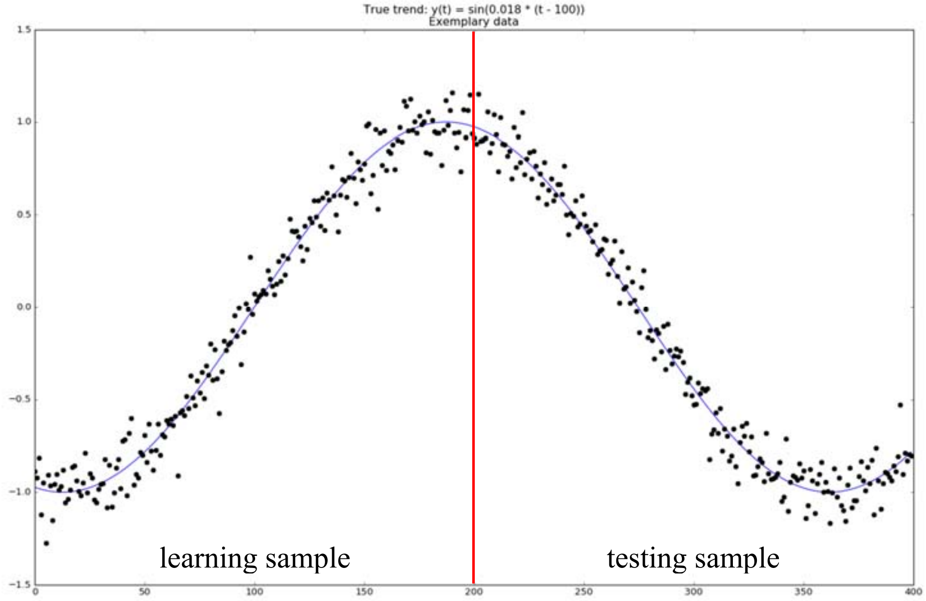


Figure 29. Exemplary data (black dots) following sinusoidal trend with long period (blue line) given by formula $f(t) = \sin(0.018\ \times (t - 100))$. Standard deviation of noise $\sigma = 0.01 \times (\max f - \min f)$.

[42] Expressed as the fraction of trend function range width – cf. Section 4.1.

Table 10 summarizes the results of experiments performed using synthetic data with a low level of noise[43] (i.e., 0.01 of width of the trend function range). For each order of the PL method the optimal LB length is used.

Table 10. Results of experiments for optimal choices of LB lengths in case of synthetic data following a periodic trend.

Method order	LB length	Noise level	Regression assumptions	EO Scores	EO lengths	Correlation: actual vs. predicted EO lengths (in sample)	Actual EO lengths (out-of-sample)	Predicted EO lengths (out-of-sample)	Correlation: actual vs. predicted EO lengths (out-of-sample)
1	30	0.01	Ok	Slowly oscillating, increasing to 395, then gradually decreasing to 10	Oscillating, increasing [1 – 70] then decreasing to 1. Most of the time below 20	0.43	Flat [1 – 11]	[7 – 14] Mode 11	-0.04 (finite EO length in 50 out of 50 runs)
2	50	0.01	Ok	Oscillating below 200, slightly increasing	Oscillating below 40, slightly decreasing, outliers up to 60	0.3	[0 – 150] Mode 100	[0 – 24]	0.14 (finite EO length in 50 out of 50 runs)
3	50	0.01	Ok (occasionally autocorrelation of residuals)	Oscillating [10 – 68], mostly undefined	Oscillating below 20, gradually decreasing outliers up to 40	0.53	[3 – 25]	[1 – 11]	-0.1 (finite EO length in 8 out of 50 runs)

As for the previous sets of experiments, the best performance is achieved for the 1st order PL method using short LBs (of just 30 points). Figure 30 shows six exemplary stages of the PL procedure. For stages of the PL method whose LBs are close to the bending points of the true trend, the EO lengths are relatively short with respect to the length of the LB. However, EOs are much longer when corresponding LBs coincide with regions in which the true trend is nearly linear—see Figure 31. The predicted EO lengths out-of-sample may be slightly over-optimistic—the range of estimated lengths is shifted to the right in comparison to the range of actual lengths of the EO starting at the end of the learning sample (cf. Figure 32). Moreover, the predicted and actual EO lengths are virtually uncorrelated. Note, however, that they are shorter than the length of LBs used in the PL procedure.

[43] For greater levels of noise the performance of the PL method deteriorates, which to certain extent may be compensated by increasing the length of the learning block.

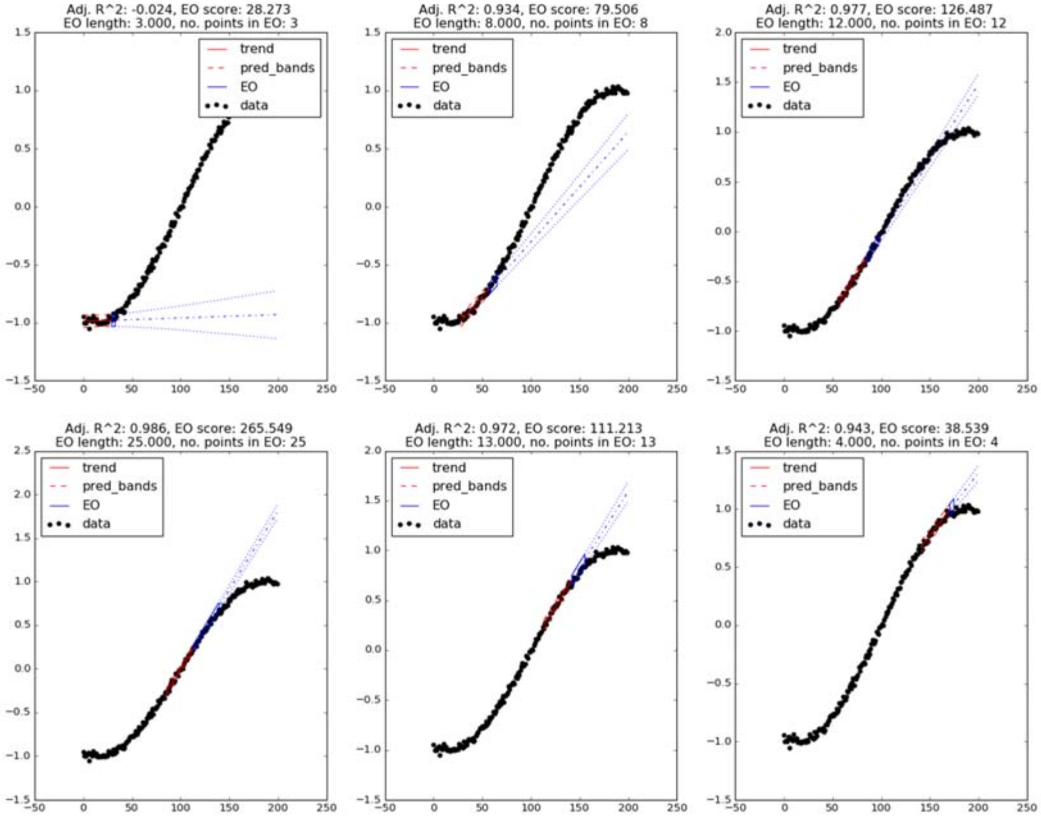


Figure 30. Six exemplary stages of the 1st order PL procedure with a LB length of 30 points. EOs are relatively short in cases when corresponding LBs are close to the bending points of the true trend and long otherwise.

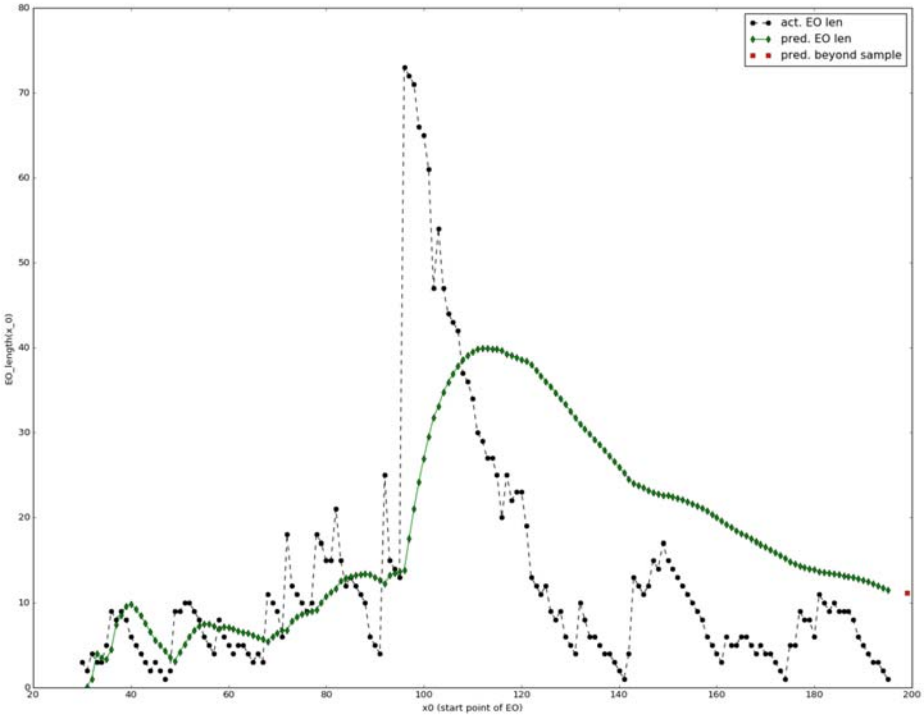


Figure 31. Actual (black dots) and predicted (green diamonds) EO lengths for all stages of the 1st order PL procedure with a LB length of 30 points. Correlation between the actual and predicted EO lengths is 0.434. The red square marks the predicted length of the EO starting at the end of testing sample. The prediction is based on all finite actual EO lengths calculated in the learning procedure (i.e., all of the black dots).

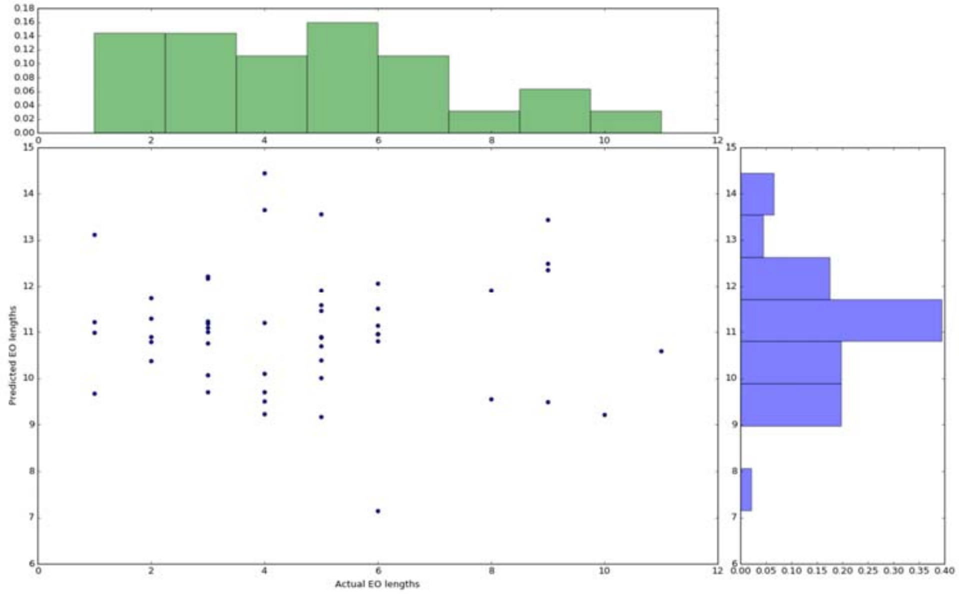


Figure 32. Estimate of a joint distribution of actual and predicted lengths of the EO starting at the end of the learning sample. Each of the 50 points on the scatter plot represents the result of one Monte Carlo run resulting in a finite actual EO length. The total number of Monte Carlo runs is 50. The histograms approximate marginal distributions of actual EO lengths (green) and predicted EO lengths (blue). Their correlation is -0.037.

Higher order polynomials are better suited to describe the local behavior of the data in the LBs than the linear functions, especially when the LB is in the vicinity of the bending points of the true trend (see Figure 33). In comparison to the 1st order method this results in longer EOs for the stages of the PL procedure when the LB coincides with the intervals in which curvature of the true trend is significant—cf. Figure 34. Nevertheless, the EO scores are worse than for the 1st order PL method. This is due to the fact that the prediction bands (defining the shape—and thus score—of the EO) for higher order polynomial regressions diverge faster than for linear regression. Moreover, the flexibility of higher order polynomial trends is not particularly advantageous when predicting the length of the EO starting at the end of the learning sample—the predicted EO lengths grossly underestimate the actual EO lengths while their correlation is weak (see Figure 35).

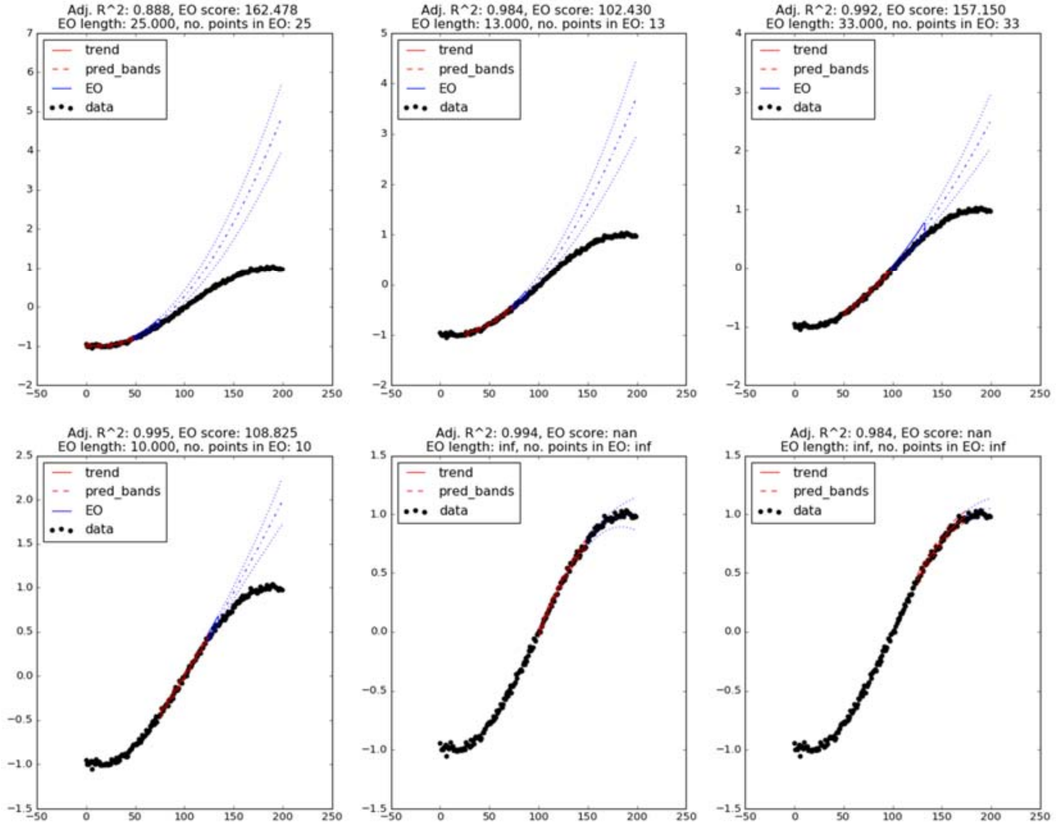


Figure 33. Six exemplary stages of the 2nd order PL procedure with a LB length of 50 points.

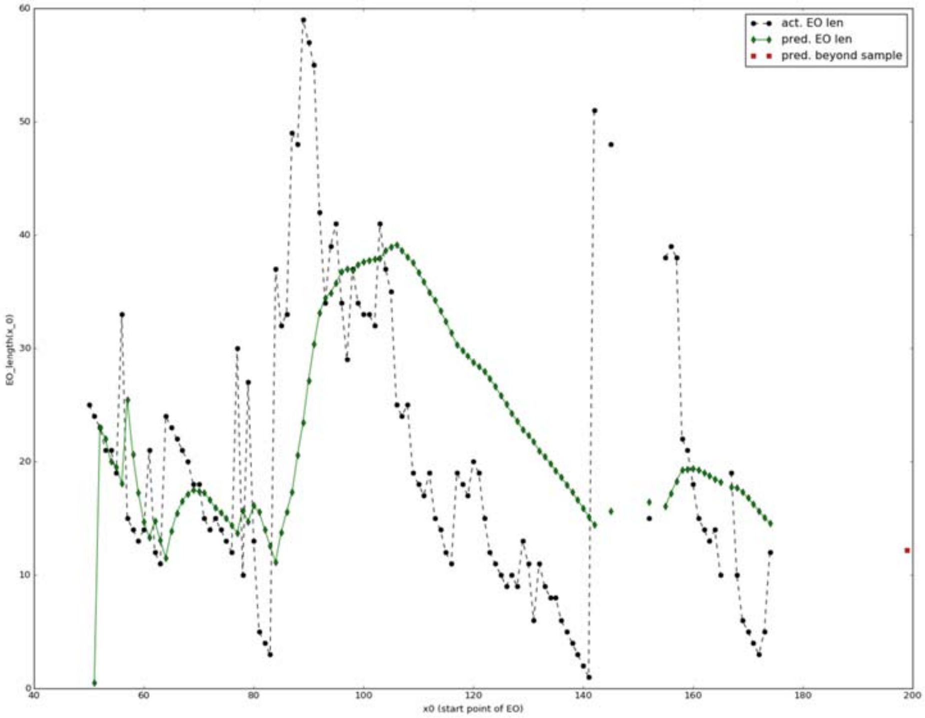


Figure 34. Actual (black dots) and predicted (green diamonds) EO lengths for all stages of the 2nd order PL procedure with a LB length of 50 points. Correlation between the actual and predicted EO lengths is 0.309. The red square marks the predicted length of the EO starting at the end of testing sample. Prediction is based on all finite actual EO lengths calculated in the learning procedure (i.e., all of the black dots). Note that majority of the EO lengths (both actual and predicted) are not longer than the length of the LB.

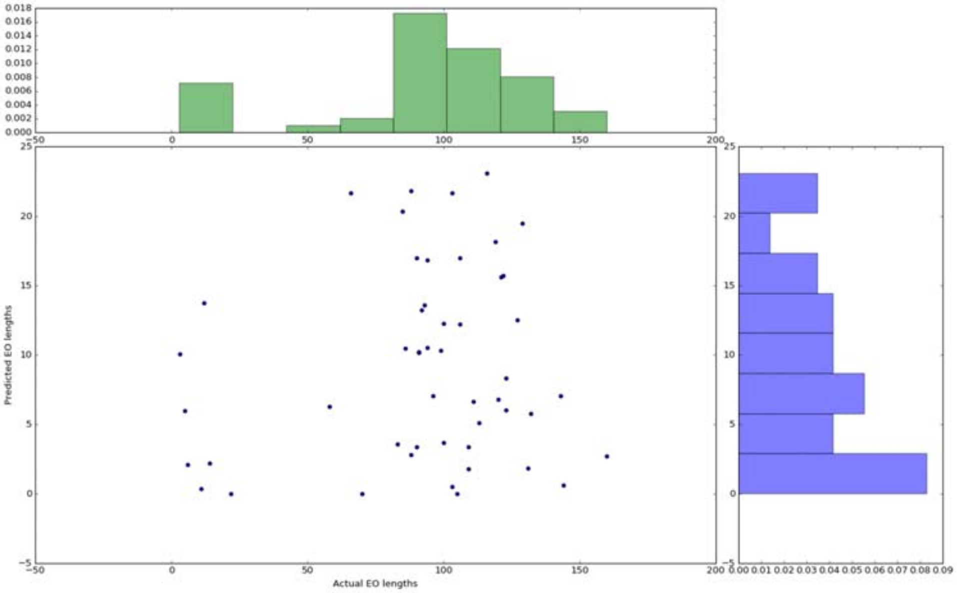


Figure 35. Estimate of the joint distribution of actual and predicted lengths of the EO starting at the end of the learning sample. Each of the 50 points on the scatter plot represents the result of one Monte Carlo run resulting in a finite actual EO length. The total number of Monte Carlo runs is 50. The histograms approximate marginal distributions of actual EO lengths (green) and predicted EO lengths (blue). Their correlation is 0.140.

4.4. Conclusions

In this section we present some general conclusions on the performance of the PL method based on polynomial regression that can be drawn the results of the experiments on synthetic datasets described in the previous two sections.

We begin with analysis of the impact of complexity of the class of regression functions (i.e. order of polynomials) used in the PL method. This factor appears to be the most important for the performance of the prognostic learning. **With increasing complexity**:

- Fulfillment of regression method assumptions does not change significantly, however, assumption violations may be slightly more frequent.
- EO scores decrease, in principle. This is due to the fact that the speed of divergence of the prediction bands—and thus the width of the EO—is of the same order as the polynomial trend used in the underlying regression model. In addition, the number of stages of the PL procedure for which EO scores are undefined (i.e., cases for which EOs have infinite length) usually increase.
- Actual in-sample EO lengths—if finite— generally decrease. Clear tendencies, such as the often-observed decrease of the EO lengths for consecutive stages of the 1st order PL procedure, gradually change to oscillations around a relatively stable level.
- Correlation between actual and predicted in-sample EO lengths typically gets weaker. This correlation is relatively strong in the presence of a clear monotonic trend in the lengths of consecutive EOs obtained in course of the learning procedure. This is most often the case for the 1st order method. As these tendencies in EO lengths change to oscillations typical for higher order methods, this correlation gets weaker.
- Actual out-of-sample EO lengths (which are determined by use of the additional testing sample back to back with the learning sample) typically decrease. This effect is especially clear for the upper limits (maximums) of the observed ranges of finite EO lengths. Moreover, for higher order methods, EOs of infinite (undefined) lengths are predominant.
- Predicted out-of-sample EO lengths decrease, in principle. Moreover, regardless of the order of method, the range of predicted EO lengths usually lies within (or at least significantly overlaps with) the range of the actual EO lengths. Thus, at least on average, predicted EO lengths out of the sample underestimate the actual ones. However, the correlation between actual and predicted EO lengths is typically weak, often negative and in principle not very reliable for higher order methods (as a result of EOs being predominantly infinite).

Increasing the level of noise in the data has, in principle, a negative impact on the performance of PL. The most apparent effect is the deterioration of EO scores as a result of the fact that higher level of noise stipulates wider EOs.

The optimal length of the LB is closely related to the order of the method used. It should not be too short or overly long (we discuss the choice of optimal LB length later in this section). Therefore, it is difficult to discriminate the marginal impact of increasing the length of the LB—what is too short for one method may be too long for another. The clearest effect one sees is for the EO scores. They may slightly improve, as a longer LB allows for better estimation of the parameters of the regression function (lower variance of estimates of regression function parameters).

Based on the experiments on synthetic data described in the previous section, we formulate the following observations about the **1st order method of PL:**

- Any true trend and any data behavior can be locally approximated by a line. This local approximation is relatively robust to the level of noise. As a consequence, ill-directed EOs (if they appear) are the result of the inability of the linear model fitted to the LB to follow the quickly changing true trend, rather than result of noisy conditions.
- Bias[44]–variance trade-offs: The 1st order method is biased—it looks only for linear trends in the data and cannot describe strongly non-linear trends well. This bias may be negligible when the true trend is slowly varying, but can be significant in the presence of a curved true trend in the data. This bias is, however, balanced by the relatively low variance of predictions made using the linear regression model, that is, slowly (at least slower than for higher order methods) diverging prediction bands determining the width (and thus the score) of the EO.
- This has two significant practical consequences:
    - If the true trend is linear then the 1st order method is optimal (prognostic uncertainty is the lowest possible).
    - If the true trend is non-linear then predictions made by extrapolating the linear trend fitted to the LB will eventually be wrong, thus the EO will **almost always have a finite length, usually not greater than the optimal length of the LB**. In this case the length of the EO informs us about the **safe lower band for the time horizon within which treating the dynamics of the data as linear is a good approximation.**
- The optimal length of the LB (and thus of the learning sample) is lowest for the 1st order PL method. This is important for the applicability of the PL method, since in practice data scarcity is a common problem.

Conclusions for the **higher order PL methods** are slightly different:

- Bias–variance trade-offs: any continuous true trend in the data over a specified interval may be well approximated with a polynomial of sufficiently high order. This ability of higher order polynomials to closely follow the data sample reduces the bias of the method. However, in noisy conditions the uncertainty in the estimates of the parameters of the polynomial regression model fitted to the data in a LB almost always results in high variance of predictions beyond the range of the LB (represented by quickly diverging prediction bands).

[44] Here the term "bias" refers to the method. It means that $\mathrm{E}\left(\hat{f}(t)\right) \neq \mathrm{E}(X_t)$ for some t within the range (period) of the sample, where $\hat{f}$ denotes the estimate of the true trend. It is not a systematic (measurement) error of analysed data.

- This has two significant practical consequences:
    - The sharp increase in the uncertainty of predictions made by extrapolation of the fitted polynomial trend beyond the range of the LB makes the usefulness of such predictions questionable.
    - More importantly, because of the flexibility of higher order polynomial trends and the quickly diverging prediction bands **in most cases** (stages of the PL procedure) **EO length is infinite**. Indeed, it is finite only in cases when the extrapolated polynomial trend around which the EO is constructed was so ill-directed that this was not offset by quickly diverging prediction bands. Thus, results of higher order PL methods should be treated somewhat differently and with more suspicion than the results of the 1st order method.
- The required length of the LB is considerably higher than for the 1st order method. A longer LB is needed to prevent overfitting—situation in which the fit of the flexible polynomial trend may be strongly impacted by random noise. This further reduces the usefulness of the higher order PL methods in analysis of relatively short real-life datasets.

We conclude this chapter with a few **rules of thumb for applying the PL method**:

1. The 1st order method should be preferred over the higher order methods.
2. The greater the noise the longer the LB required and the more difficult it is to use the higher order methods.
3. The higher the order of method the longer the LB required. In any case there should be at least 10 points in the LB per each parameter of the regression model to be estimated.
4. Given the data and the order of the PL method one should follow the following guidelines when selecting the **optimal length of the LB**:
    a. Choose the LB length for which the EO score is the highest (or slightly longer).
    b. Choose the LB length for which the EO length exhibits stable behavior in course of the PL procedure (oscillating with few small outliers) or when trends in the behavior of the EO lengths change (e.g., from clear decrease of EO length in course of the PL method to oscillations around a certain level or when a tendency of oscillations becomes apparent).
    c. Choose the LB length for which correlation between actual and predicted EO lengths in-sample is relatively strong and positive.

Ideally these criteria should be fulfilled simultaneously. Choice of the optimal LB length usually coincides with a good behavior of the predicted length of the EO starting at the end of the learning sample (i.e., a good overlap of the ranges of the actual and predicted EO lengths and a relatively strong correlation between them).

5. Real-life case studies

In the present chapter we test the applicability of the PL method in determining the limits of our understanding of the dynamics of the real-life data (i.e., their EO). In finding the optimal parameters of the PL method we draw on the insights of the previous chapter.

As examples we chose two datasets reflecting the dynamics of two processes of fundamental importance for our understanding of the impact of humans on the climate: namely the anthropogenic CO_2 emissions and the increase of CO_2 concentration in the atmosphere. Knowledge about the dynamics of these processes is also necessary to run integrated assessment models (IAMs, such as IMAGE[45]). Hence, estimation of the temporal limits of our understanding of these dynamics may also shed some light on the time horizons beyond which projections of the abovementioned IAMs may be unreliable.

The datasets we use contain the annual global CO_2 emissions from the technosphere[46] (i.e., from fossil fuel burning and cement production) and the annual average concentration of CO_2 in the atmosphere measured at the Mauna Loa station[47]. As CO_2 concentrations are influenced by anthropogenic CO_2 emissions the analyzed datasets cover the same period: 1959 – 2011.

5.1. Global CO2 emissions from technosphere

In case of anthropogenic CO_2 emissions, the best performance is achieved for the 1st order PL method with LBs of length of 25 points (which is roughly half the size of the learning sample). This is consistent with our observations from the experiments on synthetic data—for them the 1st order PL method was also the best choice. The optimal length of the LB was chosen according to the guidelines provided at the end of the previous chapter. Exemplary stages of the optimal PL procedure are presented on Figure 36. As one can see, the data follow a roughly linear trend[48], although three segments of slightly different slopes can be seen. These segments are of similar lengths to the LBs used in the learning procedure. Hence, two types of configurations of the LB with respect to the abovementioned segments are possible—and each of these constellations has a negative impact on the length of the EO. If the LB strongly overlaps with one of these segments, then the linear model describes the data in the learning data well. However, the EO representing the expected future behavior of emissions is then compared against the data in the TB which follows a different regime (i.e., an increase of a different slope) to the data in the LB. As a consequence, the EO is relatively short. The other possibility is that the moment of regime change lies well within the LB. This renders the linear model less suitable to represent the data behavior within the LB and thus in the increase of autocorrelation of model residuals. Such a strong violation of the PL method assumptions results in a shorter EO. Analysis of both actual and predicted lengths of the EOs for different stages of the 1st order learning procedure—cf. Figure 36—confirms these

[45] For brief synopsis of the IMAGE model see e.g., http://unfccc.int/adaptation/nairobi_work_programme/knowledge_resources_and_publications/items/5396.php

[46] Source: CDIAC http://cdiac.ornl.gov/trends/emis/overview_2011.html

[47] Source: NOAA http://www.esrl.noaa.gov/gmd/ccgg/trends/full.html

[48] Taking a broader perspective the overall trend in CO_2 emissions over the last 200 years is approximately exponential, but the steep growth over the last six decades alone is roughly linear.

observations. It shows that, in principle, one should not expect the EO to be much longer than about five points[49], while very short EOs for some of the stages of the learning procedure indicate that the analyzed process occasionally undergoes sudden regime changes.

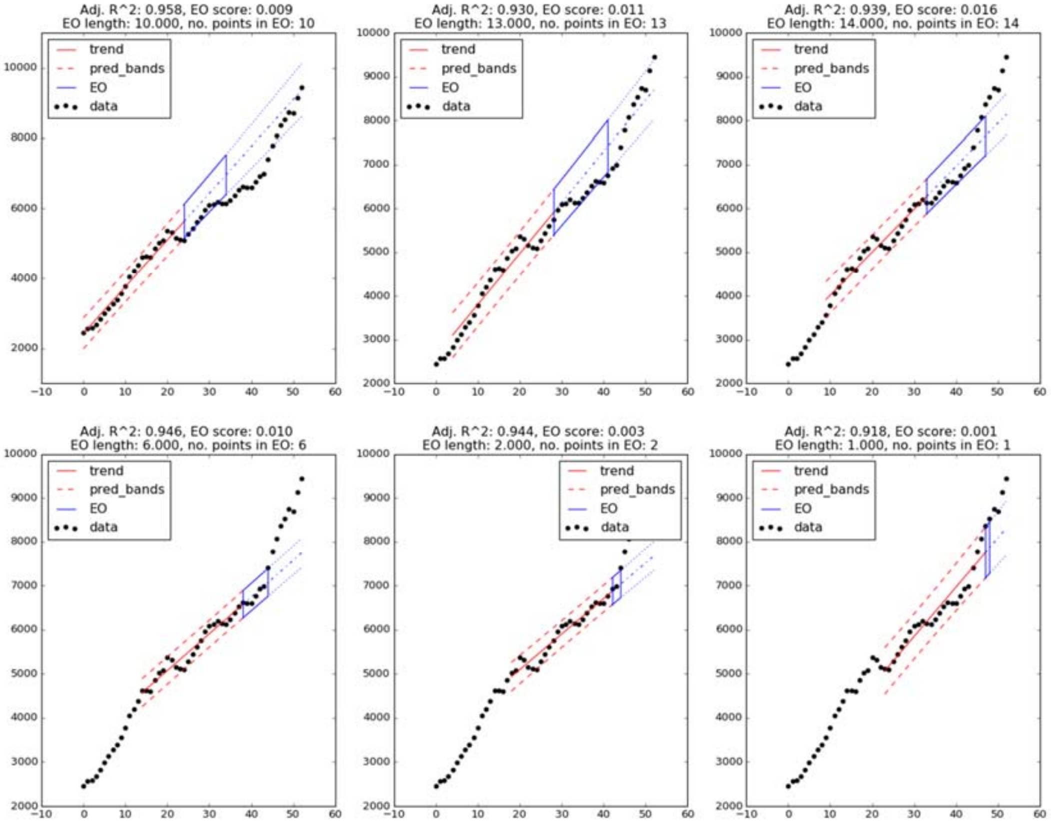


Figure 36. Six exemplary stages of the 1st order PL procedure with a LB length of 25 points.

[49] Note that the EOs are shorter than the used learning blocks. This is in agreement with what we have observed for the synthetic datasets (c.f. Chapter 4).

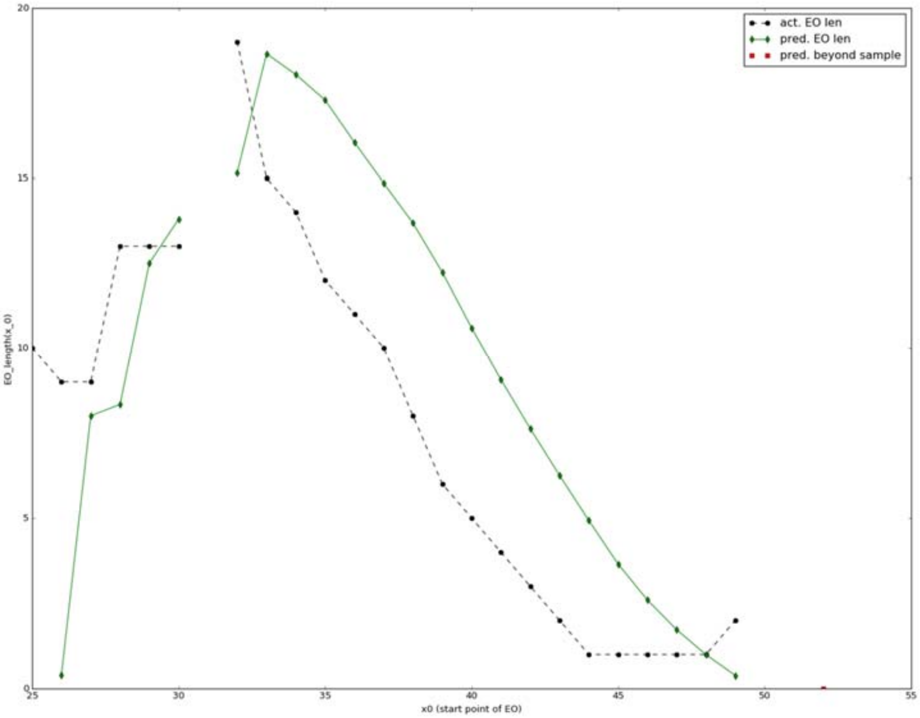


Figure 37. Actual (black dots) and predicted (green diamonds) EO lengths for all stages of the 1st order PL procedure with a LB length of 25 points. Correlation between the actual and predicted EO lengths is 0.777. The red square marks the predicted length of the EO starting at the end of testing sample. Prediction is based on all finite actual EO lengths calculated in the learning procedure (i.e., all of the black dots). Note that all of the EO lengths (both actual and predicted) are shorter than the length of the LB.

Higher order PL procedures do not yield better results. As they require longer LBs, at each stage of the PL procedure the LB contains the moment of regime (slope of local trend) change. Although polynomial trends are more flexible than the linear trend, they too are unable grasp slight but sudden regime changes—as demonstrated on the example of the 2nd order PL method (cf. Figure 38). As a result, the EOs constructed with use of the 2nd order method are only wider (since prediction bands for higher order polynomial regression diverge more rapidly than for linear case) but not longer—see Figure 39.

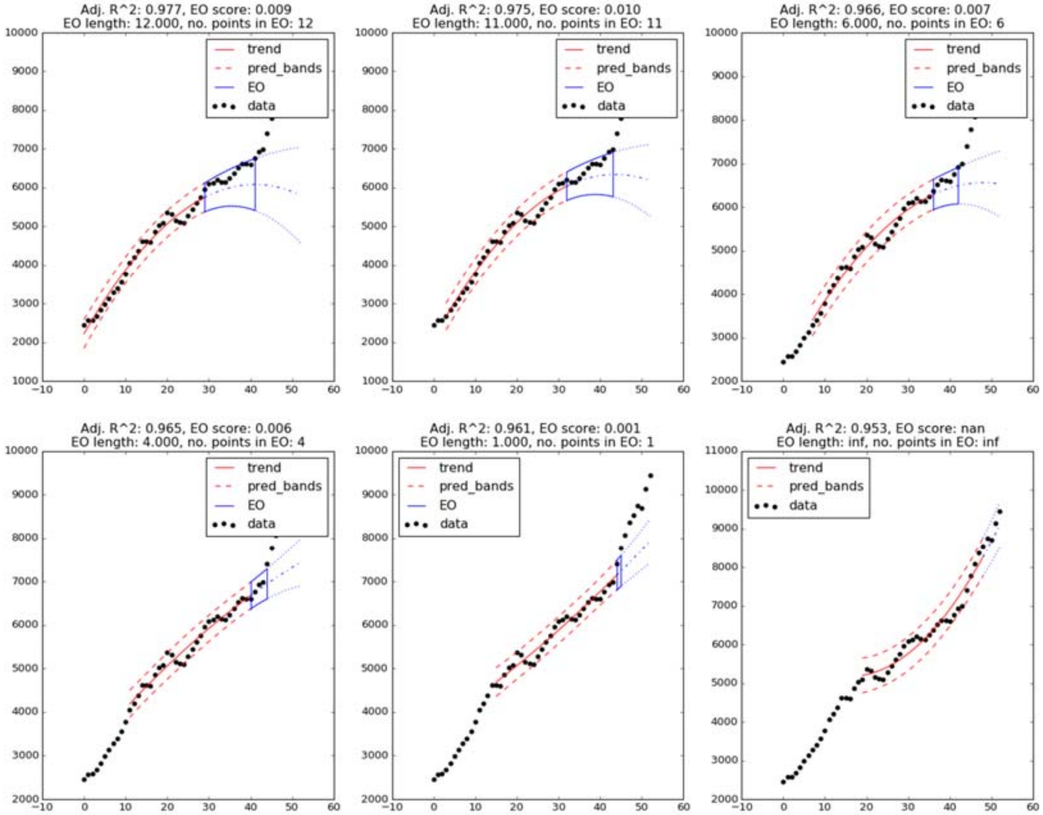


Figure 38. Six exemplary stages of the 2nd order PL procedure with a LB length of 30 points.

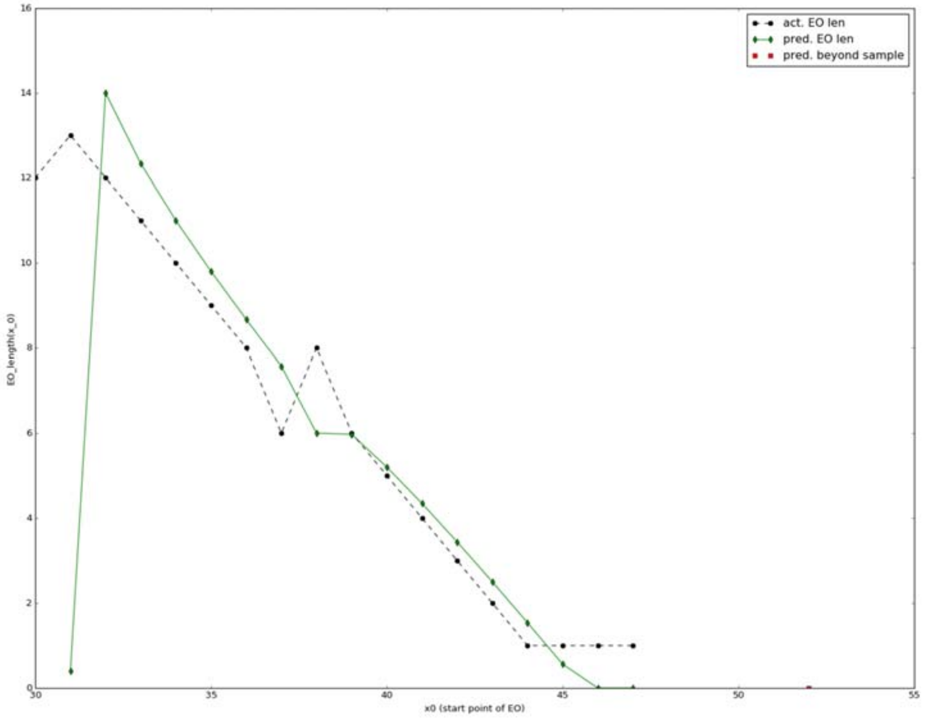


Figure 39. Actual (black dots) and predicted (green diamonds) EO lengths for all stages of the 2nd order PL procedure with a LB length of 25 points. Correlation between the actual and predicted EO lengths is 0.713. The red square marks the predicted length of the EO starting at the end of testing sample. The prediction is based on all finite actual EO lengths calculated in the learning procedure (i.e., all of the black dots). Note that all of the EO lengths (both actual and predicted) are not longer than the length of the LB.

5.2. Concentration of CO_2 in the atmosphere

Time evolution of the CO_2 concentrations over time is smooth (in comparison to that of anthropogenic CO_2 emissions) and follows a clear, exponential–like deterministic trend. The analyzed sample resembles the synthetic data with a low level of noise following an exponential trend which we analyzed in Chapter 4. Similarly to that case, the 1st order PL method proves to be the best choice among the PL methods based on polynomial regressions. The optimal length of the LB in this case is 20 points. As one can see in Figure 40, the EOs constructed using this method are narrow (because of the low variance of the residuals for the linear models fitted to the LBs) but relatively short. Indeed, for most of the PL procedure stages the EOs are not longer than three points (cf. Figure 41). This is caused by the curvature of the trend in the data.

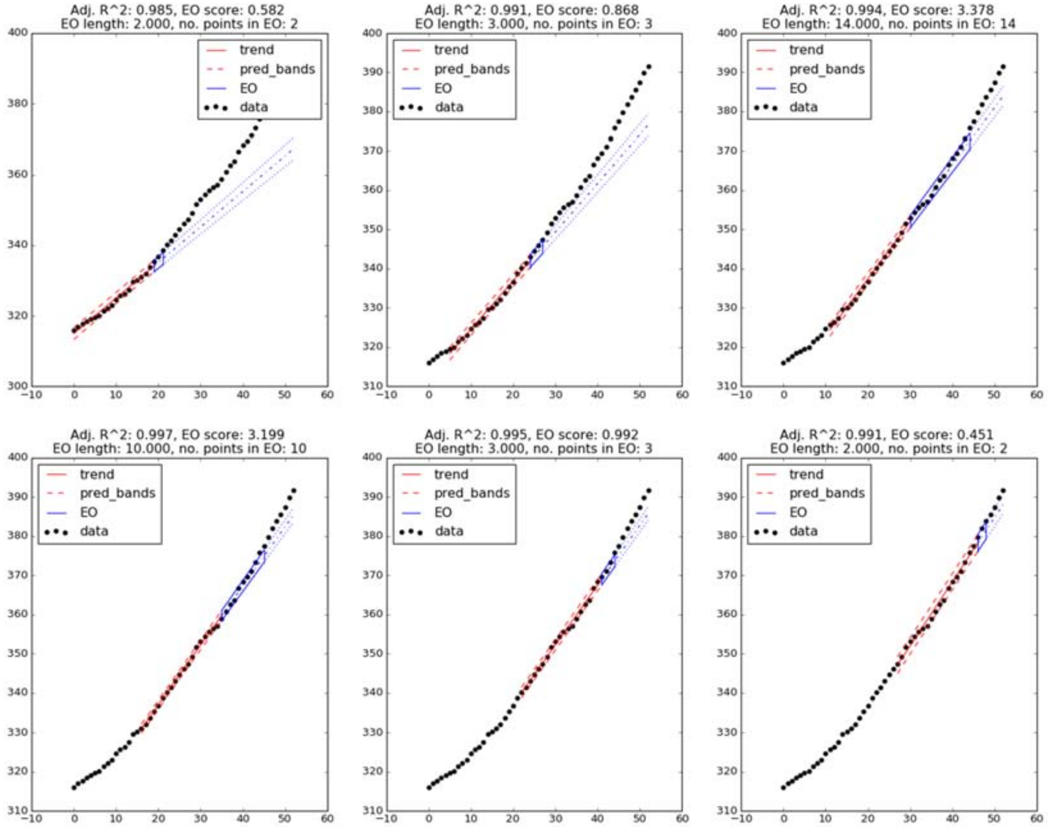


Figure 40. Six exemplary stages of the 1st order PL procedure with a LB length of 20 points.

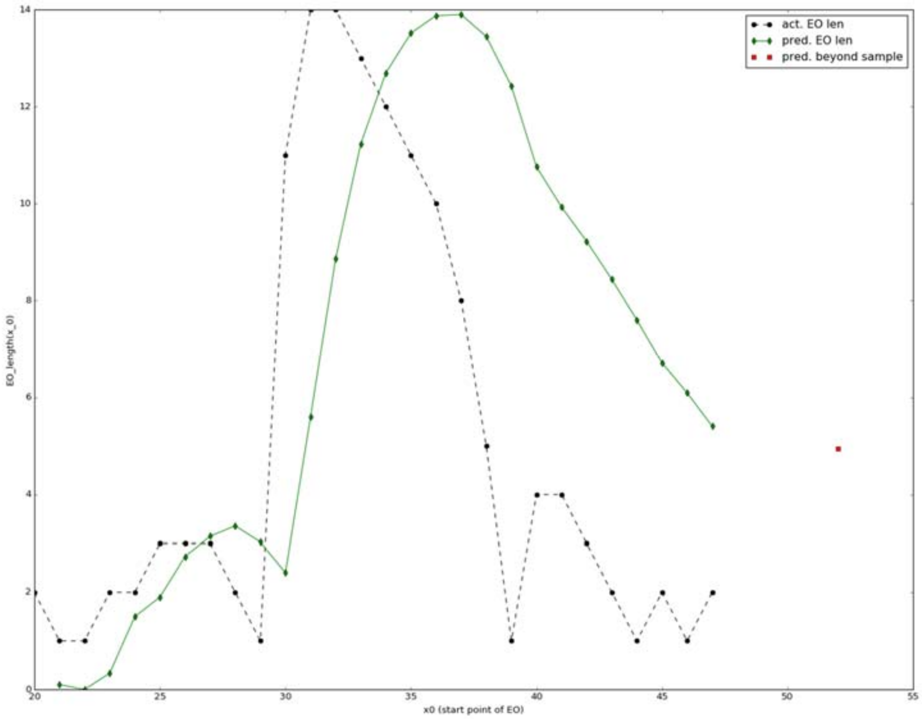


Figure 41. Actual (black dots) and predicted (green diamonds) EO lengths for all stages of the 1st order PL procedure with a LB length of 20 points. Correlation between the actual and predicted EO lengths is 0.461. The red square marks the predicted length of the EO starting at the end of testing sample. The prediction is based on all finite actual EO lengths calculated in the learning procedure (i.e., all of the black dots). Note that all of the EO lengths (both actual and predicted) are not longer than the length of the LB.

Quadratic trends are more suitable to approximate data following a curved trend (cf. Figure 42). However, in case of atmospheric CO_2 concentrations, applying the 2nd order method does not result in a longer EO. Indeed, although EOs constructed around a quadratic trend have a curved shape and are narrower than those for the 1st order method, they are still unable to follow the true trend in the long run (see Figure 43).

Applying the 3rd (or higher) order PL method to the data is not feasible, as the minimal length of LB for those methods is comparable to the size of the whole learning sample.

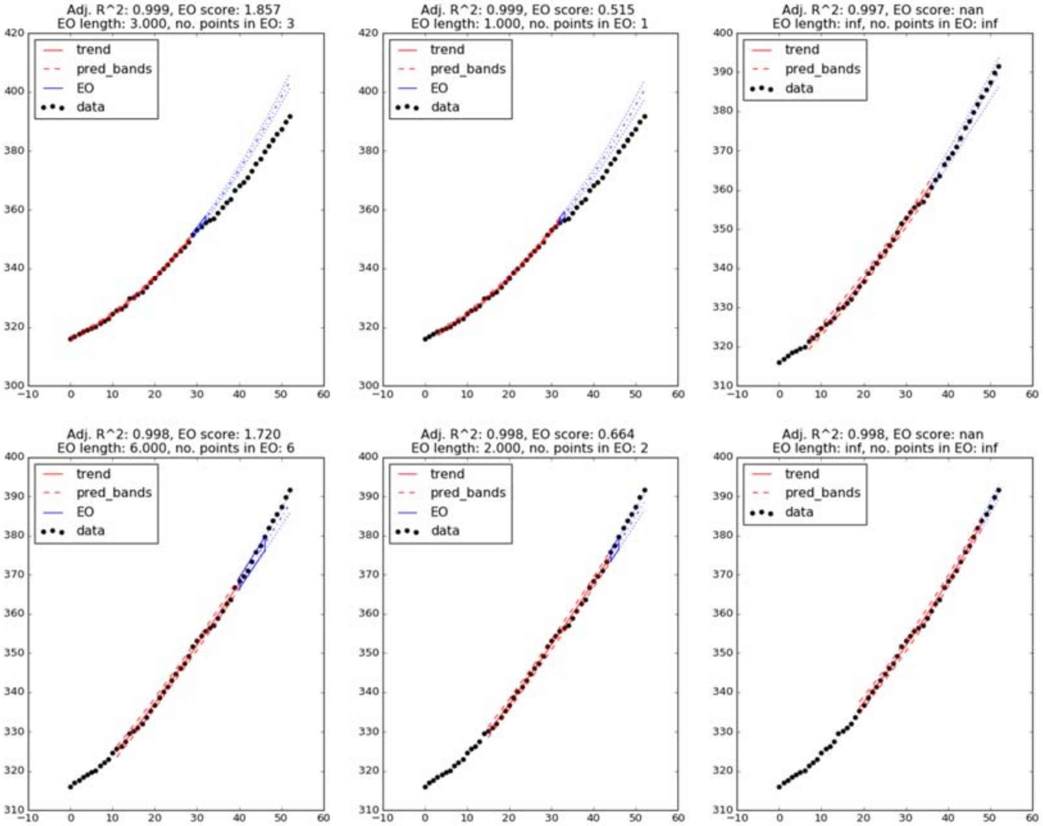


Figure 42. Six exemplary stages of the 2nd order PL procedure with a LB length of 30 points.

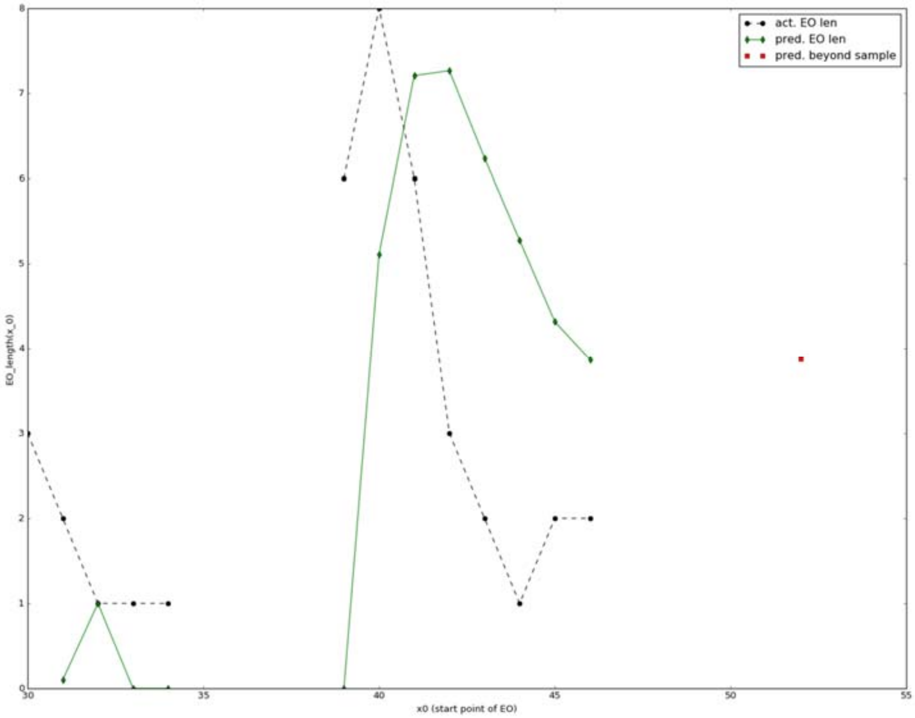


Figure 43. Actual (black dots) and predicted (green diamonds) EO lengths for all stages of the 2nd order PL procedure with a LB length of 50 points. Correlation between the actual and predicted EO lengths is 0.302. The red square marks the predicted length of the EO starting at the end of testing sample. The prediction is based on all finite actual EO lengths calculated in the learning procedure (i.e., all of the black dots). Note that all of the EO lengths (both actual and predicted) are shorter than the length of the LB.

5.3. Conclusions

The temporal dynamics of both considered processes (i.e., anthropogenic CO_2 emissions and CO_2 concentrations in the atmosphere) are essentially nonlinear. The typical time horizons within which linear predictions of the behavior of upcoming data are credible is indicated by the lengths of the EOs obtained by applying the 1st order PL method. These limits for credible linear predictions are rather short.

For anthropogenic CO_2 emissions it is at most 15 points (years), but linear predictions for the immediate future are expected to be credible over a much shorter time horizon. This is due to the fact that the linear regression model employed in the learning procedure is not able to describe or anticipate regime changes (i.e., sudden changes of slope).

The more regular behavior of the atmospheric CO_2 concentrations results in slightly better, yet still short, horizons for credible linear approximation of process dynamics — the typical length of the EOs for the 1st order PL method is 2 to 6 points (years).

Approximations of the local dynamics of the considered processes by polynomial regression functions of higher orders are better in comparison to linear ones. However, predictions made by extrapolations of such trends are more uncertain, and thus it is often impossible to assess their credibility by means of EO.

Finally, it is important to emphasize that the limits of credibility assessed by means of the 1st order PL method should be treated as the lower bound for the period within which our understanding of the system's past may be used for making reliable predictions. In principle, there may be a more suitable method than polynomial regression to explain data behavior. A PL procedure based on such a method would most likely yield better (i.e., longer but still relatively narrow) EOs, thus improving the lower bounds for the horizons of credibility.

6. Outlook

The research presented in this report is a feasibility study based on the notions of prognostic learning and explainable outreach of the data. As such, it pursues the two objectives: (1) to frame the idea of the PL and place it in a broad context of Earth system sciences; and (2) to develop and implement a PL procedure allowing us to test the PL concept in practice.

For the first objective we have restricted ourselves to analyzing data forming a time series and describing the temporal evolution of the analyzed system. Our focus was on detecting the system's dynamics (i.e., the deterministic part of the analyzed time series) represented by the prevailing trend and on understanding the relationship between the uncertainty of the estimates of this trend and the credibility of our projections based on this trend about the future system's behavior.

Understanding the temporal dynamics of the system and indicating the extent of credible predictions based on this understanding is just a first step in development of the paradigm of learning in a controlled prognostic context. However, the proposed PL method concentrates on grasping the temporal dynamics revealed by a single time series (using the time as the only explanatory variable) while hiding the explicit dependence of the system on external forcing. For example, anthropogenic CO_2 emissions exhibit roughly linear temporal dynamics over the last five decades (cf. Section 5), but they also strongly depend on the trends and disturbances of the global economy (such as the energy crises in the 1970s, the economic collapse of the soviet bloc in the 1990s or increased consumption in developing countries in recent years). We envisage a modification of the PL method by introducing additional explanatory variable(s) representing the external forcing of the system (in the context of anthropogenic CO_2 emissions this could be, for example, GDP) or dependence on some additional factors (e.g., carbon intensity of production processes). We speculate that explicit use of additional explanatory variables in the PL method will result in longer horizon of credible predictions (i.e., longer EOs).

Another challenge related to objective (1) is to demonstrate the ability of the PL method to support a modeling exercise by realizing the "model performance assessment" track (cf. Figure 2) for a suitably selected climate or integrated assessment model.

Pursuing objective (2) we have proposed a way of implementing the prognostic learning concept which is based on the ordinary least squares (OLS) polynomial regression technique. This regression method was selected for its simplicity and relatively good performance. However, the results presented in Sections 4.3 and 5 indicate the need for development of analogous versions of the PL method based on regressions using other parametric trends (e.g., exponential or power functions).

Moreover, we expect that the performance of the PL method based on higher order polynomials may be improved by application of the regularization techniques (Hastie

2009, Murphy 2012). In principle, regularization penalizes the trend functions which are overly "wiggly". It would allow us to strike a balance between the flexibility of the high order polynomials and the robustness of the predictions based on their extrapolations. We speculate that this would result in EOs that are longer and not too much wider than those obtained for the 1st order PL method.

Another way of improving the regression-based PL is to replace the OLS polynomial regressions with some more robust methods of fitting the trend, such as ridge regression or support vector regression (Hastie 2009, Murphy 2012) or nonparametric regressions (Wasserman 2006). Some preliminary results obtained by using the PL method based on selected nonparametric regression techniques are presented in the appendix. This research direction is particularly interesting for the following reasons: (1) nonparametric methods do not confine us to any specific class of regression functions; (2) nonparametric methods offer a promising link between the memory of the system (described by means of bandwidth parameter, which determines how many previous data points influences the present one) and the EO (defined as extrapolated prediction bands) and (3) flexibility of the nonparametric regression curve results in longer (yet equally robust) EOs than the ones obtained with OLS linear regression.

Note that the PL method presented in Chapter 3 relies heavily on assumption of independence of the points in the learning sample[50]. However, by making such assumption (which we do deliberately for the sake of simplicity) we ignore the fact that the patterns of behavior of the stochastic part (such as autocorrelation structure of residuals) may also be of a significant importance. Simply assuming that the stochastic part is just uncorrelated noise may result in underperformance of the EO[51]. In future research we plan to address this problem by modifying the construction of the EO to account for the autocorrelation structure of the data.

PL techniques discussed in this report identify the dynamics of the system of interest by means of a regression function. Yet, trend functions are not the only way of expressing patterns of data behavior. Therefore, alternative[52] approaches to learning in a controlled prognostic are conceivable. For example, the techniques of granular computing such as quantization or clusterization (Pedrycz 2013) may be employed to understand the patterns of data behavior. These techniques are based on assigning each of the data points to one member of a discrete collection of classes (called also information granules) in order to reduce the level of detail which may blur the more fundamental features of the data (which are represented by these classes). The patterns in data behavior may then be expressed as transition rules from one information granule to the other, or more broadly by transition probabilities, that is, the likelihood that an observation taken at certain time belongs to a certain information granule given the class into which the previous observation falls. This approach is currently being explored (Puchkova et al).

[50] It is required by both the OLS method of fitting a regression function to the data and by the way we determine the length of the EO (cf. Section 3.2).

[51] Recall that we decide to end the EO in the first moment for which layout of observations in period between the end of the learning block and this moment is unlikely under assumption that the extrapolated regression function fitted to the learning block is also a good approximate of the true trend in the testing block and the observations in the testing block are independent. However, if the observations were correlated then encountered layout of points might be not so unlikely and the actual EO length should be greater.

[52] I.e. alternative to the regression-based method presented in this report.

7. Summary

In this report we introduce the paradigm of learning in a controlled prognostic context. It is a data-driven, exploratory approach to assessing the limits to credibility of any expectations about the future system's behavior which are based on a time series of historical observations of the analyzed system. The aim of the proposed method is to indicate the typical length of time over which the trends in the historical data sample persist, as well as the level of uncertainty in identifying these trends.

The key idea of learning in a controlled prognostic context is to deduce directly from the data their EO, that is, the spatio-temporal extent for which, in lieu of the knowledge contained in the historical observations, we may have a justified belief contains the system's future evolution. The length of such EO indicates the time horizon within which predictions based on our current understanding of the system are credible. The initial width of the EO reflects the diagnostic uncertainty inherent to our imperfect understanding of the system, while the shape of the EO informs us about the strength of measures required to overcome the system's inertia.

We propose a method of constructing the explainable outreach based on the polynomial regression technique. The data sample is split into two parts: the LB and the TB. The dynamics of the system in the period covered by the LB is identified by means of a polynomial regression model and the EO expressing our expectations about the system's evolution beyond the LB is constructed by extrapolating the prediction bands of the fitted regression model. These prediction bands represent both our expectations about the future system's dynamic and its uncertainty. The EO is then tested against the remainder of the data (i.e., the TB) in order to indicate the time horizon within which predictions based on the fitted regression model are believed to be credible.

We also propose a PL procedure which supports (with the use of an EO score) selection of the most appropriate type of regression model to represent the system's dynamic. In addition, the PL procedure also allows us to derive an indicator of the typical length of the time interval within which predictions made using the regression model credibly match the actual future observations.

The proposed PL method was tested on various sets of synthetic data in order to identify its strengths and weaknesses, formulate guidelines for optimal selection of the method parameters (the order of the polynomial regression and the length of the LB), and check how useful the proposed construction of the EO is in informing us about the immediate future of the observed system. We also indicate how the PL method can be applied in the context of Earth system sciences applying it to analyze historical anthropogenic CO_2 emissions and atmospheric CO_2 concentrations. We conclude that the most robust of the analyzed methods is the one based on linear regression. However, the EOs obtained using this method and expressing horizons within which linear projections are credible are rather short.

8. Acronyms

EO	Explainable outreach
GHG	Greenhouse gases
LB	Learning block (part of the learning sample to which regression model is fitted)
OLS	Ordinary least squares method of fitting a regression function to the data
PL	Learning in a controlled prognostic context (prognostic learning for short)
TB	testing block (part of the learning sample used to test the EO in order to determine its length)
TSA	Time series analysis (statistical techniques of analysis of time series)

9. Literature

Brockwell, P.J., Davis, R.A. (2002): Introduction to Time series and Forecasting, Second Edition. Springer, ISBN 0-387-95351-5

Hastie, T., Tibshirani, R., Friedman, J. (2009): The Elements of Statistical Learning. Data Mining, Inference and Prediction. Springer, ISBN 978-0-387-84858-7

IPCC (2007: FAQ 1.2): What is the Relationship between Climate Change and Weather? In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 104–105.

IPCC (2007: FAQ 8.1): How Reliable Are the Models used to Make Projections of Future Climate Change? In: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change* [S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor and H.L. Miller (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 600–601.

IPCC (2013: Box 11.1): Climate Simulation, Projection, Predictability and Prediction. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [T.F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 959–961.

Meinshausen M., N. Meinshausen, W. Hare, S.C.B. Raper, K. Frieler, R. Knutti, D.J. Frame, M.R. Allen (2009): Greenhouse-gas emission targets for limiting global warming to 2 °C. *Nature*, **458**(7242), 1158–1162; doi: 10.1038/nature08017.

Murphy, K.P. (2012): Machine learning. A Probabilistic Perspective. MIT press, ISBN: 9780262018029

NSF (2012): Decadal and Regional Climate Prediction using Earth System Models (EaSM). National Science Foundation, Arlington VA, USA; Solicitation: http://www.nsf.gov/pubs/2012/nsf12522/nsf12522.pdf; FAQs: http://www.nsf.gov/pubs/2012/nsf12029/nsf12029.jsp.

Otto, F.E.L., C.A.T. Ferro, T.E. Fricker and E.B. Suckling (2015): On judging the credibility of climate predictions. *Clim. Change*, **132**(1–2), 47–60, doi 10.1007/s10584-013-0813-5.

Pedrycz, W. (2013): Granular computing. Analysis and design of Intelligent systems, CRC press, ISBN 9781439886816.

Puchkova, A., A. Kryazhimskiy, E. Rovenskaya, M. Jonas and P. Żebrowski (2016): Cells (working title). Manuscript, International Institute for Applied Systems Analysis, Laxenburg, Austria (Manuscript under preparation for submission to a scientific journal).

Wolberg, J. (2006): Data Analysis Using the Method of Least Squares. Springer, ISBN 978-3-540-31720-3

Wasserman, L. (2006): All of Nonparametric Statistics, Springer, ISBN 978-0-387-30623-0

Appendix: Nonparametric kernel-based regression

Nonparametric regression is an alternative to conventional parametric methods. It can be used when we do not want to be limited to the predetermined form of the estimated regression function; when we need to relax some assumptions from the regression analysis while maintaining a good estimate; or simply when the nature of the data analysed does not allow for selection of a reasonable model.

To a rich family of nonparametric regression methods (Wasserman 2006, Härdle 1990, Fan 1992, Green & Silverman 1994, Györfi et al. 2002) belong for example, local averaging, regression and smoothing splines (Rice & Rosenblatt 1981, Rice & Rosenblatt 1983, Stone 1994, Eubank 1999), wavelets (Nason 1996, Johnstone & Silverman 1997, Wang 1996), or orthogonal series (Green & Silverman 1994). However, the *kernel estimation* is especially noteworthy. It belongs to popular smoothing techniques (Simonoff 1996, Silverman 1986) that allow for estimation even in the case of complicated relationships between explanatory and response variables.

This appendix is dedicated to the application of the prognostic learning method to nonparametric kernel-based regression in real-life case studies from Chapter 5:

(1) Global CO_2 emissions from technosphere.

(2) Concentration of the CO_2 in the atmosphere.

A.1 Kernel functions

The kernel estimation (see e.g., Wasserman 2006, Green & Silverman 1994, Hart 1991), is an extension of local averaging and involves the use of the so-called *kernel function K*, being nonnegative, symmetric, square integrable, and satisfying the conditions

$$\int_{-\infty}^{+\infty} K(t)dt = 1, \quad \int_{-\infty}^{+\infty} tK(t)dt = 0, \text{ and } \int_{-\infty}^{+\infty} t^2K(t)\,dt < \infty.$$

Given these characteristics the specific choice of a kernel function is not of critical importance. One can take any symmetric probability density function (PDF) of a continuous random variable with zero mean and finite variance[53].

The most popular choices of kernel functions (Figure A.1) are the *Gaussian* (normal) *kernel* (i.e. PDF of the standard normal distribution), and a few kernels with compact support, like rectangular (uniform), tricube, or the Epanechnikov kernel.

[53] The choice of the kernel *K* may slightly affect the asymptotic properties of the kernel estimator. For results in finite samples, the difference is negligible.

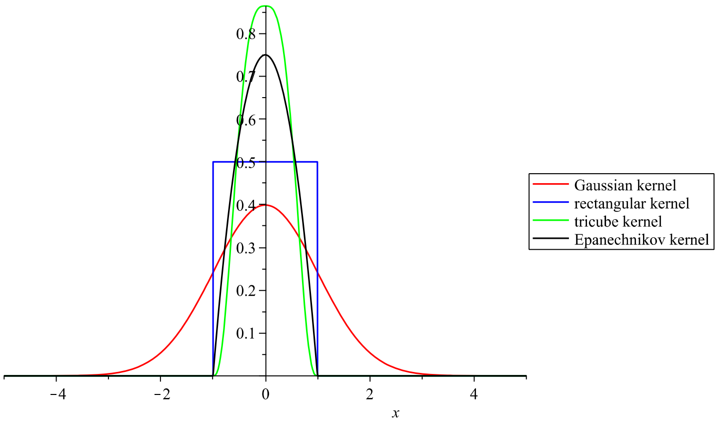


Figure A.1. Four most popular kernel functions.

A.2 Kernel-based regression methods

Kernel regression has been known for many years and various *kernel estimators* (KE) have been used. The most important (see Table A.1 for overview) are:

- Nadaraya-Watson KE (Nadaraya 1964, Watson 1964),
- k-nearest neighbours KE and its modifications (Wasserman 2006),
- Priestley-Chao KE (Priestley & Chao 1972),
- Gasser-Müller KE (Gasser & Müller 1984),
- Local polynomial regression, in particular local linear KE (Li & Racine 2004, Ruppert & Wand 1994, Fan & Gijbels 1997).

Some of them have also been considered and analysed in the case of *time series data* or correlated errors (see e.g., Hart 1991, Opsomer et al. 2001, Altman 1990). In Section A.4 two kernel estimators are used: the Nadaraya-Watson KE (NWKE)—mostly because of its simplicity in applications, and the local linear KE (LLKE)—because of its properties and good results, even for small samples.

Each of the aforementioned KEs (except the local polynomial KE) can be considered a *linear smoother* of the form

$$\hat{r}(x) = \sum_{i=1}^{n} l_i(x) Y_i \qquad (A.1)$$

where $(x_{1,}Y_1), (x_{2,}Y_2), \dots, (x_{n,}Y_n)$, denote the bivariate data, corresponding to continuous random variables *x* and *Y*,

$$\widehat{Y_\iota} = \hat{r}(x_i) + \varepsilon_i, \;\; i = 1,2, \dots n,$$

and residuals ε_i, *i=1,2,…n,* are assumed to be independent[54] and normally distributed, with zero mean and standard deviation $\sigma > 0$.[55]

Functions $l_i(x)$, *i=1,2,…n,* satisfy condition

$$\sum_{i=1}^{n} l_i(x) = 1$$

and take various forms, depending on the estimator considered (Table A.1).

Table A.1. Overview of the most popular kernel regression estimators. The methods used in this appendix are marked in green.

KE	l_i(x) in *(A.1)*	Properties & Remarks
Nadaraya-Watson (NWKE)	$l_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{h}\right)}$	- local constant estimator - can be adopted for (discrete) time series case - several 'rules of thumb' for selection of bandwidth *h* - biased (design bias and strong boundary bias) - requires large samples
k-nearest neighbours (weighted) (k-NNKE)	$l_i(x) = \frac{K\left(\frac{x-x_i}{R}\right)}{\frac{1}{n}\sum_{j=1}^{n} K\left(\frac{x-x_j}{R}\right)}$ where *R* denotes the distance between *x* and its k-nearest neighbour;	- for rectangular kernel, it reduces to NWKE - $k = 2nhf(x)$, where *f* denotes the PDF of the explanatory variable -biased (both design and boundary bias) - various modifications and simplifications; various weights - require large samples
Priestley-Chao (PCKE)	$l_i(x) = \frac{x_i - x_{i-1}}{h} K\left(\frac{x-x_i}{h}\right)$	- applicable to compactly supported data (rescaling option, with good results) - requires kernel function with compact support - no design bias, but strong boundary bias - requires large samples
Gasser-Müller (GMKE)	$l_i(x) = \frac{1}{h} \int_{v_{i-1}}^{v_i} K\left(\frac{x-u}{h}\right) du$ where $x_i \leq v_i \leq x_{i+1}$	- continuous version of PCKE - partition $\{v_i\}$, *i=1,..n-1* required - applicable to compactly supported data (rescaling option with good results) - requires kernel function with compact support - no design bias, but boundary bias - requires large samples

[54] For some kernel-based methods the independence assumption can be relaxed, especially when applying KE to time series data (Section A.3).

[55] In general, standard deviation σ does not need to be constant. Sometimes $\sigma(x) > 0$, is considered instead.

Local linear (LLKE)	$l_i(x) = \frac{b_i(x)}{\sum_{j=1}^{n} b_j(x)}$, where $b_i(x) = K\left(\frac{x-x_i}{h}\right)(S_{n,2}(x) - (x_i - x)S_{n,1,}(x))$ $S_{n,j}\,(x) = \sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)(x_i - x)^{j,}$	- particular case of local polynomial regression - local linear smoother - can be adopted for (discrete) time series cases - no boundary nor design bias - requires large samples, although thanks to good local fit, better results for smaller samples
Local polynomial KE	Estimate locally (at a point *x*) that polynomial of degree *p*, which approximates $\hat{r}\,(x)$ in a small neighbourhood of the point *x*, in the best way.	- becomes NWKE for *p=0*, and LLKE for *p=1* - in general cannot be represented as a linear smoother given by (A.1) - no boundary nor design bias - require large samples, although thanks to good local fit, reasonable results for smaller samples; - for larger *p* requires larger samples

A.2.1 The problem with bandwidth selection

Weights $l_i(x)$, *i=1,…,n,* in formula (A.1) depend on kernel function *K,* and a *smoothing parameter* $h > 0$ (also called a *bandwidth*) [56], such that

$$h \to 0 \text{ but } nh \to \infty, \qquad \text{as } n \to \infty.$$

The choice of optimal value for the smoothing parameter is crucial[57] and corresponds to a problem of finding the "golden mean", by minimizing the *mean squared error* (MSE), being the sum of squared bias[58] and sampling variance

$$MSE(\hat{r}(x)) = bias(\hat{r}(x))^2 + Var(\hat{r}(x)),$$

or its asymptotic and integrated versions.

The bandwidth parameter h controls the smoothness of estimated regression function. Larger h results in a smoother curve, but sometimes with a worse fit and hence a larger variance. Smaller h in turn means a better fit, with smaller variance, it may, however, cause a greater bias (see Figure A.2). A h that is too large therefore means *oversmoothing* (possibly failing to reflect the character of the data analysed), while too small leads to *undersmoothing.*

[56] There are also methods involving variable bandwidths. Here, we focus on methods with fixed bandwidth.

[57] See e.g. (Wasserman 2006), (Simonoff 1996), etc.

[58] $bias\ (\hat{r}(x)) = E(\hat{r}(x)) - \hat{r}(x)$

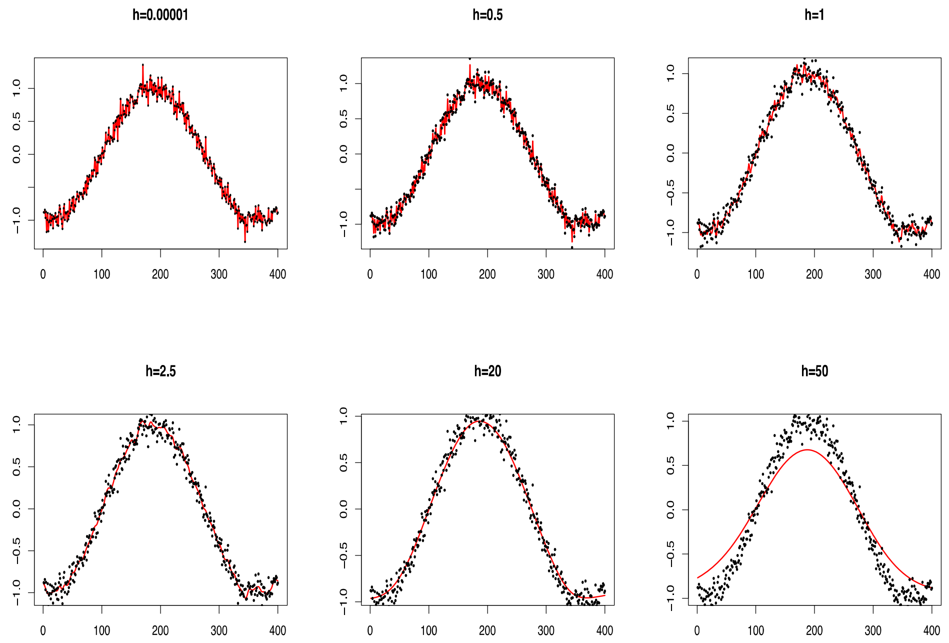


Figure A.2. Varying the smoothing parameter: examples of the NWKEs fitted to the data following sinusoidal trend (from Section 4.3.5) given by $g(x) = \sin(0.018 \times (x - 100))$, with standard deviation of noise $\sigma = 0.01 \times (\max g - \min g)$, where n=*400*, for various values of h, and using the Gaussian kernel.

The shape of $\hat{r}(x)$ changes for various values of h. The plots in the first row illustrate what happens when the smoothing parameter is too small. The variance in that case is very small, which results in a good fit, but it is at the price of an undersmoothed and strongly fluctuating regression curve. The sample is relatively large (n=*400*), so the 'noisy' shape of the estimator is caused by overfit. Increasing h gives a smoother $\hat{r}(x)$, as can be seen for h=*2.5* and *20*. The plot in the lower right corner of Figure A.2 illustrates the evident underfit (resulting in large variance)—the curve is oversmoothed and does not grasp the behaviour of the data.

It is worth noting that, despite the problem with bandwidth selection, even the simple NWKE approximates the regression function fairly well. Despite the almost 10-fold difference between the values of h, the two figures at the bottom left look satisfactory. To assess which of them really performs better, one can look at confidence or prediction intervals (the latter works better in this regard, because of more emphasis on the standard error).

Since the degree of smoothing corresponds to the variance of $\hat{r}(x)$, it also affects the width of prediction intervals[59]. Oversmoothing leads to intervals that are too wide (interpreted as large uncertainty of results), while undersmoothing means the intervals are too narrow (Figure A.3).

[59] see Section A.2.2

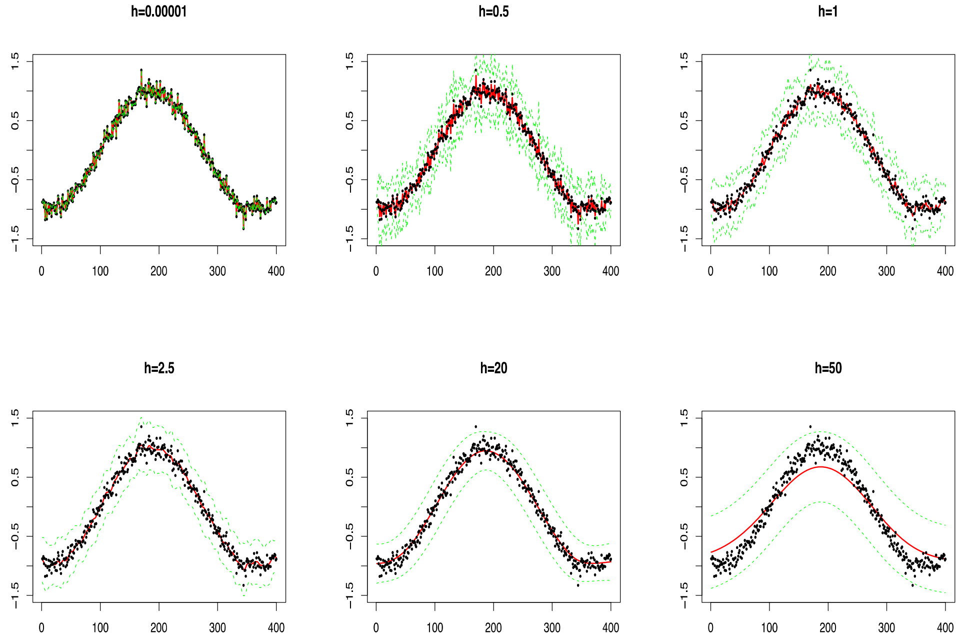


Figure A.3. Varying the smoothing parameter and illustrating its impact on 95% prediction intervals (green dashed lines): examples of the NWKEs fitted (red solid lines) to the data following a sinusoidal trend (from Section 4.3.5) given by $g(x) = \sin(0.018 \times (x - 100))$, with a standard deviation of noise $\sigma = 0.01 \times (\max g - \min g)$, where *n=400*, for various values of h, and using a Gaussian kernel.

In general, h depends on the sample size n, and asymptotically $h \propto n^{-\frac{1}{5}}$. The formulas for optimal h are different for different kernel methods. For instance, the optimal value of the smoothing parameter[60] in the case of the NWKE satisfies the following formula[61]

$$h = \left(\frac{\sigma^2 \int_{-\infty}^{+\infty} K^2(x)dx \int_{-\infty}^{+\infty} f(x)^{-1}dx}{n \int_{-\infty}^{+\infty} x^2 K(x)dx \int_{-\infty}^{+\infty} \left(\hat{r}''(x) + \hat{r}'(x)\frac{f'(x)}{f(x)} \right)^2 dx} \right)^{\frac{1}{5}} \qquad (A.2)$$

while for the LLKE[62]

$$h = \left(\frac{\sigma^2 \int_{-\infty}^{+\infty} K^2(x)dx \int_{-\infty}^{+\infty} f(x)^{-1}dx}{n \int_{-\infty}^{+\infty} x^2 K(x)dx \int_{-\infty}^{+\infty} \left(\hat{r}''(x) \right)^2 dx} \right)^{\frac{1}{5}}. \qquad (A.3)$$

[60] See e.g. (Wasserman 2006), (Green & Silverman 1994), etc.

[61] The term $\hat{r}'(x)\frac{f'(x)}{f(x)}$ in (A.2) denotes the *design bias*, typical for the NWKE (it is not present for the LLKE).

[62] See e.g. (Ruppert & Wand 1994), (Fan & Gijbels 1997), etc.

The values $\int_{-\infty}^{+\infty} K^2(x)dx$ and $\int_{-\infty}^{+\infty} x^2K(x)dx$ depend on the kernel used. For the Gaussian kernel $\int_{-\infty}^{+\infty} K^2(x)dx \cong 0.28$, while the latter one represents the variance of the standard normal distribution, i.e. $\int_{-\infty}^{+\infty} x^2K(x)dx$=1. But formulas (A.2) and (A.3) also involve unknown regression function $\hat{r}(x)$, that needs to be estimated, unknown variance σ^2, as well as $f(x)$, that is, the PDF of the explanatory variable. The methods to estimate them depend on problem requirements, the data to be analysed, and on the KE considered. In particular, for the LLKE or the GMKE, σ^2 can be estimated by an (asymptotically unbiased) estimator of the form (Gajek & Kaluszka 1993)

$$\widehat{\sigma}^2 = \frac{1}{6(n-2)} \sum_{i=1}^{n-2} (Y_{i+2} - 2Y_{i+1} + Y_i)^2$$

For the NWKE, the much simpler

$$\widehat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} (Y_{i+1} - Y_i)^2$$

can also be used. However, both formulas work well mostly for large samples.

The density function of the explanatory variable can be estimated using nonparametric methods, like kernel density estimation (Silverman 1986), or (less often) applying parametric methods (e.g., MLE, provided that, we have additional information on that variable and its distribution). In complicated cases, semiparametric methods can also be used (e.g., Jarnicka 2009). To estimate $\hat{r}''(x)$ and $\int_{-\infty}^{+\infty}\left(\hat{r}''(x)\right)^2 dx$ additional information on the data is required, since the latter one corresponds to the curvature of the estimated regression curve, or approximation by the curvature of some known curve can be used. Similarly, the term $\hat{r}'(x)\frac{f'(x)}{f(x)}$, which is responsible for the bias.

For some estimators, like the NWKE, there are a few 'rules of thumb' for finding reasonable value of h, which work well in most cases, especially for large samples (but are less useful when applied to time series data or in the case of correlated errors). Moreover, the smoothing parameter can also be chosen by the *cross-validation* (CV) criterion[63]

$$CV(h) = \sum_{i=1}^{n} (Y_i - \hat{r}(x_i))^2 \, \theta\big(z(x_i)\big),$$

where

$$z(x_i) = \frac{K(0)}{\sum_{j=1, j\neq i}^{n} K\left(\frac{x_i - x_j}{h}\right)}.$$

The penalizing function $\theta(\cdot)$ takes various forms, e.g., $\theta(z) = \frac{1}{(1-z)^2}$, (generalized CV), or $\theta(z) = e^{2z}$ (AIC – Akaike's Information Criterion), and ensures various properties (e.g., stipulating small bias or low variance)[64]. The values of h obtained using the CV criteria are usually close to the MSE-optimal ones. The problem starts with a violation of the assumption of independence of the residuals, as correlation may decrease the bandwidth

[63] See e.g. (Wasserman 2006), etc.

[64] This refers to finite samples, as they all guarantee the same asymptotic properties.

indicated by the CV criterion, so the curve obtained is undersmoothed (Opsomer et al. 2001, De Brabanter et at 2011).

A.2.2 100%(1-α)-Prediction Intervals

Choosing the right bandwidth h is of great importance for the expected estimation result. Since this choice compromises between maximizing the variation of the KE and its bias, it depends on a particular application which one of these two is more important and should be emphasized by h. In this report, we focus primarily on the variance which determines the prediction intervals (analysing it, but not trying to make it as small as possible, as this may affect the EO). According to the Central Limit Theorem (CLT), regression estimates $\hat{r}(x)$ in (A.1) have an asymptotic normal distribution

$$\frac{\hat{r}(x) - bias(\hat{r}(x))}{\sqrt{Var(\hat{r}(x))}} \to N(0,1)$$

Assuming no bias, the asymptotic *100%(1-α) - prediction interval* is of the form

$$\hat{r}(x) \pm z_{1-\frac{\alpha}{2}}\sqrt{Var(\hat{r}(x)) + \widehat{\sigma}(x)^2}$$

where $z_{1-\frac{\alpha}{2}}$ denotes the $(1-\frac{\alpha}{2})$th quantile of the standard normal distribution. For in-sample points, that is, for points from the LB, $\hat{r}(x)$ denotes the KE, and $\widehat{\sigma}^2(x)$ an estimate of the variance of residuals (corresponding to the standard error), while for new observations x^*, $\hat{r}(x^*)$ denotes the prediction at x^*, and $\widehat{\sigma}(x^*)^2$ prediction error. For the NWKE and the LLKE the variance is asymptotically equal

$$Var(\hat{r}(x)) \approx \frac{\widehat{\sigma}^2(x)\int_{-\infty}^{+\infty} K^2(t)dt}{nhf(x)},$$

which gives the in-sample *prediction bands (PB)* of the form

$$\hat{r}(x_i) \pm z_{1-\frac{\alpha}{2}}\sqrt{\frac{\widehat{\sigma}^2(x_i)\int_{-\infty}^{+\infty} K^2(t)dt}{nhf(x_i)} + \widehat{\sigma}(x_i)^2} \qquad (A.4)$$

and

$$\hat{r}(x^*) \pm z_{1-\frac{\alpha}{2}}\sqrt{\sum_{i=1}^{n}\frac{\widehat{\sigma}^2(x_i)\int_{-\infty}^{+\infty} K^2(t)dt}{nhf(x^*)} + \widehat{\sigma}(x^*)^2} \qquad (A.5)$$

for a new observation x^* (Green & Silverman 1994).

Formula (A.4) was used to construct the prediction intervals in Figure A.3. It is worth mentioning that the approximately optimal value of the smoothing parameter is *h=7.72,* while for the LLKE applied to the same data, *h=8.06* (see Figure A.4 for 95% in-sample PBs). Formula (A.5) will in turn be used to construct the EO in the procedure described in Section 3.2.

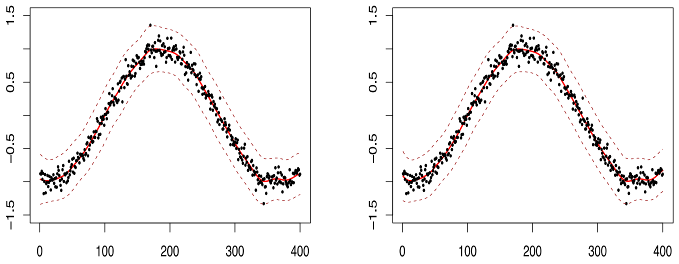


Figure A.4. 95% in-sample (LB) prediction bands (dashed) for the NWKE (left) and the LLKE (right) with the Gaussian kernel and approximately optimal bandwidths *h=7.72* and *h=8.06* for NWKE and LLKE respectively; Thanks to a large sample (LB=400 dataset from Figure A.2 and A.3) and independent observations the results are almost identical. The residual standard error is equal *0.094* and *0.093* for NWKE and LLKE respectively.

A.3 Kernel estimation of time series data

In this section we focus on *time series*, where time points are fixed and equally spaced. Following the notation from Section 3.1, let the learning block (LB) contain n observations $X_{1,}$ $X_{2,}$ …, X_n, taken at the time points $t_1, \ldots, t_n$, where $t_i = i$, *i=1,…n*.

Consider

$$\hat{x}(t) = \hat{r}(t) + \varepsilon_t,$$

where $x(t)$ = X_t is a value of the observation taken at time *t*, and the noise term ε_t is normally distributed with zero mean and standard deviation $\sigma > 0$.[65] We assume that residuals ε_t, t = 0, 1, 2, …, are correlated and their correlation decreases in inverse proportion to the distance between them[66].

[65] Assumptions on residuals, when compared to parametric regression techniques can be relaxed. Two scenarios are considered in the literature: (1) allowing non-normal distribution, but ensuring covariance stationarity and possibly weak correlation (Brabanten et al. 2011, Opsomer et al. 2001), or (2) ensuring normality and analyzing correlation structure, e.g. (Li & Li, 2009). Both lead to problems with appropriate bandwidth selection, the second one, however, allows for asymptotically better results, in particular in view of predictions and the EO.

[66] This assumption corresponds to the condition $Corr\left(\varepsilon_{t_i}, \varepsilon_{t_j}\right) = \rho(t_i - t_j)$, based on unknown stationary correlation function ρ(.). This allows for correlation decaying, when $n \to \infty$, and hence better results for large samples. We will not however be interested in analysing the correlation structure in detail, using only ‘independence-like’ approximations.

When analysing a time series, one has to deal with specific nature of the data, resulting in a need for modifications in optimal bandwidth selection methods. Moreover, the problem with applying the kernel methods to time series data is also connected to the discrete distribution of the explanatory variable *t* (discrete time), which has to be approximated by a continuous estimate.

A.3.1 Bandwidth selection in the time series case

The problem of optimal bandwidth selection, described and illustrated in Section A.2, is now more visible. The time points are equally spaced, and more importantly, the data points (and hence the residuals) are correlated, so the shape of the estimated regression function changes considerably as the smoothing parameter changes (see Figures A.5 (NWKE) and A.6 (LLKE) for examples).

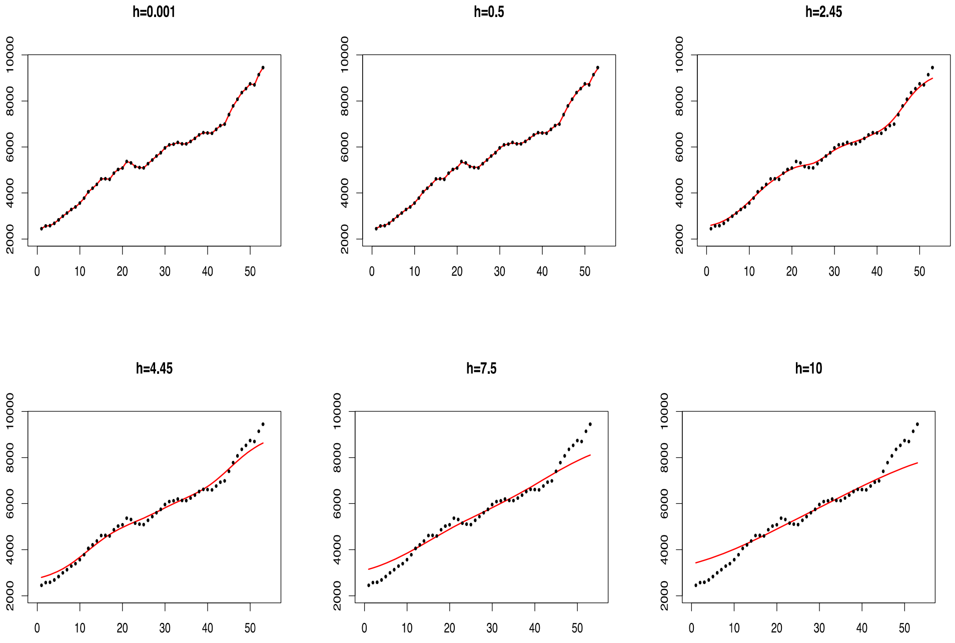


Figure A.5. Varying the smoothing parameter: examples of the NWKEs fitted to the data on global CO_2 emissions from technosphere (*n=53)* for various values of *h,* and the Gaussian kernel.

The NWKE is fitted to the data on global CO_2 emissions from technosphere. To illustrate the problems with finding the optimal bandwidth for time series, we take the whole sample, consisting of *n=53* data points and consider six exemplary values of *h*.

It is easy to see that the values *h=4.45, 7.5*, and *10* are too large, resulting in oversmoothing, which means that only the central part of the data is estimated, and the result is rather poor. On the other hand, *h=0.001* is too small, showing a perfect fit, with no visible uncertainty. Both *h=0.5* and *2.45* seem to be quite good. *h=0.5* seems to better describe the behaviour of the data. *h=2.45*, however, results in a slightly looser fit, which may be better from the EO perspective.

Note that, in four of the six examples given, we have to deal with the boundary bias, which is characteristic for the NWKE. It can significantly affect the length of the EO, since it cannot be overcome by slightly stronger smoothing, and greater variance. Therefore the LLKE is used for the EO analysis, as it is free from boundary bias. For comparison, in Figure A.6, the LLKE is fitted to the same data series, using the Gaussian kernel, and taking the same exemplary values of *h*.

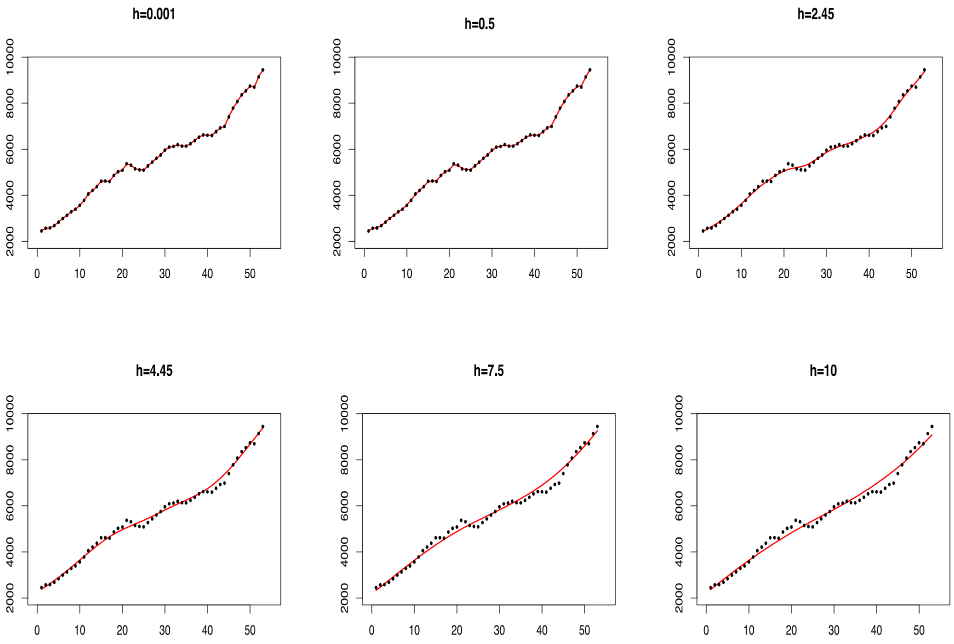


Figure A.6. Varying the smoothing parameter: examples of the LLKEs fitted to the data on global CO_2 emissions from technosphere (*n=53)* for various values of *h,* with Gaussian kernel.

It is easy to observe that the LLKE (Figure A.6) gives better results than the NWKE (Figure A.5). This is primarily related to the lack of boundary bias. Because the estimator is fitted to the data locally, even when the smoothing parameter *h* is too large (e.g. for *h=4.5* or *7.5*) the LLKE seems to properly identify the general shape of the estimated relationship.

This is also reflected in the variation of the standard error in those cases (Figure A.7), as the standard error (SE) increases much faster in the case of the NWKE.

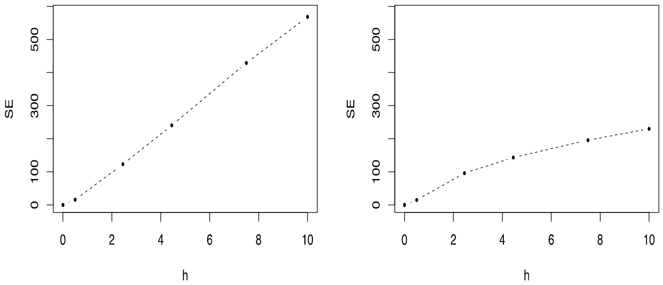


Figure A.7. The relationship between the smoothing parameter and the standard error for the NWKE (left) and the LLKE (right) considered in Figures A.5 and A.6.

Optimal bandwidth parameter is dataset-specific. Repeating the same analysis as above for the concentration of CO2 in the atmosphere (second dataset from Chapter 5) gives slightly different results (Figures A.8 and A.9).

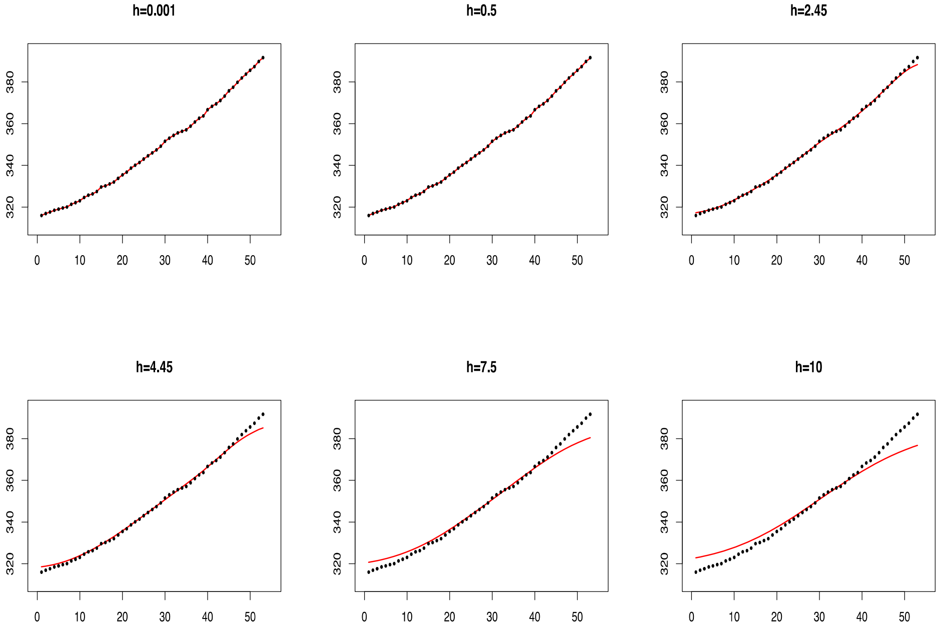


Figure A.8. Varying the smoothing parameter: examples of the NWKEs fitted to the data on concentration of the CO_2 in the atmosphere (*n=53)* for various values of *h*, with Gaussian kernel.

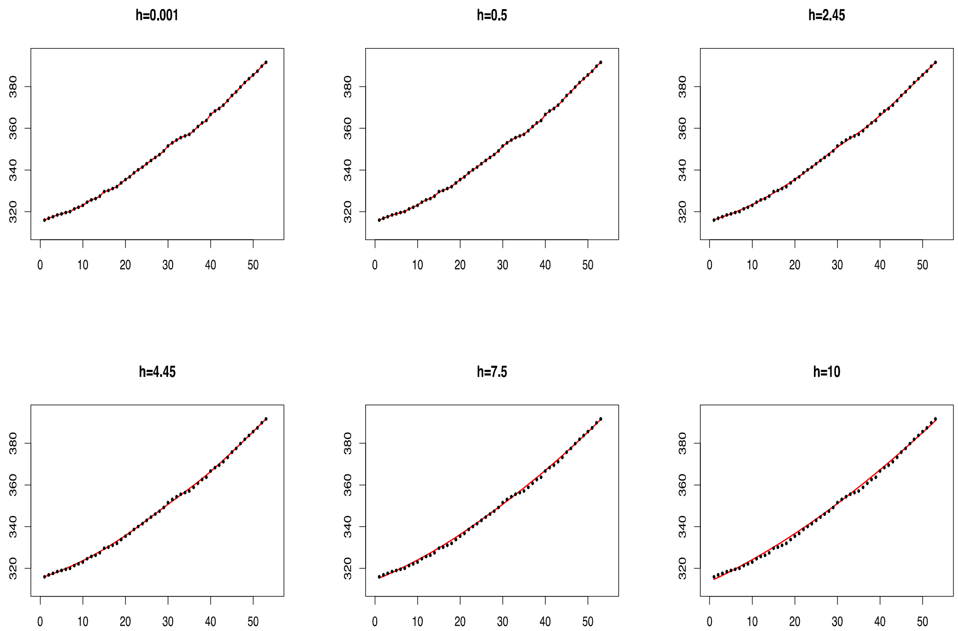


Figure A.9. Varying the smoothing parameter: examples of the LLKEs fitted to the data on concentration of the CO_2 in the atmosphere *(n=53)* for various values of *h,* with Gaussian kernel.

Although varying the smoothing parameter changes the results, the KEs used to estimate the regression function seem to work well. As above (Figures A.5 and A.6) the LLKE performs better, but the difference is not as evident as for the CO_2 emissions data. The main reason is the scale of the standard errors. The comparison of standard errors shows that the results of the NWKE are better (Figure A.10), that is, the standard errors of the LLKE are smaller and the difference is significant, as presented in Figure A.7.

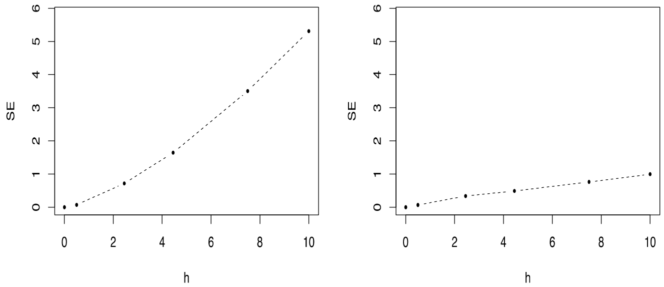


Figure A.10. Relationship between the smoothing parameter and the standard error for the NWKE (left) and the LLKE (right) considered in Figures A.8 and A.9.

Since the smoothing parameter cannot be chosen using the CV criterion for time series (usually correlation causes oversmoothing (Opsomer et al. 2001)), formulas (A.2) and (A.3) should be used.

To estimate unknown factors in (A.2) and (A.3), some additional assumptions are required.

- As a kernel function *K*, we take the Gaussian kernel, so

$$\int_{-\infty}^{+\infty} K^2(x)dx \cong 0.28, \qquad .\int_{-\infty}^{+\infty} x^2K(x)dx = 1.$$

- The explanatory variable has a discrete uniform distribution, and can therefore be roughly approximated by its continuous version. In particular, the PDF of the uniform distribution over an interval is nonzero only over this interval. For simplicity, the factor related to that PDF is constant and can therefore be omitted. To estimate the PDF of the explanatory variable in PB, we use kernel density estimation with the bandwidth chosen by the Silverman's rule of thumb $h = (\frac{1.06\sigma}{n})^{\frac{1}{5}}$ (Silverman 1986).
- For simplicity, we assume that, the unknown regression function is close to a straight line. The factor $\int_{-\infty}^{+\infty}(\hat{r}''(x))^2dx$ is constant and can also by omitted.
- The variance $\hat{\sigma}^2$is assumed constant, and is estimated by

$$\hat{\sigma}^2 = \frac{1}{6(n-2)}\sum_{i=1}^{n-2}(Y_{i+2} - 2Y_{i+1} + Y_i)^2 \qquad (A.6)$$

Therefore, in Section A.4, to find the bandwidth *h*, we use the following rule of thumb

$$h = \left(\frac{\hat{\sigma}^2 0.28}{n}\right)^{\frac{1}{5}} \qquad (A.7)$$

This corresponds to known rules of thumb for NWKE (Green & Silverman 1994), and is used for both NWKE and the LLKE. In this case, formula (A.7) corresponds rather to the optimal bandwidth for the LLKE (no design bias factor), but assuming no bias in the NWKE and approximating h by the same formula, (as for the LLKE) leads to a slight oversmoothing (and hence that assumption becomes reasonable).

A.3.2 In-sample prediction bands – the time series case

For time series data the independence assumption is not satisfied, and, in general, some asymptotic properties of the KE may not be satisfied (Hart 1991). However, for some cases of correlation structure, especially assuming the correlation decays in inverse proportion to the distance between observations (Opsomer et al. 2001), or for the AR correlation structure (Li & Li 2009), asymptotic properties of the KE are close the ones that hold in the independent case. Moreover, generalized version of the CLT, indicates the asymptotic normal distribution, which allows for the use of formulas (A.4) and (A.5) to find the asymptotic prediction bands.

The construction of the PBs is connected with the choice of the smoothing parameter. Adding 95% prediction bands helps in illustrating differences between the results obtained in Section A.3.1 for various values of h.

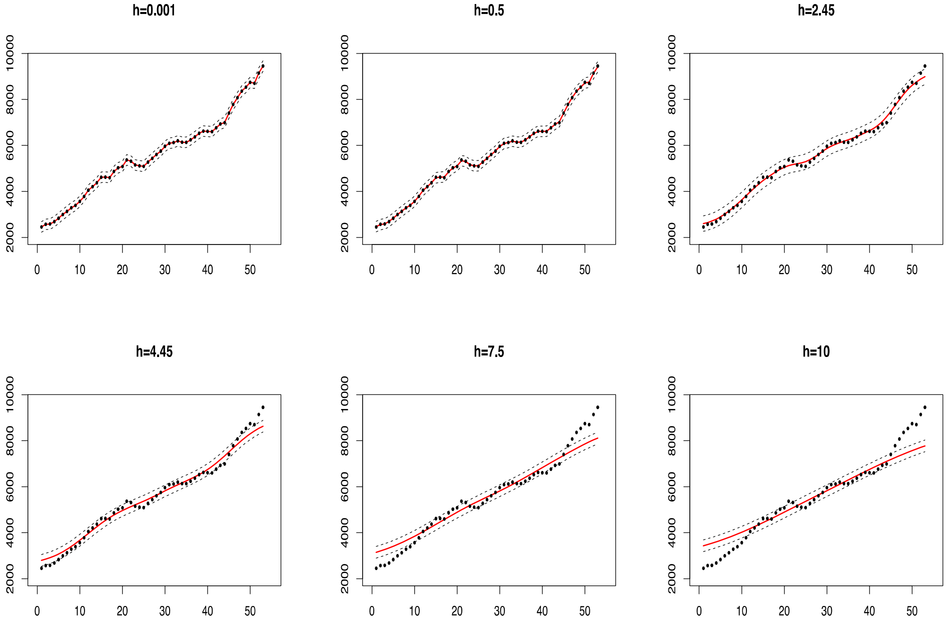


Figure A.10. Varying the smoothing parameter and illustrating its impact on the variance in terms of 95% prediction bands (black dashed lines): examples of the NWKEs fitted to the data on global CO_2 emissions from technosphere ($n=53$) for various values of h, with a Gaussian kernel.

For $h=0.001$, prediction bands do not cover all the data points depicted, since the variance of the estimated regression function is too small and the prediction interval too narrow. Values $h=0.5$ and 2.45 provide different results—the latter appears to be slightly too large, increasing the variance and causing the wider prediction interval. For $h=10$, the regression estimate is obviously oversmoothed. The shape of the data is not properly reflected, and despite the large variance, only few data points fall within the prediction bands[67].

[67] That effect is partly connected with boundary bias of the NWKE.

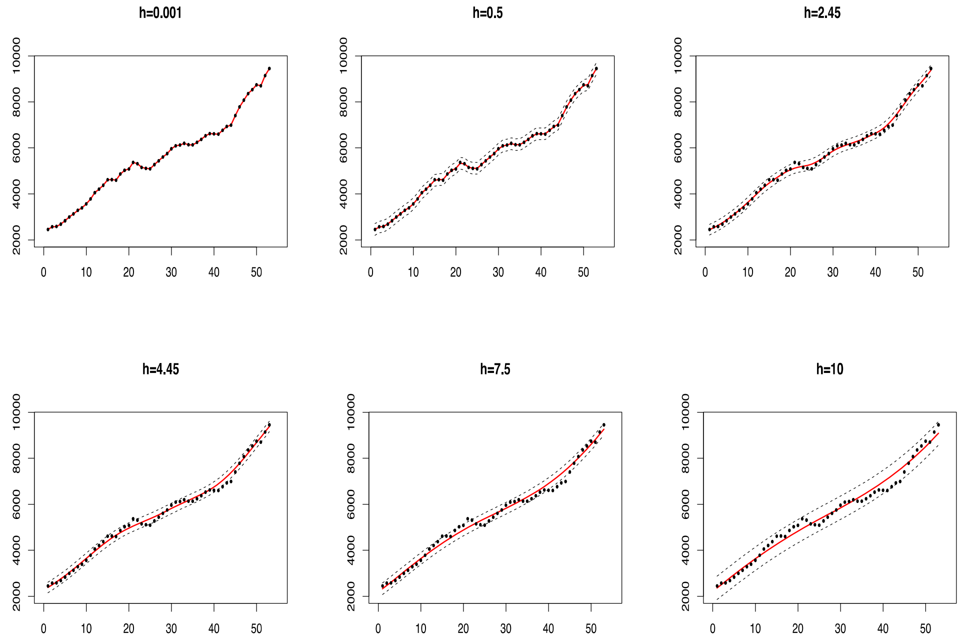


Figure A.11 Varying the smoothing parameter and illustrating its impact on the variance in terms of 95% prediction bands (black dashed lines): examples of the LLKEs fitted to the data on global CO_2 emissions from technosphere (*n=53)* for various values of *h,* with a Gaussian kernel.

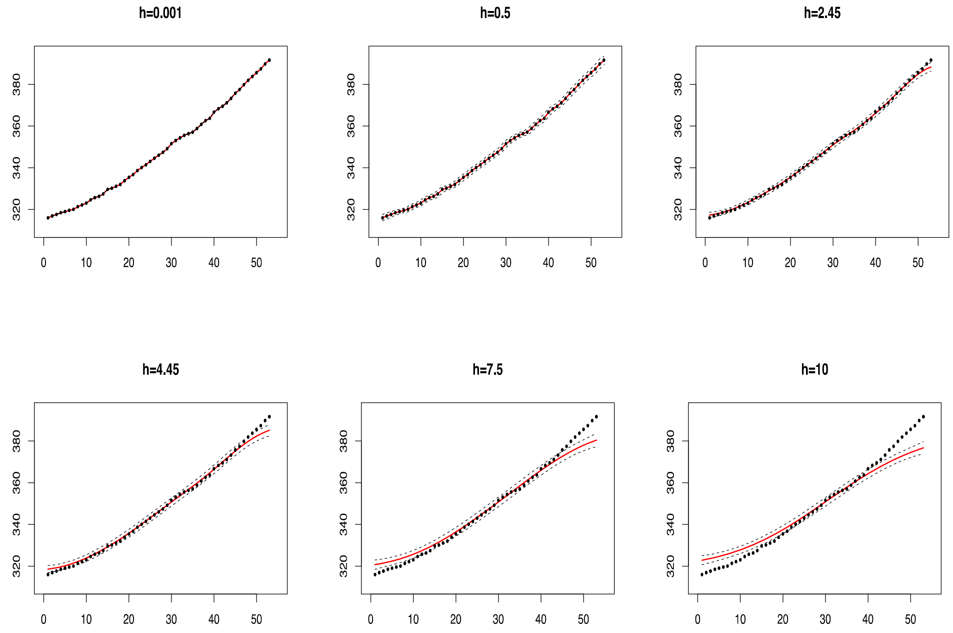


Figure A.12 Varying the smoothing parameter and illustrating its impact on the variance in terms of 95% prediction bands (black dashed lines): examples of the NWKEs fitted to the data on concentration of the CO_2 in the atmosphere (*n=53)* for various values of *h,* with a Gaussian kernel.

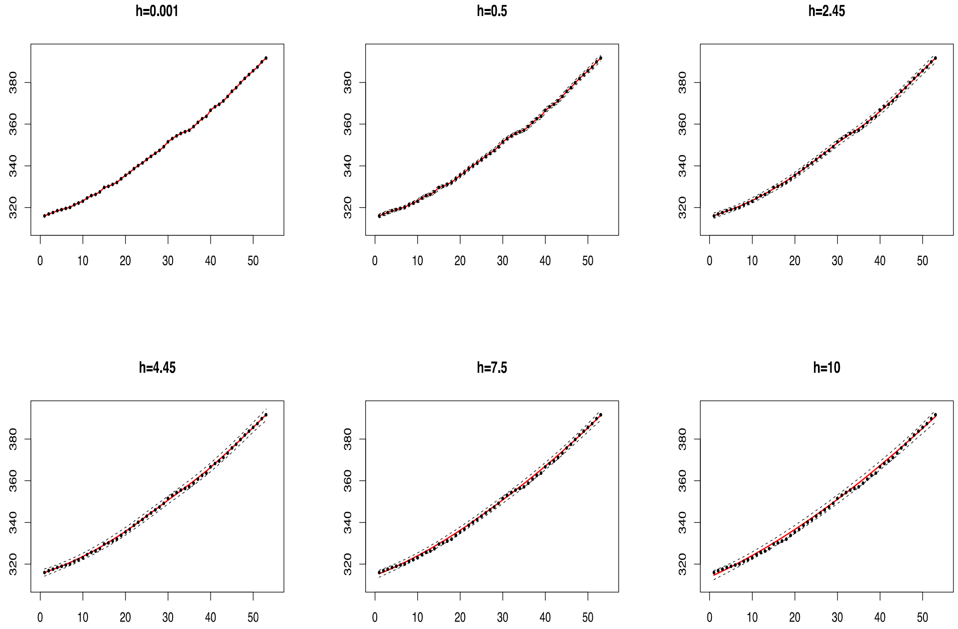


Figure A.13 Varying the smoothing parameter and illustrating its impact on the variance in terms of 95% prediction bands (black dashed lines): examples of the LLKEs fitted to the data on concentration of the CO_2 in the atmosphere (*n=53)* for various values of *h,* with a Gaussian kernel.

A.4 Real-life case studies

The methods of PL from Chapter 3, in particular the procedure for assessing the EO, are applied to real-life case studies, considered in Chapter 5: (1) global CO_2 emissions from the technosphere, and (2) concentration of CO_2 in the atmosphere.

PL is tested in terms of the EO (described in Section 3.2) for both aforementioned kernel regression estimators: LLKE and the much simpler NWKE.

The most problematic aspect of using nonparametric methods is their requirement of a large sample size, but each of them (including kernel regression) depend on the sample size in a different way. Because of the asymptotic properties of kernel estimators, the sample should be sufficiently large, although it is difficult to specify the threshold above which the results will be good. The conducted analyses and simulations (Wasserman 2006, Green & Silverman 1994) indicate that this depends on the type of data, in particular on their distribution. Also, correlation of data (as in the time series case) requires a larger number of test points (Opsomer at al. 2001, Hart 1991). It can therefore be expected that for LBs of 25 or slightly more training points, the results may not be satisfactory, which will influence the EO in some way.

A.4.1. Procedure for analysing the EO, in the case of the kernel regression

To test the PL method, the following procedure is considered:

Given the sample of $n = n_1 + n_2$ data points, we perform the following steps.

Step 1. We take the LB of n_1 data points.

- The unknown variance of residuals is estimated by (A.6)
- The smoothing parameter is found by (A.7)
- The NWKE and the LLKE are used.
- The model assumptions are verified.
- The in-sample 95% prediction bands are found for both NWKE and LLKE, using (A.4)

Step 2. We take the testing sample of n_2 data points.

- The out-of-sample 95% prediction bands are found for both the NWKE and LLKE, using (A.5)
- The length and the score of the EO are found, using the procedure described in Section 3.2.

Step 3. We increase the LB by one and repeat Step 1 and Step 2.

A.4.2. Global CO_2 emissions from the technosphere

The procedure described in Section A.4.1 is applied, starting with $n_1 = 25$. The six exemplary stages are presented in Figures A.14 (for the LLKE) and A.15 (NWKE).

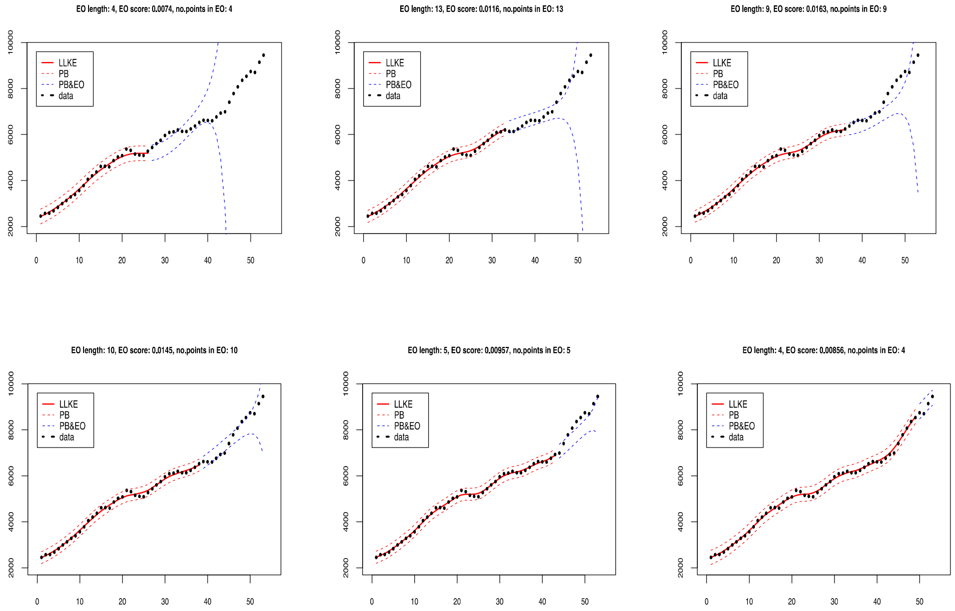


Figure A.14. Six exemplary stages of the PL procedure (LB lengths: 26, 33, 36, 38, 43, 49): the LLKE using the Gaussian kernel.

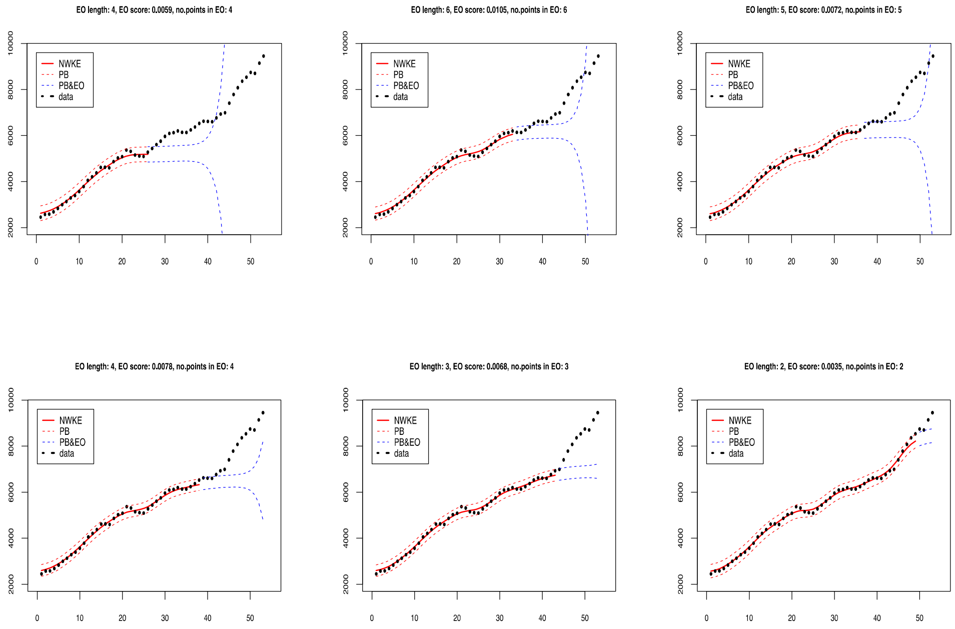


Figure A.15. Six exemplary stages of the PL procedure (LB lengths: 26, 33, 36, 38, 43, 49): the NWKE using a Gaussian kernel

For both estimators the 95% out-of-sample PBs (for the shortest LBs open quite fast[68], but the PBs for the NWKE, in particular for the shortest LBs, seem to stabilize at first, increasing rapidly after a few out-of-sample points. This is related to the boundary bias of the NWKE, in particular for small samples.

The PBs for the LLKE better reflect the estimated relationship between explanatory and response variables, which also results in the longer EO. The prediction intervals for the NWKE are wider, which is connected with the greater standard errors, and results in lower EO scores (Figure A.17).

In contrast to the EO lengths presented in Figure 37—as a result of using parametric linear regression—no decreasing trend can be observed, for LB>30. The EO lengths decrease and increase, for the LLKE having peaks at LB=32 (local maximum), 34 (local minimum), 37 (max), and then 42 (min), 43 (max), 44 (min) and 47 (max). For LB>48, all the remaining data points are within the PBs, giving the infinite length.

It is worth mentioning, that in spite of differences in the EO lengths, the results obtained using both estimators show similar monotonic behavior (Figure A.16). A similar effect can be observed for the EO scores (Figure A.17). This means that the EO depends on the data. Since in the case of the LLKE standard errors are smaller than for the NWKE, the prediction intervals for LLKE are narrower, and the data type affects the EO outcome more strongly.

[68] This is connected with prediction errors increasing very fast. The in-sample errors behavior is completely different (Figures A.10 and A.11), as they seem to be constant.

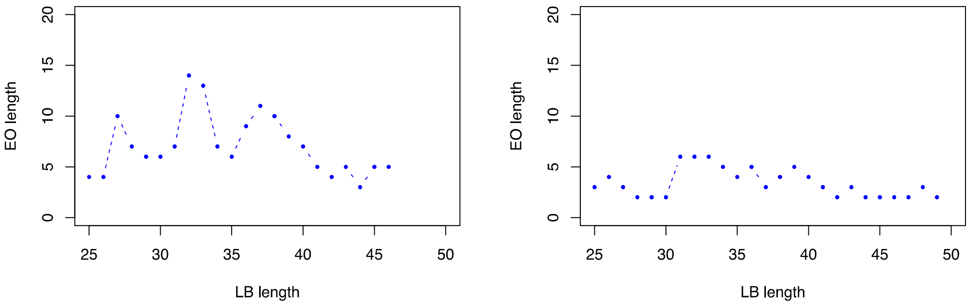


Figure A.16. The EO length as a function of the LB, in the case of the LLKE (left) and the NWKE (right).

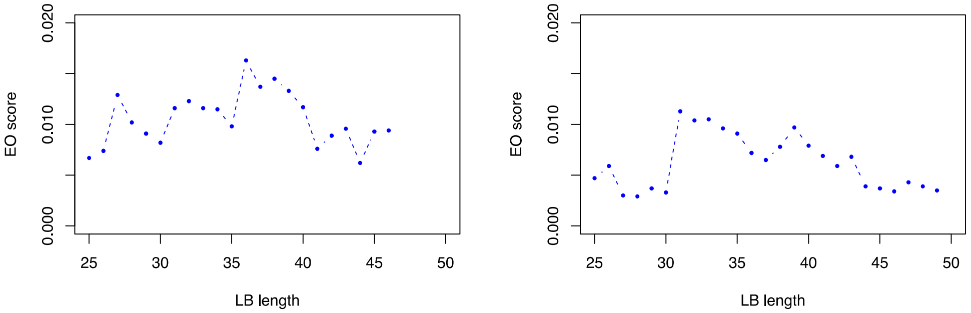


Figure A.17. The EO score as a function of the LB, in the case of the LLKE (left) and the NWKE (right).

The comparison of the results for the LLKE and NWKE is presented in Table A.2. The conducted analysis shows that the LLKE performs better, giving longer EOs—between 4 and 14 data points (Figure A.16).

Moreover, starting with an LB of 47 points, all the remaining data points fall within the PBs. The resulted EO lengths for the NWKE are in turn more stable, giving values between 2 and 6.

Table A.2 Prognostic learning—a comparison of the LLKE and NWKE results when applied to the data on CO_2 emissions from the technosphere.

Results		**LLKE**	**NWKE**
EO	max length	finite: 14 (for LB=32) ∞ (for LB≥47)	finite: 6 (for LB=31, 32, and 33) ∞ for LB≥50
	min length	4 (for LB=25, 26, 42 and 44)	2 (for LB=28, 29, 30, 41, 44-47, and 49)
	infinite length	for LB≥47 all tested data points fall within the PBs	for LB≥50 all tested data points fall within the PBs
	score	0.0062 – 0.0163 for LB<47 ∞ for LB≥47	0.0029 – 0.0113 for LB<50 ∞ for LB≥47
Residuals	normality	ε_t normally distributed (Shapiro-Wilk test, p-values>0.2)	ε_t normally distributed (Shapiro-Wilk test, p-values>0.1)
	zero mean	ok (t-test, p-values>0.2)	ok (t-test, p-values>0.2)
	correlation	autocorrelation at lag 1 and 2, (ACF, Box-Pierce test)	autocorrelation up to lag 5 or 6 (ACF, Box-Pierce test)

A.4.2 Concentration of CO_2 in the atmosphere.

Now the procedure described in Section A.4.1 is applied to the second dataset. As previously, we start with $n_1 = 25$ and then increase the LB length by one. The six exemplary stages of the procedure are presented in Figures A.18 (for the LLKE) and A.19 (for the NWKE).

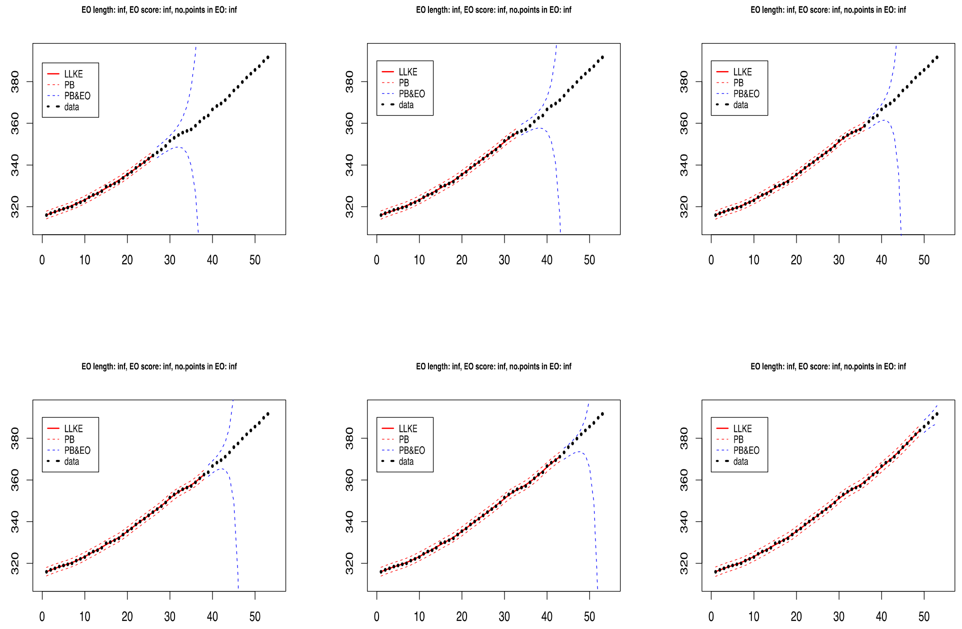


Figure A.18. Six exemplary stages of the PL procedure (LB lengths: 26, 33, 36, 38, 43, 49): the LLKE using a Gaussian kernel

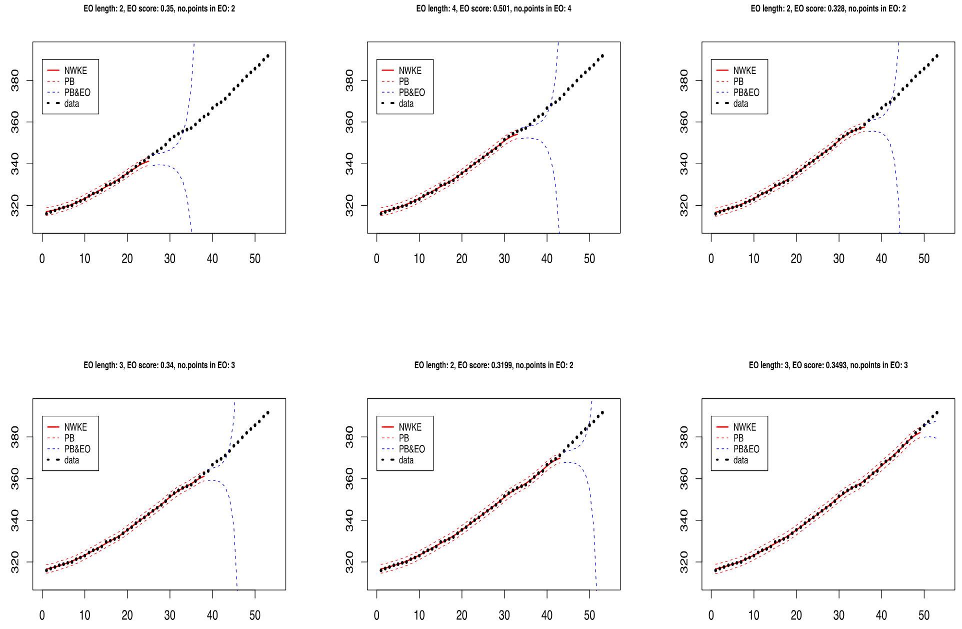


Figure A.19. Six exemplary stages of the PL procedure (LB lengths: 26, 33, 36, 38, 43, 49): the NWKE using a Gaussian kernel

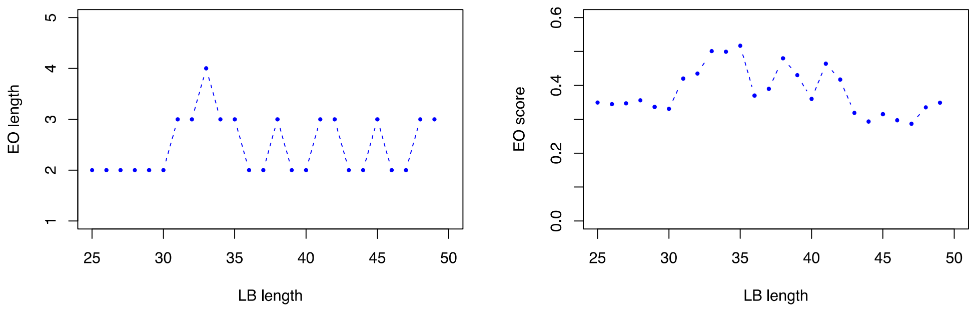


Figure A.20. The EO length (left) and EO score (right) as a function of the LB, in the case of the NWKE.

The comparison of the results for the LLKE and NWKE is presented in Table A.3. The conducted analysis shows that, the PL method based on LLKE fails to establish the length of the EO. As a result of quickly diverging PB, all testing points fall within them and the resulting EO lengths are infinite (i.e., undefined).

The NWKE method on the other hand performs poorly. This is caused by the boundary bias resulting in horizontal EO while the testing points continue to follow an increasing trend.

Table A.3 Prognostic learning—comparison of the LLKE and NWKE results when applied to the data on concentration of CO_2 in the atmosphere.

Results		**LLKE**	**NWKE**
EO	max length	∞ (all tested points fall within the PBs)	4 (for LB=33)
	min length	no finite EO length	2 (for LB=25-31,36-37, 39-40, 43-44, and 49)
	∞	for LB≥25 all tested data points fall within the PBs	for LB≥50 all tested data points fall within the PBs
	score	∞ (no finite EO score)	finite: 0.287 – 0.517 or ∞ (for LB≥50)
Residuals	normality	ε_t normally distributed (Shapiro-Wilk test, *p*-values>0.1)	ε_t normally distributed (Shapiro-Wilk test, *p*-values>0.09)
	zero mean	ok (t-test, *p*-values>0.2)	ok (t-test, *p*-values>0.2)
	correlation	autocorrelation at most at lag 1 or none (ACF, Box-Pierce test)	autocorrelation at lag 1 (at most 2) or none (ACF, Box-Pierce test)

A.5 Conclusions

Analysis of the performance of the PL method based on nonparametric regression applied to real-life datasets of anthropogenic CO_2 emissions and atmospheric CO_2 concentrations leads to the following conclusions:

- The use of the LLKE regression performs better than the NWKE. Since it does not exhibit the boundary bias it has smaller prediction errors. This results in longer prediction errors.
- The method based on nonparametric regression easily adapts to the data behaviour, reflecting fluctuations and peaks (for CO_2 emissions dataset) while being more stable for data exhibiting regular behaviour (as for the CO_2 concentrations dataset).
- Autocorrelation of residuals (more pronounced for the NKWE method than for the LLKE method) has a negative impact on the performance of the PL procedure, that is, it results in shorter EOs.

Acronyms

ACF – autocorrelation function

AR - autoregression

CLT – central limit theorem

CV – cross-validation

GMKE – Gasser-Müller kernel estimator

KE – kernel estimator

k-NNKE – k-nearest neighbour kernel estimator

LLKE – Local linear kernel estimator

MLE – maximum likelihood estimation

MSE – mean squared error

NWKE – Nadaraya-Watson kernel estimator

PB – prediction bands

PCKE – Priestley-Chao kernel estimator

PDF – probability density function

SE – standard error (i.e. residual standard error)

Literature

Altman N.S., *Kernel Smoothing of Data with Correlated Errors*, J. Amer. Statist. Assoc., 1990, 85, 749-759.

K. De Brabanter, J. De Brabanter, J. A. Bart De Moor, *Kernel Regression in the Presence of Correlated Errors*, J.Machine Learn. Research 12 (2011), 1955-1976.

Eubank R.L., *Nonparametric regression and spline smoothing*, Marcel Dekker Inc., New York, 1999.

Fan J., *Design-adaptive Nonparametric Regression*, Journal of the American Statistical Association, Vol. 87, 1992.

Fan J., Gijbels I., *Local Polynomial Modeling and Its Applications*, Chapman & Hall, London, 1997.

L. Gajek, M. Kałuszka, *Wnioskowanie statystyczne: modele i metody*, Wydawnictwa Naukowo-Techniczne, Warszawa, 1993.

Gasser T., Müller H.G, *Estimating Regression Functions and Their Derivatives by the Kernel Method*, Scand. J. Statist., 1984, **11**:171-185.

Green P.J., Silverman B.W., *Nonparametric Regression and Generalized Linear Models: a Roughness Penalty Approach*, Chapman & Hall, London, 1994.

Györfi L., Kohler M., Krzyżak A., Walk H., *A Distribution-Free Theory of Nonparametric Regression*, Springer, NewYork, 2002.

Hart J.D., *Kernel regression estimation with time series errors*, J. Royal Statist. Soc. B. 1991, **53**(1):251-259.

Härdle W., *Applied Nonparametric Regression*, Cambridge University Press, 1990.

Johnstone I., Silverman B.W., Wavelet threshold estimators for data with correlated noise, J. Royal Statist. Soc., B., 1997, **59**, 319-351.

Jarnicka J., *Multivariate kernel density estimation with a parametric support*, Opuscula Math. **29**, no. 1 (2009), 41-55.

Nadaraya E.A., *On estimating regression*, Theory Prob. Appl.1964, **9**(1): 141-142.

Nason G.P., Wavelet shrinkage using cross-validation, J. Royal Statist. Soc. B, 1996, **58**, 463-479.

Opsomer J., Wang Y, Yang Y., *Nonparametric Regression with Correlated Errors*, Statist. Sci. 2001, **16**(2): 134-153.

Priestley M.B., Chao M.T., *Non-parametric function fitting*, J. Royal Statist. Soc. B, 1972, **34**: 385-392.

Rice J., Rosenblatt M., *Integrated mean squared error of a smoothing spline*, J. Approx. Theory, 1981, **33**, 353-369.

Rice J., Rosenblatt M., Smoothing splines: regression, derivatives and deconvolution, Ann. Statist, 1983, **11**, 141-156.

Ruppert D., Wand M.P., *Multivariate Locally Weighted Least Squares Regression*, The Annals of Statistics, 1994, Vol. 22, p. 1346-1370.

Silverman B.W., *Density Estimation for Statistics and Data Analysis*, Champan & Hall, New York, 1986.

Simonoff J.S., *Smoothing Methods in Statistics*, Springer, 1996.

Stone C.J., *The use of polynomial splines and their tensor products in multivariate function estimation*, Ann. Statist., 1994, **22**, 118-184.

Wang Y., Function estimation via wavelet shrinkage for long-memory data. Ann. Statist**. 24**, 1996, 466-484.

Wasserman L., *All of Nonparametric Statistics*, Springer Texts in Statistics, New York, 2006.

Watson G.S., Smooth regression analysis, Sankhya Ser. A, 1964, **26**(4): 359-372.

PART II. The crux of reducing emissions in the long-term: The underestimated “now” versus the overestimated “then”

Matthias Jonas[1], Piotr Żebrowski[1]

[1] IIASA, Advanced Systems Analysis Program

Abstract

This article is a perspective piece. It aims to elaborate on the usefulness of GHG emission inventories by obtaining deeper insights into their inherent systemic features—we focus on memory, persistence, and uncertainty regarding the near-term future. We conjecture that, as is typical in forced systems with memory, prognostic emission reduction scenarios underestimate not only the degree to which GHG emissions will continue on their historical path beyond “today” but also the extent of their persistence on that path. This leads to overestimation of the amount of reductions that might be achieved in the future.

Memory allows us to reference how strongly a system’s past can influence its near-term future. We consider memory to be an intrinsic property of a system, retrospective in nature; and we consider persistence to be a consequential (i.e., observable) feature of memory, prospective in nature, and reflecting the tendency of a system to preserve its current state (including trend).

There are different approaches to capturing memory. For example, we can capture memory using three of its characteristics: its temporal extent, its weight over time, and its quality over time. In the example provided, we focus on systemic insight. Our intention is to illustrate one way (among others) of reflecting memory and thereby understand how persistence plays out and how it determines the system’s near-term future. However, the example we provide does not exhibit fundamental shortfalls, and it does not restrict generalization. We use it to look into the following three questions: (i) How well do we need to know the characteristics of memory mentioned above (and/or possibly others) to be able to delineate a system’s near-term future by means of what we call its explainable outreach (EO)? (ii) Can we differentiate between and specify the various characteristics of memory (i.e., those mentioned above) by way of diagnostic data-processing alone? Or, to put it in another way, how much systems understanding do we need to have, and need to inject into, the data-analysis process to be able to make that differentiation? Moreover, (iii) can, or even should, the derivation of EOs become an integral part of model building?

The insights gained from answering these three questions indicate that our conjecture has a high chance of proving true. Being ignorant of memory and persistence, we underestimate, probably to a considerable extent, the momentum with which GHG emissions will continue on their historical path beyond today, and therefore overestimate the reductions that we might achieve in the future.

Keywords**:** Greenhouse gas emissions, emissions inventory, memory, persistence, explainable outreach, uncertainty

Contents

1. Introduction

1.1 Acknowledging memory and recognizing persistence

This article is a perspective piece. It aims to elaborate on the usefulness of greenhouse gas (GHG) emission inventories (cf. Lieberman *et al.* 2007; White *et al.* 2011; Ometto *et al.* 2015). Emissions experts are trying to understand how helpful GHG emission inventories are in reconciling short- and long-term emission estimates. Given that emission paths are sensitive to historical conditions, as well as uncertain with respect to their mandated reduction levels, experts are grappling with the probability that long-term targets will be missed (Jonas *et al.* 2010a: 4.3; 2010b; 2014).

We deal with this question **from a data-processing perspective**. To obtain promising outcomes, we lower our ambitions by posing two conditions. First, we restrict ourselves to **systems with memory**. These systems, which are typical in Earth system sciences, include the emissions system. Memory allows us to reference how strongly a system's past can influence its **near-term future** (e.g., Wikibooks 2013; Wikipedia 2017; Joshi 2016; Brockwell & Davis 2002). This brings us to our second condition: we make only a "small" step into the future. This step is sufficient to show that the way a data analyst and a prognostic modeler understand uncertainty, when stepping across "today"—the interface between past and future—fundamentally differ.

When we speak of memory, we refer to the momentum of a system over a historical time range. In the physical world of Newtonian systems, momentum indicates a system's inertia, or, if the momentum is not constant, its deviation from an inert state due to an external force. Likewise, systems with memory also exhibit inertia—called **persistence** below—under the impact of an external force. (For inertia in the climate system, refer to, e.g., IPCC 2001: SPM: Q5.) The notion of memory suggests approaches that attempt to characterize the dynamics of a system in terms of trend and fluctuations around that trend (e.g., Kantelhardt 2004). However, separating the two, as we will show, may not always be possible.

Different approaches are able to capture memory. (We look into how memory is understood by the scientific community in Section 1.3.) By way of example, we capture memory with the help of three of its characteristics: **its temporal extent, its weight over time, and its quality over time**. The extent of memory quantifies how many historical states directly influence the current state, while the weight of memory describes the strength of that influence. The quality of memory steers how well we know the latter.

One important question **is how well we need to know these (and/or possibly other) characteristics of memory in order to delineate a system's near-term future**, which we seek to do by means of what we call the system's explainable outreach (EO). In our perspective piece, we argue that we have reasons for optimism that the system's EO can be derived under both incomplete knowledge of memory and imperfect understanding of how the system is forced.

Our focus is on **forced systems**. In many cases, we know that a system possesses memory, for example, because it does not respond instantaneously to the forcing it experiences. But we find it difficult to quantify memory, let alone visualize it, in a way that is easy to understand, particularly for practitioners and decision-makers.

Figure 1 is as a major example of a forced system. Here, the forcing is due to anthropogenic activities (e.g., fossil-fuel burning, cement production, and land use). The figure informs us of the emission reduction paths we would have to follow almost instantaneously at the global scale to keep global warming at or below 2ºC to prevent the most dangerous impacts of climate change. However, the figure does **not** inform us about the "degree and extent of persistence" with which GHG emissions will continue along their historical path into the future, for example, due to old, carbon-intensive furnaces remaining in operation; such knowledge is crucial for the design, implementation, and effectiveness of realistic emission-reduction policies and for overcoming path-dependences caused by memory.

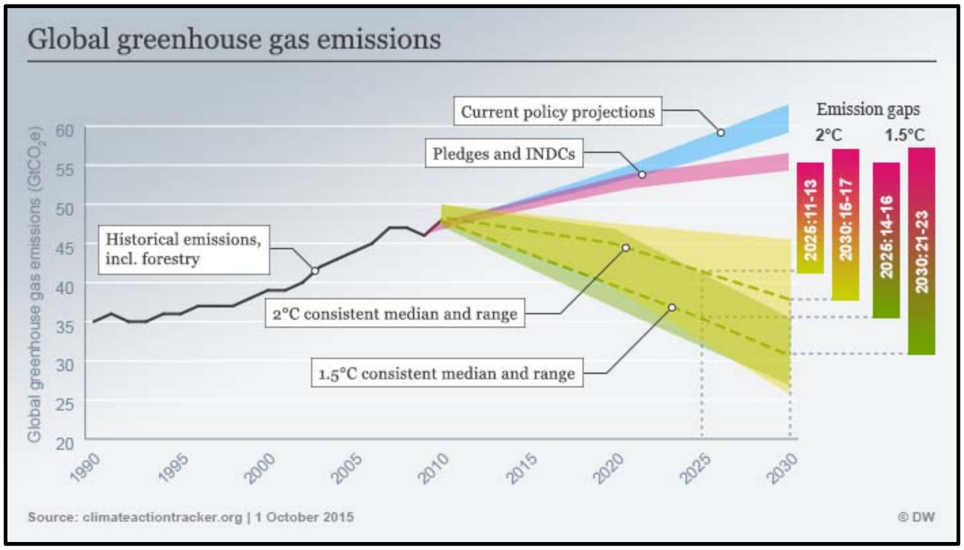


Figure 1. Illustrating the effect of implementing pledges and the so-called intended, nationally determined contributions of 146 governments, and comparing these with the expected absolute emissions in 2020, 2025, and 2030 with respect to the 1.5ºC and 2ºC benchmark emission pathways in accordance with the UNFCCC Paris Agreement (DW, 2015).

Is it then possible to distinguish between the various characteristics of memory and specify them (e.g., those mentioned above) through diagnostic data-processing alone? Or, to put it another way, **how much systems understanding do we need to possess and also to inject into the data-analysis process to enable those distinctions to be made?** This question, which is directly related to our previous question, is of considerable interest. As yet, we do not see any possibility of it being answered theoretically. However, we do see a value in exploring it in order to identify approximate, yet sufficiently robust, modi operandi to identify EO concepts that are easy to apply in practice.

The main objective of this perspective piece is to explore and reflect on an approach that allows a system's EO to be derived and to demonstrate the usefulness of doing so (Section 2). To that end, we start from a simple synthetic data (time) series example—our control—which we equip, step by step, with realistic physical features, such as memory and noise, while exploring the system's persistence—the counterpart to memory—and deriving its EO (**forward mode**). The main reason for providing the example is to better understand memory and persistence and to consolidate our systems thinking. Nonetheless, our perspective piece should be considered as explorative only. Systemic insight is more valuable than mathematical rigor. The example is geared to making the concept of EOs applicable (Sections 2.1 to 2.3). We discuss how consequential the example is, where it underperforms, and the questions it raises in Section 2.4.

The remaining two sections of the introduction serve to deepen our understanding of memory and to frame our mindset for the following chapters. In Section 1.2 we expand on memory, persistence, and EO; and in Section 1.3 we provide an overview of how memory and persistence are understood and interpreted by various scientific communities.

In Chapter 3 we discuss the problems that we envisage in quantifying persistence without having a priori knowledge about memory and its major characteristics (**backward mode**). In Chapter 4 we expand on the issue of applicability by asking **whether the derivation of EOs can, or even should, become an integral part of model building**. We see good reasons for answering in the affirmative. Chapter 5 summarizes our insights.

1.2 Our understanding of memory, persistence, and explainable outreach

In this section, we gather our initial understanding of memory, persistence, and explainable outreach, which varies according to scientific community (cf. Section 1.3). This explains why some leeway exists in understanding (and defining) these terms.

1. We consider memory to be an intrinsic property of a system, retrospective in nature, and persistence to be a consequential (i.e., observable) feature of memory, prospective in nature. Persistence is understood to reflect the tendency of a system to preserve a current state (including trend) and depends on the system's memory which, in turn, reflects how many historical states directly influence the current state. The nature of this influence can range from purely deterministic to purely stochastic.

2. Deriving an EO should not be confused with prediction (or with perfect forecasting). In statistics, predictability is used in the context of in-sample and out-of-sample predictability, neither of which interests us. However, there is a potential for misunderstanding which is rooted in the way an EO is applied as a measure of reference. Deriving a time series EO requires the historical data of the series to be evaluated by applying learning **and** testing (what we call **learning under controlled prognostic conditions**). Ultimately, shifting the EO to the end of the series' historical data (= today) has to happen untested and would therefore **not** be permitted. However, this forward shift is still done only to provide a bridge into the immediate future (see Point 3), and thus act as **a reference measure for prognostic modelers and decision-makers**. Shifting the EO to today requires a conservative systems view which ensures that the system is not exposed to any surprises it had not previously experienced, as these can cause the system to fall outside the EO.

3. Figure 2 visualizes the idea of using the EO as a reference measure for prognostic modelers and decision-makers. An EO is derived only from the historical data of a time series and then shifted to its end (= today). Prognostic scenarios falling outside (above or below) the EO, as well as scenarios falling within it but eventually extending beyond the EO, are no longer in accordance with the series' past—allowing a decision-maker to enquire about the assumptions made in constructing a forward-looking scenario and to interpret them in terms of how effective the planned measures (e.g., emissions reductions) need to be and/or for how long the effectiveness of those measures will remain uncertain. We consider an EO taking the form of an uncertainty wedge to be a more appropriate reference measure for the immediate future than a single, model-dependent, business-as-usual scenario used as a reference by modelers.

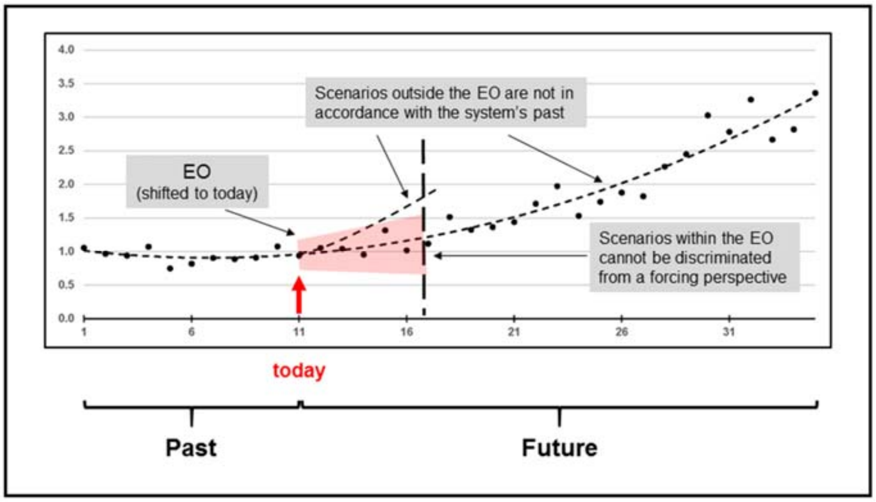


Figure 2. Illustrating why knowing the EO of a time series is important (see text). For convenience in constructing the figure, we assumed that a future is known (see black dots in the future part of the time series).

4. Deriving the EO of a time series must not be confused with signal detection. The term "signal" encourages thinking in terms of deviations from a predefined baseline (which can also be the zero line). We practice signal detection elsewhere to evaluate GHG emissions in an emissions change-versus-uncertainty context (cf. Jonas *et al.* 2010b). Figure 2 helps to understand why deriving the EO of a data series and detecting its signal fundamentally differ. Signal detection requires determination of the time at which changes in the data series outstrip uncertainty—which is not done here. Take, for example, the scenario falling above the EO. The time at which its signal steps out of the uncertainty wedge does not coincide with the temporal extent of the EO.

5. Assuming persistence to be an **observable** of memory, Figure 2 allows persistence to be quantified (**not** defined). Given its directional positioning, the red-shaded EO in the figure may be described with the help of two parameters, its extent L, and its aperture A at the end. We would then say that a time series with a long and narrow EO (the ratio L/A would

be large) exhibits a greater persistence than a time series with a short and wide EO (the ratio L/A would be small).

L and A are characteristics of the model selected to analyze the given time series. We expect that increasing the resolution ($R\uparrow$) with which the system is observed will result in a decreasing extent of EO ($L\downarrow$). The EO of the selected model is likely to prove less suitable for capturing the more fluctuating time series, which discloses more detail (surprises) than before (resulting in $A\uparrow$). To increase persistence to its prior value, our model would have to be modified so that it smooths the time series more than it did before.

1.3 Memory and persistence understood by various scientific communities

It is in the context of analyzing the structural dependences in the **stochastic component** of a time series that the terms "memory" and "persistence" commonly appear and are widely discussed. In modeling the **deterministic component** of a time series, memory effects are typically believed **not** to be of major concern and that the component can be captured by means of suitable regression methods (e.g., linear/polynomial or non-parametric)—something which is easy to refute (cf. example in Chapter 2).

Memory and persistence are not strictly defined, as they are regarded as statistical properties resulting from the time series' structural dependences, of which there is a great variety. This thus leaves room for interpretation—however, scientific communities understand these terms in different ways and may also apply different methods to analyze them. Table 1 gives an overview of the terminology used by various scientific communities when referring to memory and persistence, and how they interpret them.

Table 1. Memory and persistence as understood and interpreted by various scientific communities. Source: Jarnicka (2017; pers. comm.)

Field	Terminology	Interpretation	Literature
Climate Analysis	**Memory**, **dependence** (distinguishing between short-term/short-range and long-term/long range)	Rate of decay of the autocorrelation function (considered geometrically bounded; but also with exponential, power rate or hyperbolic decay)	Caballero *et al.* (2002); Palma (2007); Franzke (2010); Mudelsee (2010); Lüdecke *et al.* (2013); Barros *et al.* (2016); Belbute & Pereira (2017)
	also, **persistence**	Long-range memory (also checked with the help of spectral or fluctuation analysis)	
Economy & Finance	**Serial dependence**, **serial correlation**, **memory**, **dependence**	Statistical dependence in terms of the correlation structure with lags (mostly long memory, i.e., with long lags)	Lo (1991); Chow *et al.* (1995); Barkoulas & Baum (1996); Dajcman (2012); Hansen & Lunde (2014)
	also, **persistence**	positive autocorrelation	
Geophys. & Physics	**Persistence**, **dependence**, also **memory** (mostly long-term)	Correlation structure in terms of Hurst exponent or power spectral density; but systems dynamics also expressed by regularities and repeated patterns	Majumdar & Dhar (2001); Kantelhardt *et al.* (2006); Lennartz & Bunde (2009, 2011)

2. Example

The example presented in Chapter 2 focuses on systemic insight. Its purpose is to illustrate one way (among others) of reflecting memory, understanding how persistence plays out, and deriving an EO. The example is discussed intensively with respect to how consequential it is, where it underperforms, and the questions it provokes, which are listed at the end of Chapter 2. However, the example does **not** exhibit fundamental shortfalls. It does not restrict generalization, while allowing us to spot the important research issues that we will be facing in terms of deriving the EO of a data series.

The example is geared toward making the concept of EOs applicable. Figure 3 visualizes the different "worlds of knowledge" with which the example confronts us. Some of its features are excessively exaggerated to better understand how memory can lead to persistence even under unfavorable conditions, for example, a forcing that is weak and a memory with an extent that is short compared to the noise which is superimposed.

The example is dealt with throughout Sections 2.1 to 2.3. Section 2.1 is composed of two steps. In the first step, we obliterate the knowledge of our control, a 2nd-order polynomial, by applying a high level of noise; in the second step, we limit and steer this obliteration back in time by introducing memory in terms of extent, weight, and quality. This allows what had previously been obliterated to be reconstructed. The qualitative and quantitative characteristics of this reconstruction remain to be investigated against a reference. The intention behind this step-wise procedure is to develop an understanding of how memory works and how it leads to persistence. Section 2.2 visualizes this process graphically; Section 2.3 offers one way of deriving an EO; and Section 2.4 summarizes important insights and questions.

2.1 Mental and numerical set-up

We work with four functions dependent on x (with $x=1,...,35$; sufficiently long for illustrative purposes) which can, but need not, be interpreted as time series dependent on time t measured in years.[1] The functions can be understood to reflect four observers (O) who perceive an otherwise accurate world differently—precisely or imprecisely, and with perfect knowledge, and limited or no memory (cf. upper half of Figure 3).

To start with, all observers have complete (but not necessarily perfect) knowledge of their worlds (i.e., x extending from 1 to 35). We introduce two additional observers later when we

[1] This choice of interpretation, although it allows our example to be simplified, is systemically important. Consider, for example, the logical link between GHG emissions—atmospheric concentration—global mean surface warming, and the lag (memory) effect between any two of them; say concentration (C) and emissions (E). It is this two-data-series perspective, here the $C=C(E)$ perspective in the t-E-C space, in which practitioners are interested. However, reducing the two-data-series perspective to the perspective of a single time series, here $E=E(t)$ or $C=C(t)$, comes in useful. It allows memory to be described deterministically **and/or** stochastically. Here we apply the single-time-series perspective. However, to acknowledge the wider perspective, we use x as variable, **not** t.

split the time series into past (x from 1 to 7) and future (x from 8 to 35). These two observers will have incomplete knowledge because they see the historical part of the time series only (cf. lower half of Figure 3).

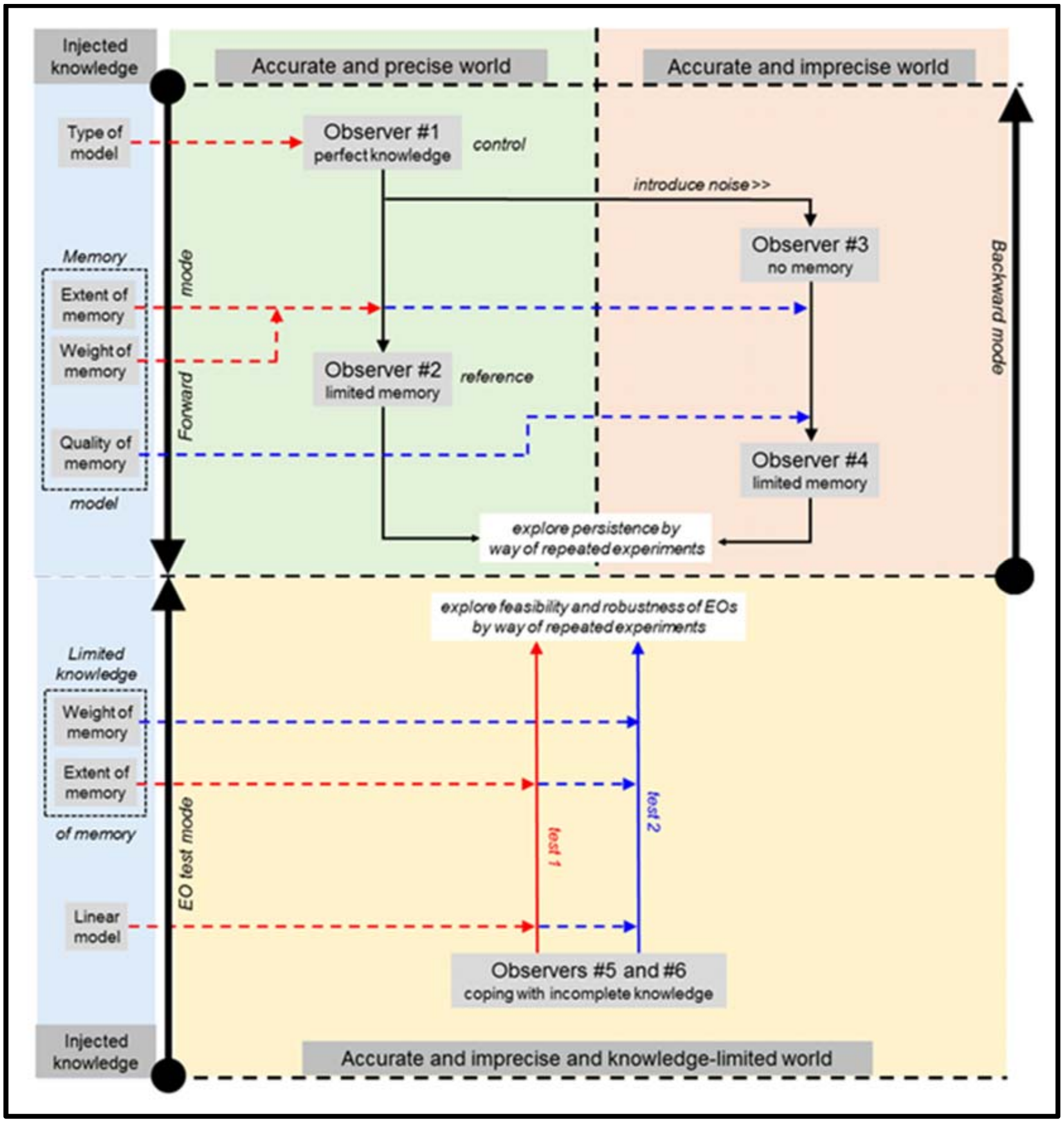


Figure 3. Graphical visualization of the different "worlds of knowledge" underlying the example discussed in Chapter 2. The main purpose of the figure is to distinguish these "worlds" by means of the knowledge that is injected into expanding the example step-by-step.

The world of observer O1 is described by

y_{Quad}: O1's observations are accurate and precise and can (in this example) be perfectly described by a 2nd-order polynomial, serving as both forcing and control in the following equation. Its coefficients are chosen such that its initial part exhibits a quasi-linear behavior:

$$y_{Quad}\left(x\right)=a_0+a_1x+a_2x^2;[2] \tag{2.1}$$

here with $a_0=1$, $a_1=-0.025$, and $a_2=0.0025$.[3]

<u>The world of observer O2 is described by</u>

$\boldsymbol{y_{Quad\_wM}}$ **:** y_{Quad} with memory (M). M is chosen by way of assumption (seven years here; justified below) but ensuring that it is shorter than the quasi-linear range of y_{Quad}. Each value of y_{Quad_wM} is constructed as a sum over the seven last values of y_{Quad} (including today), the weights of which decrease exponentially back in time:

$$y_{Quad\_wM}\left(x_k\right)=\sum_{j=0}^{6}e^{-cj}y_{Quad}\left(x_{k-j}\right) \tag{2.2}$$

for $x_k=k$ $\left(k=1,...,35\right)$ and $y_{Quad}=0$ for $x_{k-j}=-5,...,0$ $\left(k-j=-5,...,0\right)$; and with e^{-cj} steering the **weight of memory** (cf. Table 2).[4] The exponential weighting is determined such that its value six years back in time (excluding today) is only 0.05, which reflects our cut-off level (**extent of memory**). That is, only 5% of a six-year old y_{Quad} value contributes to constructing the y_{Quad_wM} value of today ($c=ln0.05/\left(-6\right)=0.50$). The weighting stays constant during the construction of y_{Quad_wM} and is not yet normalized (which we leave for later).

Table 2. The weights of M over seven years back in time (including today).

$-x_j$	$e^{(-cj)}$
0	1.00
-1	0.61
-2	0.37
-3	0.22
-4	0.14
-5	0.08
-6	0.05
Total	**2.47**

[2] To make it easier to follow the fate of our control, we use descriptor type of indices (such as "Quad"). Otherwise, our mathematical terminology is standard (cf., e.g., Wolberg 2006): small letters are used for model-related variables, while capital letters are used for values that are observed (or estimated).

[3] We see the use of the term "full memory" (erroneously for the term "memory-less") as a consequence of fitting a model to the data. Knowing the model's coefficients (three in our case) perfectly well leads us to believe that we know the time series all the way from its beginning (past) to its end (which we may even extrapolate into the future). Perfect knowledge makes it particularly difficult to define memory properly.

[4] The way y_{Quad_wM} operates falls under smoothing, namely, techniques for smoothing time series data. Here, we assign weights to past observations or estimates which decrease exponentially over time. It is noted that smoothing introduces a phase shift into the data (cf., e.g., https://en.wikipedia.org/wiki/Exponential_smoothing).

Three important comments must be made with respect to the definition of y_{Quad_wM} : (1) The exponential weighting appears to be a natural choice. With reference to (what we term) **learning in a diagnostic context**, we see in retrospect that, at the scale of countries, learning (or, conversely, the decrease of uncertainty) in reporting GHG emissions happens exponentially (Hamal 2010; Halushchak *et al.* 2018)—leading us to start out here with exponential weighting as well. (2) The notion of memory in connection with y_{Quad_wM} may not appear straightforward because, ideally, y_{Quad} requires the values of only three points (years) to be entirely determined for all time, all the way from the beginning to the end. On the other hand, we use a memory extent of seven years when we construct y_{Quad_wM} with the help of y_{Quad} . Thus, it may be argued that a finite memory becomes meaningless because each individual point of y_{Quad_wM} carries "full memory." However, the situation changes if y_{Quad_wM} is perceived as the extreme outcome of a thought experiment in which the noise surrounding each point of y_{Quad_wM} eventually decreases to zero (3). It is important to note that how we formalize memory is crucial for how we proceed during the backward mode when we want to quantify persistence without having a priori knowledge of memory and its major characteristics (Chapter 3).

The world of observer O3 is described by

$\boldsymbol{Y_{QwN}}$ **:** y_{Quad} with noise. Y_{QwN} is derived not only by blurring but by obliterating the 2nd-order polynomial character of y_{Quad} by means of great noise, here expressed in relative terms:

$$Y_{QwN}(x) = \left(a_0 + a_1 x + a_2 x^2\right)\left(1 + N u\right) = y_{Quad}(x)\left(1 + N u\right) \tag{2.3}$$

where *N* is a scaling factor and the values u_k are taken randomly from the *u* (standard normal) distribution. The equation describes a parabola with a noise component of $N*100\%$ of the "true" values of y_{Quad} .

In general, we deal with noise in the order of $N \approx 0.10$ (that is, $N*100\% \approx 10\%$).[5] Here, however, we increase N by one order, namely, to $N = 3.0$ (that is, $N*100\% = 300\%$); this may result in Y_{QwN} being perceived as a whole as random noise with some directional drift, if at all, rather than a signal that is clearly visible, albeit superimposed, by noise. The almost complete obliteration of Y_{QwN} is why we argue here that we can freely choose the extent of memory in constructing y_{Quad_wM} (observer O2 above) and Y_{QwN_wM} (observer O4 below).

[5] As a real-world pendant with noise in the order of $N \approx 1$ one may think of, e.g., the net biome production of the terrestrial biosphere.

The world of observer O4 is described by

$\boldsymbol{Y_{QwN\_wM}}$: Y_{QwN} with M (seven years). Y_{QwN_wM} is given by:

$$Y_{QwN\_wM}\left(x_k\right)=\sum_{j=0}^{6} e^{-cj}\, y_{Quad}\left(x_{k-j}\right)\left[1+\left(1-De^{-dj}\right)N\,u_{k-j}\right] \tag{2.4}$$

with $1-De^{-dj}$ steering the **quality of memory** (cf. Table 3). This term is determined in such a way that (i) it allows only 0.05 parts (5%) of random noise for today, meaning that our memory is fairly precise, while (ii) it allows 0.95 parts (95%) of random noise when our memory gets as old as six years (excluding today), meaning that our memory is highly imprecise ($D=0.95$ and $d=ln\left(0.05/0.95\right)/\left(-6\right)=0.49$). Or, if interpreted, for example, in the context of furnace-generated GHG emissions, the contribution of old, still-operating furnaces to today's emissions is not only smaller than that of more recent furnaces, but also less well known. The quality stays constant during the construction of Y_{QwN_wM} and can easily be refined.[6]

Table 3. Quality of M over seven years back in time (including today). The last column shows the interaction of both weight (Table 2) and quality of memory (last column) over time in the case that $a_0=1$, $a_1=a_2=0$, and $N\,u_{k-j}=1$ for all *k* and *j* as specified in the text.

$-x_j$	$De^{(-dj)}$	$1-De^{(-dj)}$	$e^{(-cj)}\left[1+\left(1-De^{(-dj)}\right)\right]$
0	0.95	0.05	1.05
-1	0.58	0.42	0.86
-2	0.36	0.64	0.61
-3	0.22	0.78	0.40
-4	0.13	0.87	0.25
-5	0.08	0.92	0.16
-6	0.05	0.95	0.10
Total	**2.37**		**3.43**

To summarize, in introducing memory we make use of three characteristics: its temporal extent (here dealt with by way of "insightful decision"), its weight over time, and its quality over time. We show in Section 2.2 that memory can, but need not, allow partial reconstruction of what had previously been obliterated.

2.2 An experimental realization

Our mental-numerical set-up allows multiple experiments. A new experiment is launched with a new set of u_k taken randomly from the standard normal distribution, while all other parameters are kept constant.[7] Each experiment consists of two parts: I) construction and

[6] For example, the quality can be made to follow an S-shape of a normalized cumulated distribution function more closely if memory accumulates over time.

[7] These are: $a_0=1$; $a_1=-0.025$; $a_2=0.0025$; $c=0.50$; $D=0.95$; $d=0.49$; $N=3.0$; $k_{min}=1$; $k_{max}=35$; $j_{min}=0$; $j_{max}=6$.

graphical visualization of y_{Quad}, y_{Quad_wM}, Y_{QwN}, and Y_{QwN_wM}; and II) linear regression of the first seven points of Y_{QwN_wM}. The deeper understanding of Part II is (i) that we now split the world with respect to time into two parts, past ($x = 1,...,7$) and future ($x = 8,...,35$), making, in particular, the step from observer O4 who has complete knowledge of his/her world—**the world which we ultimately experience and have to deal with**—to observers (O5 and O6; cf. also lower half of Figure 3) who have incomplete knowledge of that world, namely, of its historical part only (seven years; in accordance with the extent of memory); and (ii) that these observers can (in this example) perceive the historical part of the "O4 world" only by way of linear regression, at best.

Part I: Construction and graphical visualization of y_{Quad}, y_{Quad_wM}, Y_{QwN}, and Y_{QwN_wM}

Figures 4a and 4b show the graphical visualization of an experiment. Figure 4a shows y_{Quad} (orange), y_{Quad_wM} (black), Y_{QwN} (blue) and Y_{QwN_wM} (red); while Figure 4b shows only Y_{QwN} (blue) and Y_{QwN_wM} (red). Dashed lines indicate 2nd-order regressions and their coefficients of determination (R^2) which were determined using Excel.[8] The purpose of showing the 2nd-order regressions *of* y_{Quad}, y_{Quad_wM} and Y_{QwN} in Figure 4a and Y_{QwN_wM} in Figure 4b, along with their R^2-values, is to facilitate understanding. Knowing that our control is a 2nd-order polynomial, these regressions and their R^2-values allows the obliteration of y_{Quad} to be followed by its incomplete reconstruction thereafter.

The experiment is very insightful because it is not (yet) as successful as we wish it to be. As expected, the application of great noise obliterates y_{Quad}. The blue points (Y_{QwN}) do not seem to follow a clear trend. Still, if one wanted to assign a 2nd-order regression to these points just for the sake of it, the regression would exhibit (here) a concave curvature—which would be opposite to the convex curvature of y_{Quad}—and a low R^2-value of 0.005 (cf. also Table 4), confirming the complete obliteration of y_{Quad}.[9]

Y_{QwN_wM} overcomes much of that obliteration, bringing the curvature back to convex and increasing the R^2-value substantially, here to greater than 0.5 (cf. also Table 4).

[8] We are aware that the coefficient of determination for a nonlinear regression exhibits limitations (http://blog.minitab.com/blog/adventures-in-statistics/why-is-there-no-r-squared-for-nonlinear-regression). This is why this coefficient should be understood as a qualitative indicator only, as done here and explained in the text. An appropriate alternative would be the use of the standard error of the regression.

[9] Concave = concave downward (or convex upward); convex = convex downward (or concave upward).

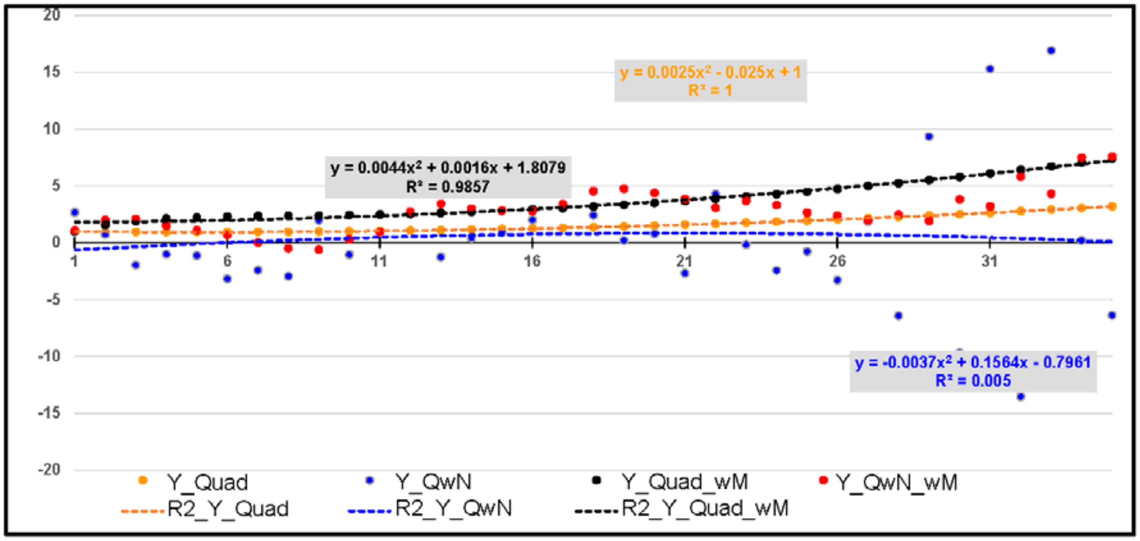


Figure 4a. An experimental realization: y_{Quad} (orange; invariant), y_{Quad_wM} (black; invariant), Y_{QwN} (blue; variable) and Y_{QwN_wM} (red; variable). Dashed lines indicate the 2nd-order regressions and their coefficients of determination $\left(R^2\right)$. Here, the regression of Y_{Quad_wM} falls above the regression of y_{Quad} because we have not yet normalized the coefficients of Y_{Quad_wM} which steer the weight of memory over time.

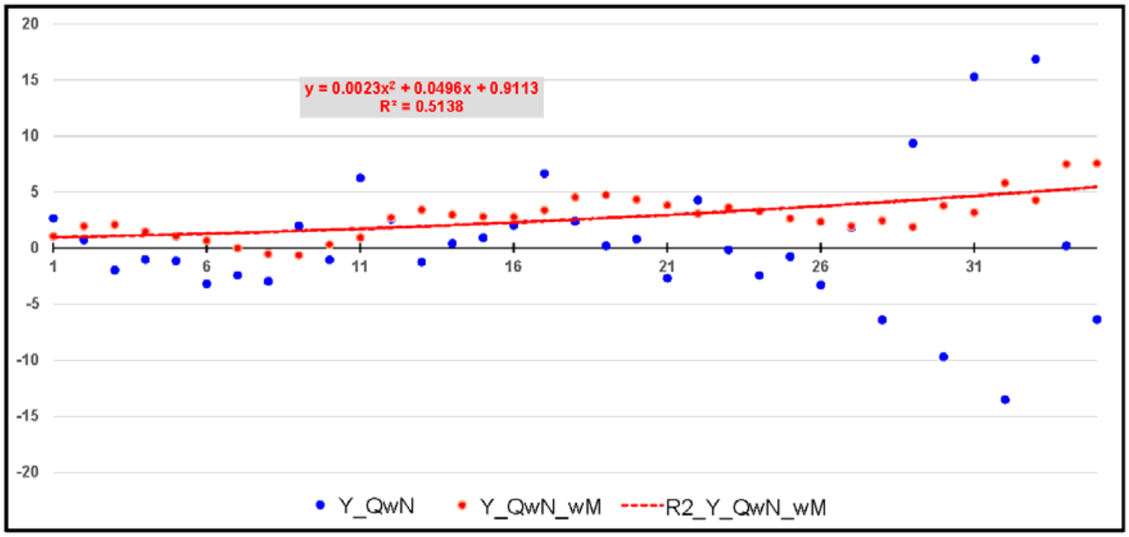


Figure 4b. Like Figure 4a, but allowing for a better overview only Y_{QwN} (blue; variable) and Y_{QwN_wM} (red; variable) with its 2nd-order regression (red solid line).

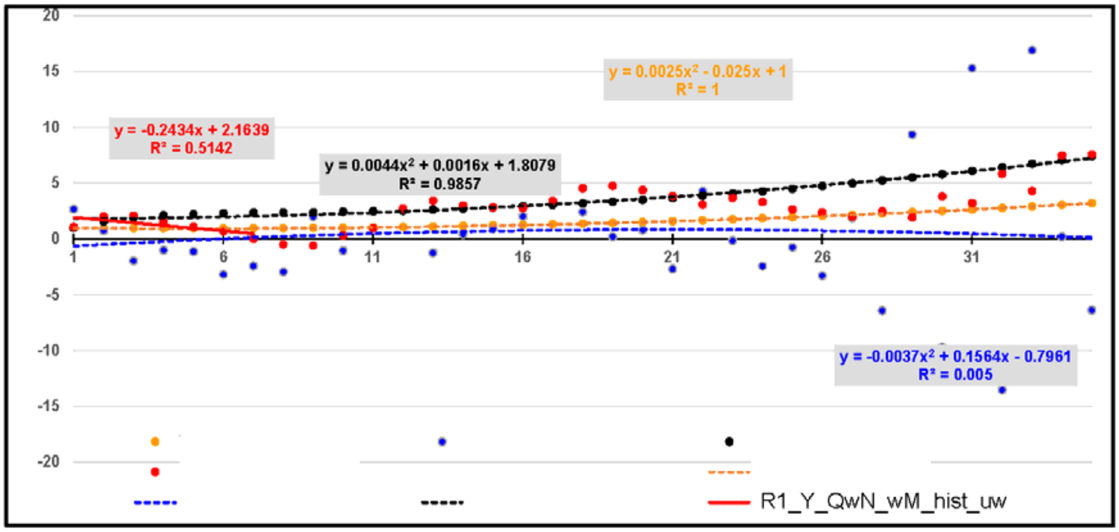


Figure 5a. Like Figure 4a, but additionally showing *R1_Y_QwN_wM_hist_uw*, a linear regression applying unit weighting (uw) back in time for the first seven points of Y_{QwN_wM} (red; variable). The assumption here is that it is only these points (i.e., the extent of memory) of Y_{QwN_wM} that observer O5 knows.

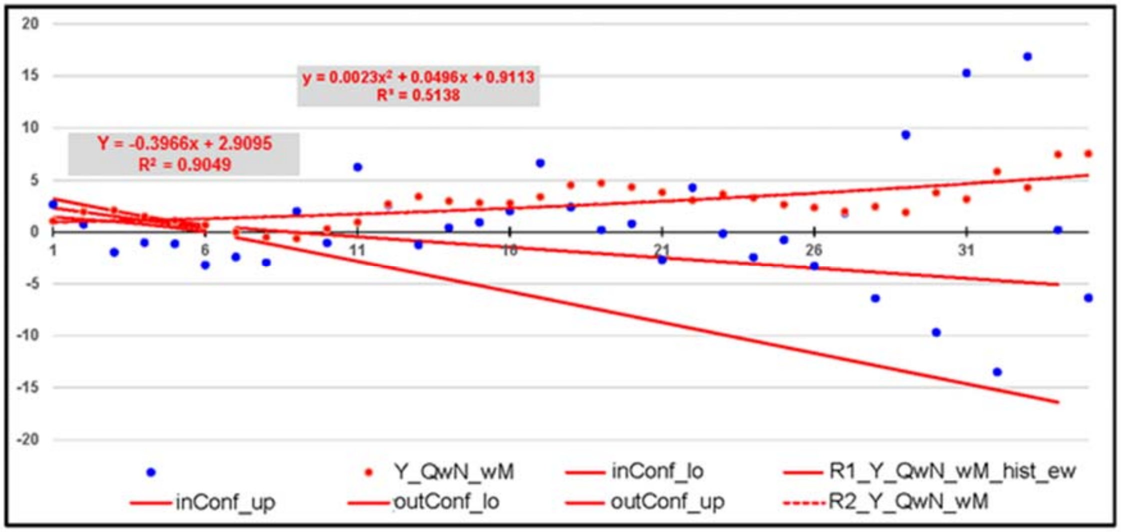


Figure 5b. Like Figure 4b but additionally showing *R1_Y_QwN_wM_hist_ew*, a linear regression applying exponential weighting (ew) back in time for the first seven points of Y_{QwN_wM}, together with its in-sample (inConf) and out-of-sample (outConf) confidence bands. The borders of the confidence bands are indicated by upper (up) and lower (lo). The assumption here is that observer O6 knows, like observer O5, only the first seven points (i.e., the extent of memory) of Y_{QwN_wM} but, in addition, also the weight of memory over time.

Table 4. Supplementing Figures 4 and 5 (while recalling Footnote 8): Compilation of regression parameters and coefficients of correlation, the latter between: (1) y_{Quad} and y_{Quad_wM} (invariant); (2) y_{Quad} and Y_{QwN} (variable); (3) Y_{QwN_wM} and Y_{QwN_wM-7yr} (variable) with Y_{QwN_wM-7yr} being identical to Y_{QwN_wM} but shifted backward in time (year 8 becomes year 1, year 9 year 2, and so on, while dropping the first seven years of Y_{QwN_wM}); and (4) y_{Quad} and Y_{QwN_wM} (variable). The first correlation coefficient indicates that limiting only the extent of memory back in time is not sufficient to overcome the "full memory" of y_{Quad}.[3] Correlation coefficients 2 and 3 seem to confirm that applying a high level of noise completely obliterates the 2nd-order polynomial character of y_{Quad} and that memory does not extend beyond seven years. Finally, correlation coefficient 4 seems to confirm that memory (that is, Y_{QwN_wM}) nullifies much of the obliteration brought about by Y_{QwN}.

Polynomial / Regression for	a_2	a_1	a_0	R^2
y_{Quad}	0.0025	- 0.0250	1.0000	1.000
y_{Quad_wM}	0.0044	0.0016	1.8079	0.9857
Y_{QwN}	- 0.0037	0.1564	- 0.7961	0.005
Y_{QwN_wM}	0.0023	0.0496	0.9113	0.5138
$Y_{Lin,7yr}$	----	- 0.2434	2.1639	0.5142
$Y_{Lin\_exp,7yr}$	----	- 0.3966	2.9095	0.9049
Coefficient of Correlation between				
1) y_{Quad} & y_{Quad_wM}		Influence of memory (w/o noise)		0.99
2) y_{Quad} & Y_{QwN}		Influence of noise (obliteration)		0.02
3) Y_{QwN_wM} & Y_{QwN_wM-7yr}		Influence of memory after 7 yr (w noise)		0.06
4) y_{Quad} & Y_{QwN_wM}		Influence of memory in the presence of noise (reconstruction)		0.71

Part II: Linear regression of the first seven points of y_{QwN_wM}

Figure 5 expands Figure 4 (cf. also lower half of Figure 3). Figure 5a shows a linear regression called *R1_Y_QwN_wM_hist_uw* (in the figure) and $Y_{Lin,7yr}$ (in Table 4) for the first seven points of Y_{QwN_wM} where we assume that it is only these seven points of Y_{QwN_wM} that an observer (O5 hereafter) knows. **It is this assumption—knowing the extent of memory—that requires discussion**. In deriving the linear regression, the seven points are weighted equally (unit weighting (uw) back in time), resulting in a low R^2-value of about 0.51 but, more importantly, in the wrong direction (downward).[10] Note that the overall direction of Y_{QwN_wM} is upward (cf. also Table 4).

By way of contrast, in deriving the linear regression in Figure 5b the first seven points are weighted exponentially (ew) over time. Here, we assume that an observer (O6 hereafter) like

[10] We are aware (indeed, tolerate) that the assumption of independent and identically distributed [IID] random variables underlying linear regression theory may be violated (cf. https://en.wikipedia.org/wiki/Independent_and_identically_distributed_random_variables).

observer O5, knows the first seven points (i.e., the extent of memory) of Y_{QwN_wM} but **also the weight of memory over time—an assumption that also requires discussion**. The exponential weighting (the same as that underlying Y_{QwN_wM}) results in a more confident linear regression called *R1_Y_QwN_wM_hist_ew* (in the figure) and $Y_{Lin\_exp,7yr}$ (in Table 4) with an R^2-value of about 0.90 and an even greater downward trend (-0.40 versus -0.24; cf. Table 4). Figure 5b also shows the confidence bands belonging to $Y_{Lin\_exp,7yr}$ for the first seven years (inConf) and beyond; the latter by means of the out-of-sample (outConf) continuation of the seven-year confidence band. As can be seen, Y_{QwN_wM} crosses the seven-year confidence band from below to above and falls above the out-of-sample confidence band.

The reason for selecting this (unsuccessful), rather than another (successful) experimental realization is to prepare for the next section where we ask if we can make use of repeated regression analyses to capture the immediate future of Y_{QwN_wM}? This will cause the experimental outlook to change from unsuccessful to promising.

2.3 Toward a robust EO

We now repeat the experiment described in Section 2.2 multiple times (cf. also lower half of Figure 3). Table 5 summarizes the results of 100 consecutive experiments where Y_{QwN_wM} falls within the (in-sample and out-of-sample) confidence band of $Y_{Lin\_exp,7yr}$ for a time that corresponds to two times the extent of memory (= 14.5 years in the numerical set-up). These experimental realizations are denoted by “1: Y_{QwN_wM} in”. All other experiments without exception by “0: Y_{QwN_wM} out”. This repetition indicates how often shifting an EO with an extent of seven years to today (here: year 7) is justified, using “one times the extent of memory” as reference for both the shift and the extent of the EO. Table 5 indicates that this is the case in 42% of all experiments.

But we can learn more than just success and failure from the statistics. Table 5 also suggests that the R^2-value of $Y_{Lin\_exp,7yr}$, as well as that of $Y_{Lin,7yr}$, seems to be the right leverage point to differentiate “0-experiments” from “1-experiments.” In the numerical set-up given here, a grouping of experiments depending on whether the R^2-value of $Y_{Lin\_exp,7yr}$ is greater or smaller than 0.50 seems to be a success. This is shown in the fact that the R^2-values of $Y_{Lin\_exp,7yr}$ and those of $Y_{Lin,7yr}$ do not overlap:

$$Y_{Lin\_exp,7yr}: \quad R^2 > 0.50: \quad 0.82 \pm 0.13 = [0.69, 0.95]$$

$$R^2 < 0.50: \quad 0.18 \pm 0.11 = [0.07, 0.29]$$

$$Y_{Lin,7yr}: \qquad R^2: \qquad 0.68 \pm 0.26 = [0.42, 0.94]$$

$$R^2: \qquad 0.19 \pm 0.15 = [0.04, 0.34].$$

Table 5. Summary of results of 100 consecutive experiments where Y_{QwN_wP} falls within the (in-sample and out-of-sample) confidence bands of $Y_{Lin\_exp,7yr}$ for a time that corresponds to two times the extent of memory (= 14.5 yr in the numerical set-up.). These experimental realizations are denoted by "1" (Y_{QwN_wM} in); all others by "0" (Y_{QwN_wM} out); indicating how often it is justified to shift the EO to today (here: year 7).

Grouping of Experiments	**Coefficient of Determination for**			**Coefficient of Correlation for**			**No. of Exp**
	$Y_{Lin\_exp,7yr}$	$Y_{Lin,7yr}$	Y_{QwN_wM}	y_{Quad} & Y_{QwN}	Y_{QwN_wM} & Y_{QwN_wM-7yr}	y_{Quad} & Y_{QwN_wM}	
No grouping	0.58 ± 0.32	0.50 ± 0.31	0.53 ± 0.22	0.10 ± 0.23	0.19 ± 0.35	0.62 ± 0.26	100
0: Y_{QwN_wM} **out**	0.72 ± 0.26	0.60 ± 0.30	0.55 ± 0.22	0.15 ± 0.22	0.20 ± 0.37	0.65 ± 0.27	58
1: Y_{QwN_wM} **in**	0.38 ± 0.30	0.35 ± 0.27	0.50 ± 0.21	0.03 ± 0.22	0.17 ± 0.32	0.59 ± 0.25	42
0: Y_{QwN_wM} **out** and **R**^{2} of $Y_{Lin\_exp,7yr}$ **> 0.30**	0.77 ± 0.19	0.64 ± 0.28	0.56 ± 0.21	0.15 ± 0.21	0.19 ± 0.37	0.67 ± 0.24	53
1: Y_{QwN_wM} **in** and **R**^{2} of $Y_{Lin\_exp,7yr}$ **< 0.70**	0.27 ± 0.22	0.25 ± 0.19	0.50 ± 0.23	0.04 ± 0.23	0.19 ± 0.34	0.61 ± 0.25	34
0: Y_{QwN_wM} **out** and **R**^{2} of $Y_{Lin\_exp,7yr}$ **> 0.50**	0.82 ± 0.13	0.68 ± 0.26	0.54 ± 0.21	0.13 ± 0.20	0.17 ± 0.36	0.66 ± 0.25	48
1: Y_{QwN_wM} **in** and **R**^{2} of $Y_{Lin\_exp,7yr}$ **< 0.50**	0.18 ± 0.11	0.19 ± 0.15	0.50 ± 0.22	0.03 ± 0.23	0.18 ± 0.33	0.61 ± 0.26	27

In addition, Table 5 indicates (while recalling Footnote 8) that the obliteration of y_{Quad} appears to be slightly greater on average for "1-experiments" than for "0-experiments" (cf. coefficients of correlation between y_{Quad} and Y_{QwN}: 0.03 ± 0.23 versus 0.13 ± 0.20). However, it seems that "1-experiments" perform, on average, slightly better from a reconstruction perspective than "0-experiments." In fact, they almost catch up (cf. coefficients of correlation between y_{Quad} and Y_{QwN_wM}: 0.61 ± 0.26 versus 0.66 ± 0.25).

In a nutshell, Table 5 confirms what common sense tells us: **a world perceived too precisely is difficult to "project" even into the immediate future.** Conversely, it is much easier to achieve if we are confronted with a highly imprecise world forcing us to acknowledge our ignorance. It is exactly this insight which tells us that (1) we should avoid following in the footsteps of "perfect forecasting" to derive the EO of a data series (cf. Section 1.2); and (2) we can even derive a robust EO if we resist trying to describe the world we perceive too precisely.

2.4 Pertinent insights and questions

Part I: Insights and questions of systemic nature

Our example raises a number of insights and questions of a systemic and mathematical nature, all of which need to be researched:

1. Is the approach of deriving EOs that deliberately perceives the historical part of a data series imprecisely (by way of linear regression in our example) robust? How imprecisely should we perceive the historical part of the data series?

2. We are confident that we can reduce the problem of studying memory and persistence systemically to, initially, studying single time series, if we allow flexible approaches to capture memory ranging from purely deterministic to purely stochastic, all the while keeping the issue of data availability in mind. In our example, we capture memory (by way of approximation) in terms of extent, weight, and quality, with the latter interacting with the data series' stochastic component. However, different approaches to capturing memory may require EOs to be derived differently.

3. Even if our understanding of how a system is forced is imperfect, we still need to know one (or more?) characteristics of memory—in our example we need to know at least the extent of memory—in order to quantify a system's EO. How well do we need to know/can we know these characteristics in the presence of great noise? How much systems understanding do we need to inject in order to specify all characteristics of memory?

4. Should we consider an upper ceiling for noise? We are aware of concerns that require observations (estimates) of systems to be pre-selected/conditioned so that their noise is ≤ 100% ($N \leq 1$); this is particularly the case for system variables which balance at/around zero under (near-) equilibrium conditions.

5. In our example, we have taken advantage of being able to repeat experiments multiple times—which we may not be able to do in reality. We would have to apply an alternative (e.g., a moving-window technique, where the length of the window coincides with the extent of memory). To start with, can we determine the extent of memory with sufficient precision under great noise? How long must a data series be to allow findings to be achieved that are as robust as those achieved by repetition?

Part II: Insights and questions of mathematical nature

6. Our example underperforms mathematically in various ways, for example: What are the consequences of applying weights of memory that are not (yet) normalized and thus come with a "phase-in" effect? Was it justified to choose the extent of memory freely in the example's forward mode? Is the R^2-value a good measure for robustly differentiating EOs, considering that the historical part of a data series can also be perceived by way of nonlinear regression? Under what conditions is the use of confidence bands more appropriate than the use of prediction bands, or vice versa, to determine the shape of EOs?

7. In our example, to derive a system's EO, we need to at least know the extent of memory. What technique(s) can be applied to determine the extent of memory in the presence of great noise? Can we think of an iterative trial-and-error procedure (including stacking) which would result in "de-noising" and, as a consequence, in determining the extent of memory?

8. Can time series analysis be applied in a flexible way so to allow testing approaches to capture memory, ranging from the purely deterministic to the purely stochastic? In this

context, it is noted that de-trending a time series, as our example shows, is not readily possible without knowing how memory plays out. (We are **not** able to make the step from Y_{QwN} to Y_{QwN_wM} if we do not inject the knowledge of how memory works.) Do other de-trending approaches exist that can be used?

3 Inverse problem: A glimpse into extracting persistence

This chapter is more hypothetical than the previous one. Its purpose is to give a brief overview of the problems that we anticipate in determining persistence without having a priori knowledge of memory and its major characteristics (cf. Section 2.4: Point 3). To this end, we proceed in two steps: the first refers to the deterministic case and the second to the stochastic case.

Case I: *From* y_{Quad_wM} *to* y_{Quad}

Here we assume that we know y_{Quad_wM} and are interested in resolving the two pertinent characteristics of memory (extent and weight) and, if possible, in reconstructing y_{Quad}. To start with, it is worth noting that we should have some a priori, if not a fairly good, understanding of the system under investigation, including the temporal extent of its memory.[11] Figure 6 seems to suggest that the coefficient of correlation between y_{Quad_wM} and y_{Quad_wM} shifted backward in time, designated $Y_{Quad\_wM-i\,yr}$ with $i = 1,...,19$ ($Y_{Quad\_wM\_shifted}$ in the figure), allows the temporal extent of memory to be detected in the vicinity around our/an insightful a priori assumption (here: seven years). The figure shows that the correlation coefficient decreases slowly during the first seven years, that is, as long as memory provides a bond between y_{Quad_wM} and $Y_{Quad\_wM-i\,yr}$ (a consequence of Equation 2.2) and decreases more strongly thereafter.

Being able to determine the temporal extent of memory is already an important first step. However, determining how much past values contribute to today's value is more difficult. We need to know **how** this happens. We recall that we had applied an exponential function to weight memory over time (cf. Equation 2.2). If, and only if, the exponential weighting approach holds—indeed, it would be good to know if this approach even holds in general—we would be able to deduce y_{Quad} value by value, starting at its beginning. The smallest weighting (we had chosen 0.05 as cut-off, leading to $c = 0.50$, the function's exponent) could be dealt with by way

[11] Emission inventory experts typically have very good knowledge of national emissions by source and removals by sink, such as the emissions over time from vehicles and furnaces in use by type and age. They estimate the emissions from these sources annually by applying appropriate emission factors. Our approach to capturing memory in terms of extent, weight, and quality is approximate only, the less so the more we disaggregate the system's emissions by source (such as vehicles and furnaces). The more aggregated an emissions approach we take, the less it refers to a particular emissions source.

of agreement, while the phase-in could be overcome, for example, by recourse to the system's equilibrium (or a quasi-equilibrium) state.

To sum up, we reiterate that the deduction of y_{Quad} will only be possible if the exponential (or another) approach holds of weighting memory back in time. That is, we are left with the challenge of acquiring a deeper systemic understanding to substantiate how memory plays out over time. However, we consider meeting that challenge as feasible.

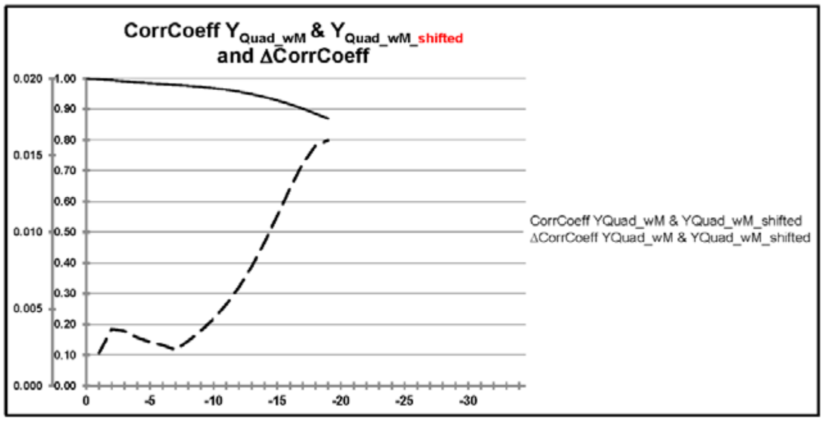


Figure 6. **Solid black line and inner (right) vertical axis:** Coefficient of correlation between Y_{Quad_wM} (invariant) and Y_{Quad_wM} shifted by $i=1,\ldots,19$ years back in time (invariant), designated $Y_{Quad\_wM-i\,yr}$. For instance, $Y_{Quad\_wM-1\,yr}$ is identical to Y_{Quad_wM} but shifted backward by one year (year 2 becomes year 1, year 3 year 2, and so on; while dropping the first year of Y_{Quad_wM}). The correlation coefficient decreases over the range of shifted years shown here. **Dashed black line and outer (left) vertical axis:** The year-to-year difference in the correlation coefficient indicates that this decrease exhibits a local minimum between years -7 and -6 (disregarding the minimum between years -1 and 0 which is an artifact resulting from how the phase-in of Y_{Quad_wM} is currently realized).

Case II: *From* Y_{QwN_wM} *to* y_{Quad}

Here we assume that we know Y_{QwN_wM} and are interested in resolving the three pertinent characteristics of memory (extent, weight, and quality) and, if possible, in reconstructing y_{Quad}. We recall that we are now confronted with random experimental realizations (depending on the u_k which are taken randomly from the standard normal distribution). Figure 7 refers to two such random realizations. Table 6 provides additional information. Figures 7a and 7b are similar to Figure 6 but show the coefficient of correlation between Y_{QwN_wM} and $Y_{QwN\_wM-i\,yr}$ with $i=1,\ldots,19$ ($Y_{QwN\_wM\_shifted}$ in the figure) and the year-to-year change in this coefficient. The figures indicate that: (1) these two quantities, the correlation coefficient, and its year-to-year change become quite variable; and (2) a temporal extent of memory of seven years cannot be as easily identified as in Figure 6. This does not come as a surprise—it is the result of allowing a high level of random noise.

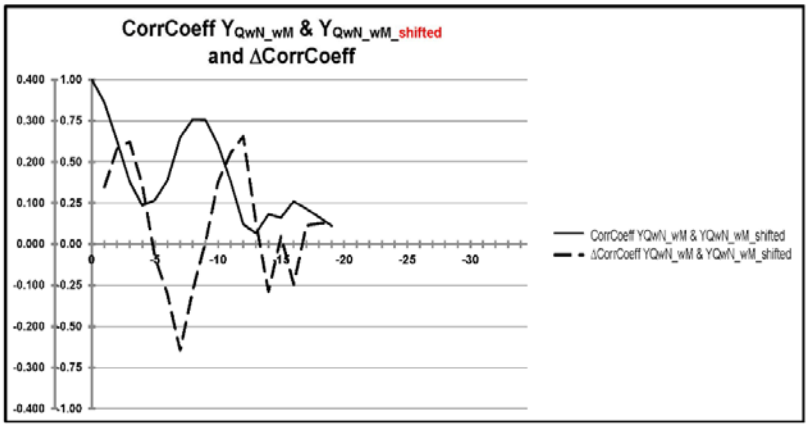


Figure 7a. Like Figure 6 but for the coefficient of correlation between Y_{QwN_wM} (variable) and $Y_{QwN\_wM-i\,yr}$ with $i = 1,\ldots,19$ (variable), and the change in this coefficient. From the perspective of Figure 5b, Y_{QwN_wM} can be described as falling within both the in-sample and the out-of-sample confidence band belonging to $Y_{Lin\_exp,7\,yr}$. For further information see Table 6.

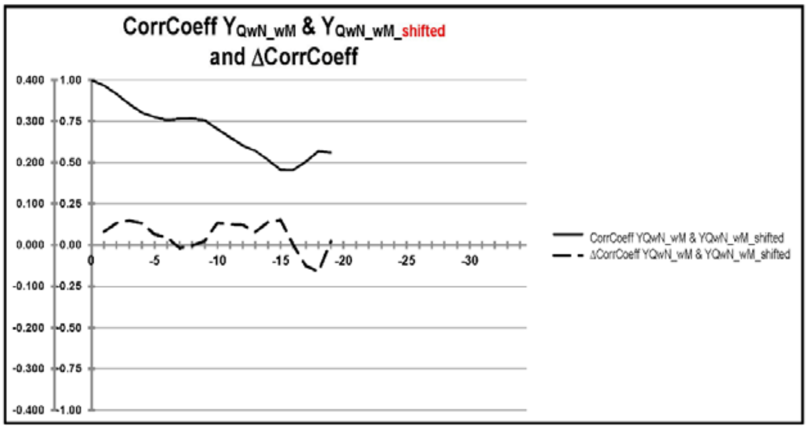


Figure 7b. Like Figure 7a. From the perspective of Figure 5b, Y_{QwN_wM} can be described as falling within and leaving the in-sample confidence band belonging to $Y_{Lin\_exp,7\,yr}$ above, afterwards continuing above its out-of-sample confidence band. For further information see Table 6.

Table 6. Additional information on the experiments underlying Figures 7a and 7b.

Additional Information to	Coefficient of Determination for			Coefficient of Correlation for		
	$Y_{Lin\_exp,7\,yr}$	$Y_{Lin,7\,yr}$	Y_{QwN_wM}	y_{Quad} & Y_{QwN}	Y_{QwN_wM} & $Y_{QwN\_wM-7\,yr}$	y_{Quad} & Y_{QwN_wM}
Fig. 7a	0.0254	0.0645	0.6818[a)]	0.29	0.65	0.82
Fig. 7b	0.9194	0.8827	0.9051[b)]	0.50	0.77	0.95

a) $Y_{QwN\_wM} = 0.0126\,x^2 - 0.2238\,x + 2.7973$; b) $Y_{QwN\_wM} = 0.014\,x^2 - 0.1934\,x + 0.6469$

Figures 7a and 7b reflect special experimental realizations: (1) the obliteration of y_{Quad} is less severe than on average—the coefficient of correlation between y_{Quad} and Y_{QwN} ranges between

0.29 and 0.50 (cf. Table 6 and compare with Table 5); and (2) Y_{QwN_wM} nullifies more of that obliteration than on average—the coefficient of correlation between y_{Quad} and Y_{QwN_wM} ranges between 0.82 and 0.95 (cf. Table 6 and compare with Table 5).

The reason behind this experimental selection is to prepare for a potential way forward. We know that, **in the case of zero noise**, Figures 7a and 7b coincide with Figure 6 (Y_{QwN_wM} coincides with y_{Quad_wM}). By contrast, it appears that **in the case of non-zero noise**, the random, Figure-7-like experimental realizations exhibit a behavior similar to that in Figure 6, on average, especially in the beginning during the first seven years when memory still provides a bond between Y_{QwN_wM} and $Y_{QwN\_wM-i\,yr}$, becoming arbitrarily variable thereafter. That is, it should be possible to overlay many Figure-7-like realizations to identify a behavior like that in Figure 6 and to determine the temporal extent of memory, **not exactly but approximately**. The option of stacking Figure-7-like realizations, however, would require a sufficiently long time series to allow application of a moving-window technique.

As in Case I, knowing the temporal extent of memory is an important step, if not the most important—it allows construction of $\boldsymbol{Y_{Lin,7yr}}$ **, the $\boldsymbol{R^2}$-value of which appears to be an appropriate means of successfully identifying robust EOs** (cf. Table 5). But we are interested in more, namely, in how memory evolves back in time in terms of both weight and quality. We recall that knowing how the weight of memory evolves back in time allows construction of $Y_{Lin\_exp,7yr}$, the R^2-value of which appears to be an even better means of identifying robust EOs (cf. Table. 5).

In our example, the two exponentials that we applied to describe weight and quality back in time are not independent—they share the same j_{max} (cf. Equation 2.4 and Table 3)—allowing us to treat them in combination and proceed as in Case I (meaning that initial and cut-off values determining c, D, and d could be dealt with by way of agreement). Of course, **even with the knowledge of the two exponential functions at hand, it will not be possible to reconstruct $\boldsymbol{y_{Quad}}$**. This is because we do not know the noise component individually at each point in time. Nonetheless, in the case where the two exponential functions can be deduced by systemic insight,[11] it should be possible—while proceeding as in Case I—to "memory-correct" Y_{QwN} point by point while moving forward in time, the best-fit regression of which would exhibit a behavior close to that of y_{Quad_wM} (i.e., ideally also of 2nd-order). As we will know c, D and d only imprecisely at best, we will (ideally) find a set of 2nd-order best-fit regressions. It remains to be seen whether Y_{QwN_wM} will turn out as the mean of that range—a challenge which we leave for later.

To sum up, it seems possible to determine the temporal extent of memory. However, the deduction of y_{Quad} will not be possible; only at best best-fit regressions centering around

y_{Quad_wM} , if the exponential approach holds of describing weight and quality of memory back in time. That is, we are still left with the challenge of acquiring a deeper systemic understanding to substantiate how memory plays out over time (exponentially as here or otherwise). However, we consider meeting that challenge feasible. Disaggregating emissions by source will help to counteract great noise.

4 Should the derivation of EOs become an integral part of model building?

In Sections 2 and 3, our main concern was to illustrate that deriving EOs is possible in principle. Persistence, the counterpart to memory, was shown to serve as a good guide into the near-term future. In this section, we are concerned with the aperture of an EO in particular.

We address the question posed above by comparing a prognostic modeler with a data analyst in terms of how they step across "today," the interface between past and future, and how they treat uncertainty during this process. The comparison shows that the understanding of uncertainty on the part of the modeler and the data analyst differs fundamentally with respect to dimensionality and magnitude. This we illustrate by means of a two-data-series example which we treat analytically under a number of simplifying assumptions. **These, however, neither compromise systemic insight nor restrict generalization**, including adjustment to physical reality. Our three key assumptions are that:

1. each of the two data series—we refer to them as (hypothetical) emission $\left(E=E(t)\right)$ and concentration data $\left(C=C(t)\right)$ the noise of which is "sufficiently" small—exhibits a linear dynamics;
2. the two data series exhibit a serial interdependence, i.e., $C=C(E)$; and that this interdependence is of linear nature (unlike in reality); and
3. memory is minimal,[12] meaning that current concentrations depend on current emissions only and not on past emissions (i.e., they exhibit an interdependence of an instantaneous nature without systemic delays, unlike in reality).

A modeler operating in the t-E-C space

Assumptions 1 to 3 come with the important consequence that, over the diagnostic range, the modeler knows the structure of the C-E model perfectly well—we may also speak of a true model—despite existing uncertainty (resulting from imprecise estimates and/or observations). In building the model, the modeler focuses on the model behaving as accurately as possible from a mean-value perspective. He starts from historic emissions as input

$$E(t)=m_{Et}t\,; \tag{4.1}$$

[12] We prefer to avoid the term memoryless.

derives, with the help of the model (i.e., the equations that it implements; here Equation. 4.1), concentrations as output

$$C(E) = m_{CE} E\,; \tag{4.2a}$$

and compares these against observations

$$C(t) = m_{Ct} t\,; \tag{4.3}$$

where m denotes the signals' linear dynamics (slopes), and indices the space (plane) we are working in. The slopes come with an uncertainty which the modeler captures simplistically, here for emissions:

$$E_{up}(t) = f_{Et,up} m_{Et} t\ ,\ E_{lo}(t) = f_{Et,lo}\, m_{Et} t\,, \tag{4.4a,b}$$

with the constants $f_{Et,up}$ and $f_{Et,lo}$ indicating the upper (up) and lower (lo) borders of the uncertainty in the slope m_{Et}.[13] The difference (Δ) between upper and lower border at any time is given by

$$\Delta E(t) = \Delta f_{Et} m_{Et} t = \Delta f_{Et} E(t) \tag{4.5a,b}$$

with $\Delta f_{Et} = f_{Et,up} - f_{Et,lo}$. To capture prognostic uncertainty, the modeler typically lets all parameters, here m_{CE} and m_{Et}, vary within lower and upper uncertainty borders considered reasonable: here formalized by letting m_{CE} and m_{Et} vary, in accordance with their respective uncertainty, in

$$C(E(t)) = m_{CE} m_{Et} t\ . \tag{4.2b}$$

Uncertainty increases from zero today at $t = 0$ (sufficient for our purposes) to

$$\delta C(E(t)) = (f_{CE} f_{Et} - 1) C(E(t)) \tag{4.6}$$

at time t, which is assumed to be known exactly. This is illustrated in Figure 8. $\delta C(E(t))$ describes a rectangle in the E-C plane at time t.

It is important to note that the modeler understands $\delta C(E(t))$ to encompass

$$\Delta C(t) = \Delta f_{Ct} m_{Ct} t = \Delta f_{Ct} C(t)\,, \tag{4.7a,b}$$

the uncertainty underlying the observations $C(t)$, which come in as the projection of $C(E(t))$ on to the t-C plane in Figure 8. Similarly, the projection of $\Delta C(E)$, the maximal extent of $\delta C(E(t))$ along the C axis, is understood as the uncertainty $\Delta C(t)$ by the modeler

[13] This (reduced) understanding of uncertainty is sufficient for the purposes of this section.

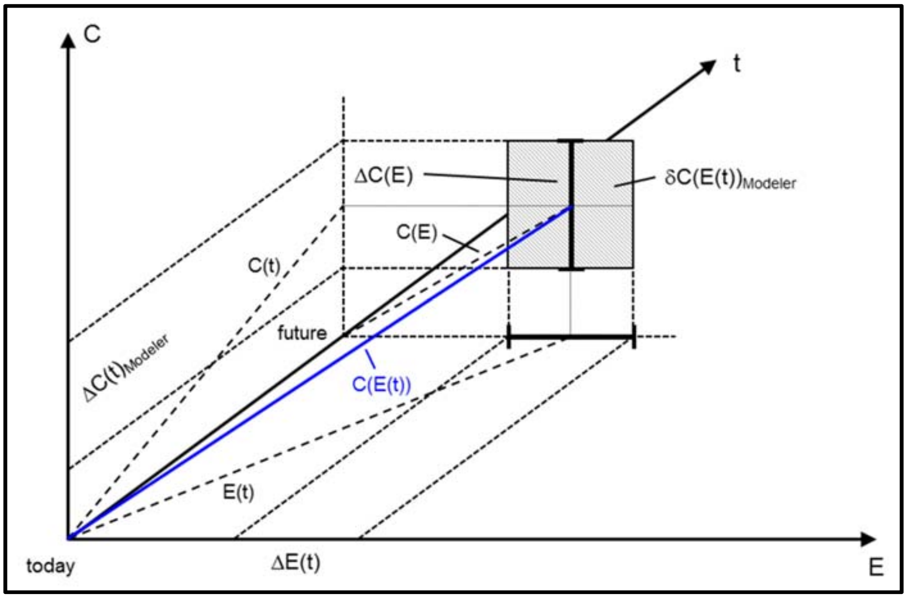


Figure 8: E-C-data-series example to illustrate the uncertainty determined by a modeler (hatched rectangle). This rectangle is spanned by the uncertainty in the slopes m_{Et} and m_{Ct} at time *t* (which is assumed to be known exactly).

A data analyst operating in the t-E-C space

Assumptions 1 to 3 come with the important consequence that, over the diagnostic range, the data analyst knows the memory effect (here without any lag) between GHG emissions and concentrations perfectly well—we may speak of a true "memory model" by analogy—despite existing uncertainty. Knowing the memory model perfectly well entails an important consequence. From Section 2 we know that determining the extent of an EO depends, in particular, on an incomplete understanding of memory and its characteristics. That is, the EO derived on the basis of a perfect memory model always reaches infinitely far into the future. (We will come back to this point at the end of this chapter.) It is important to recognize that, in contrast to the modeler, the data analyst's focus is on capturing both mean value and uncertainty.

With the focus on the first, the data analyst proceeds like the modeler; that is, he starts from historic emissions as input ($E(t)$; Equation 4.1); derives, with the help of the memory model and its individually known characteristics, concentrations as output ($C(E)$; Equation 4.2a); and compares these against observations ($C(t)$; Equation 4.3).[14]

[14] Note that here, deliberately by design, the mean-value perspectives of both the modeler and the data analyst are identical. However, it can be shown (but is not done here) that their mean-value perspectives can differ—and typically do so for multiple reasons. For instance, if memory includes more historical values (i.e., not only today's), the data analyst seeks to capture their contributions to today's value individually, while a modeler would seek to capture their contribution as a whole, possibly by also prioritizing alternative analytical expressions (such as differential equations) which are more convenient from a model-building point of view.

With the focus on uncertainty, we recall that the data analyst does not attempt to capture prognostic uncertainty—he derives the EO from the historical data of a time series and then shifts the EO to its end (= today). To capture the uncertainty underlying $C\big(E(t)\big)$, the data analyst also acknowledges the serial interdependence in uncertainty, here formalized by applying the law of error propagation to Equation (4.2b), finding

$$\Delta f_{CE(t)} = \sqrt{\Delta f_{CE}^{2} + \Delta f_{Et}^{2}} \quad ; \qquad (4.7)$$

where, as before, time is assumed to be known exactly. The meaning of Equation (4.7) is illustrated in Figure 9: The modeler's two-dimensional uncertainty space, a rectangle in the E-C plane at time t, reduces to a one-dimensional space, here given by either one of the rectangle's two diagonals. That is, uncertainty reduces in terms of dimensionality, but it becomes maximal in terms of aperture.

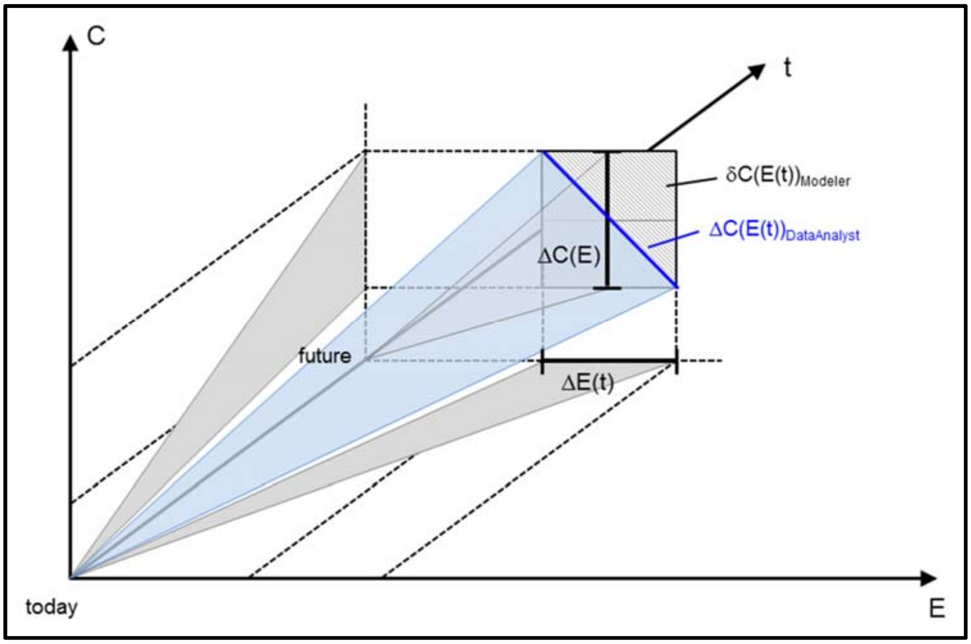


Figure 9: E-C-data-series example to illustrate the uncertainty determined by a data analyst (blue uncertainty wedge). The aperture of this uncertainty wedge is one-dimensional. It is given by either of the two diagonals which span the uncertainty rectangle determined by the modeler.

Figure 9 may lead us to conclude that the maximal uncertainty seen by the modeler coincides with that seen by the data analyst and there is no need for a modeler to process data similarly to a data analyst. However, this conclusion would be misleading. We recall that our two-data-series example involves a number of simplifying assumption:

- It does not consider memory which also includes past values (i.e., prior to today's); and, in consequence, it does not exhibit persistence.
- The uncertainty wedges of the time series involved still reach infinitely far into the future; that is, we are not yet confronted with EOs whose extents are limited because of our

incomplete understanding of memory and its characteristics (which is especially so in the presence of nonlinear data-series interdependences).

We anticipate that unless the modeler adopts the same methods as the data analyst uses to process data, he will not be able to derive EOs and thus unable to appropriately address memory and persistence in the model.

5. Summary and outlook

Our perspective piece seeks to expand on the usefulness of GHG emission inventories. Its intention is to acquire deeper insight into memory, persistence, and explainable outreach. We conjecture that prognostic emission reduction scenarios underestimate the degree and extent of persistence with which GHG emissions will continue on their historical path into the future—which is typical of forced systems with memory—thus, also leading to the amount of reduction that can be achieved in the future being overestimated.

We opted for a holistic perspective piece because this research is generally new, faces a number of open research questions, requires multiple approaches to be tested and described appropriately in scientific journals in the future, and because a reader benefits most from an initial synopsis of insights and conjectures.

Memory allows reference to how strongly a system's past can influence its near-term future. We consider memory to be an intrinsic property of a system, retrospective in nature; and we consider persistence to be a consequential (i.e., observable) feature of memory, prospective in nature and reflecting the tendency of a system to preserve a current state (including trend). Persistence depends on the system's memory which, in turn, reflects how many historical states directly influence the current one. The nature of this influence can range from purely deterministic to purely stochastic.

Different approaches exist to capturing memory. For example, we capture memory with the help of three characteristics: its temporal extent, its weight over time, and its quality over time. The extent of memory quantifies how many historical states directly influence the current state (here, limited by means of a threshold for practical reasons), while the weight of memory describes the strength of this influence. The quality of memory steers how well we know the latter.

In the example, we focus on systemic insight. Our intention is to illustrate one way (among others) of reflecting memory by using a one-data-series perspective to understand how persistence plays out, and to derive a system's EO. The example presented in Section 2 is discussed intensively with respect to how consequential it is, where it underperforms, and the research questions it provokes. However, the example does **not** exhibit fundamental shortfalls and does not restrict generalization. We make use of the example to look into the following three questions:

Q1. How well do we need to know the aforementioned three (and/or possibly other) characteristics of memory in order to delineate a system's near-term future, which we do by means of the system's EO?

Q2. Can we differentiate and specify the various characteristics of memory (e.g., those mentioned above) by way of diagnostic data-processing alone? Or, in other words, how much systems understanding do we need to have and to inject into the data analysis process in order to enable such differentiation?

Q3. Can, or even should, the derivation of EOs become an integral part of model building?

We consider the general insights gained above and beyond the example and the degree of systemic complexity involved in them. We argue the following for each question:

Q1: We have reasons for optimism that the system's EO can be derived under both incomplete knowledge of memory and imperfect understanding of how the system is forced. However, we learn (1) that we should avoid following the footsteps of "perfect forecasting" to derive the EO of a data series; and (2) that we can derive a robust EO even if we resist attempting to describe the world we perceive too precisely.

Q2: Determining the temporal extent of memory in the presence of (possibly great) noise is an important step, if not the most important. But we are interested in more, namely, in how memory evolves back in time in terms of both weight and quality. That is, we are left with the challenge of acquiring a deeper systemic understanding to substantiate how memory plays out over time (exponentially, as in our example, or otherwise). However, we consider meeting that challenge to be feasible.

Q3: We broadened our initial example further in Section 4 (switching perspectives from one to two data series), while simplifying it systemically (linearizing both the dynamics of the data series and their serial interdependence, reducing our uncertainty perspective, and assuming memory to be minimal). This we did to compare a prognostic modeler with a data analyst in terms of how they step across "today," the interface between past and future, and how they treat uncertainty during this process. This comparison is complementary to our investigations above with the focus on the degree and extent of persistence in stepping across "today." It shows that the modeler's and the data analyst's understanding of uncertainty differs fundamentally concerning dimensionality and magnitude. We conclude that, without adopting the way of how the data analyst processes data, the modeler will not be able to derive EOs and, in consequence, address memory and persistence appropriately in the model.

Although the prime intention of our perspective piece to expand on the usefulness of GHG emission inventories in terms of memory, persistence, and explainable outreach, our insights indicate the high chance of our conjecture proving true: being ignorant of memory and persistence, we underestimate, probably considerably, the momentum with which GHG emissions will continue on their historical path beyond today and thus overestimate the amount of reductions that we might achieve in the future.

References

Barkoulas, J.T., and C.F. Baum, 1996: Long-term dependence in stock returns. *Econ. Lett.*, **53**, 253–259, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.470.3891&rep=rep1&type=pdf.

Barros, C.P., L.A. Gil-Alana and P. Wanke, 2016: Brazilian airline industry: Persistence and breaks. *Int. J. Sustain. Transp.*, **10**(9), 794–804, doi: 10.1080/15568318.2016.1150533.

Belbute, J.M., and A.M. Pereira, 2017: Do global CO_2 emissions from fossil-fuel consumption exhibit long memory? A fractional integration analysis. *Appl. Econ.*, doi: 10.1080/00036846.2016.1273508.

Brockwell, P.J. and R.A. Davis, 2002: Introduction to Time Series and Forecasting. Springer-Verlag, New York, NY, United States of America, pp. 434.

Caballero R., S. Jewson and A. Brix, 2002: Long memory in surface air temperature: Detection, modeling, and application to weather derivative valuation. *Clim. Res.*, **21**, 127–140, doi: 10.3354/cr021127.

Chow, K.V., K.C. Denning, S. Ferris and G. Noronha, 1995: Long-term and short-term price memory in the stock market. *Econ. Lett.*, **49**, 287–293, doi: 10.1016/0165-1765(95)00690-H.

Dajcman S., 2012: Time-varying long-range dependence in stock market returns and financial market disruptions – a case of eight European countries. Appl. Econ. Lett., 19(10), 953–957, doi: 10.1080/13504851.2011.608637.

DW, 2015: Draft climate text released - but what about the emission gap? Deutsche Welle, Bonn, Germany, 05 October 2015, http://dw.com/p/1Giw2 (accessed 13 Sep. 2017).

Franzke C., 2010: Long-range dependence and climate noise characteristics of Antarctic temperature data. J. *Climate*, **23**(22), 6074–6081, doi: 10.1175/2010JCLI3654.1.

Halushchak, M., M. Jonas, P. Żebrowski, J. Jarnicka, R. Bun, Z. Nahorski and E. Rovenskaya, 2018. (Manuscript describing how learning can be distilled from revised greenhouse gas emission time series; under preparation).

Hamal, K., 2010: Reporting GHG Emissions: Change in Uncertainty and its Relevance for the Detection of Emission Changes. Interim Report IR-10-003, International Institute for Applied Systems Analysis, Laxenburg, Austria, pp. 34, http://webarchive.iiasa.ac.at/Publications/Documents/IR-10-003.pdf.

IPCC, 2001: *Climate Change 2001: Synthesis Report. A Contribution of Working groups I, II and III to the Third Assessment Report of the Intergovernmental Panel on Climate Change* [R.T. Watson and the Core Writing Team (eds.)]. Cambridge University Press, Cambridge, United Kingdom, and New York, NY, United States of America, pp. 398.

Hansen P.R. and A. Lunde, 2014: Estimating the persistence and the autocorrelation function of a time series that is measured with error. *Economet. Theor.*, **30**(1), 60–93, doi: 10.1017/S0266466613000121.

Jonas, M., G. Marland, W. Winiwarter, T. White, Z. Nahorski, R. Bun and S. Nilsson, 2010a: Benefits of dealing with uncertainty in greenhouse gas inventories: Introduction. Clim. Change, 103(1–2), 3–18, doi: 10.1007/s10584-010-9922-6.

Jonas, M., M. Gusti, W. Jęda, Z. Nahorski and S. Nilsson, 2010b: Comparison of preparatory signal analysis techniques for consideration in the (post-) Kyoto policy process. Clim. Change, 103(1–2), 175–213, doi: 10.1007/s10584-010-9914-6.

Supporting online material: (1) Mathematical background and numerical tables (pp. 26; Doc file); (2) Numerical results (Excel file). International Institute for Applied Systems Analysis, Laxenburg, Austria, http://webarchive.iiasa.ac.at/Research/FOR/unc_prep.html.

Jonas, M., G. Marland, V. Krey, F. Wagner and Z. Nahorski, 2014: Uncertainty in an emissions-constrained world. *Clim. Change*, **124**(3), 459–476, doi: 10.1007/s10584-014-1103-6.

Joshi, P., 2016: Measuring the memory of time seris data. Perpetual Enigma, Blog, https://prateekvjoshi.com/2016/11/30/measuring-the-memory-of-time-series-data/ (accessed 12 Sep. 2017).

Kantelhardt, J.W., 2004: Fluktuationen in komplexen Systemen. Habilitationsschrift, Justus Liebig University Giessen, Germany, pp. 217, http://www.physik.uni-halle.de/Fachgruppen/kantel/habil.pdf.

Kantelhardt J.W., E. Koscielny-Bunde, D. Rybski, P. Braun, A. Bunde and S. Havlin, 2006: Long-term persistence and multifractality of precipitation and river runoff records, *J. Geophys. Res.*, **111**(D01106), doi: 10.1029/2005JD005881.

Lennartz, S. and A. Bunde, 2009: Trend evaluation in records with long-term memory: Application to global warming. *Geophys. Res. Lett.*, **36**, L16706, doi: 10.1029/2009GL039516.

Lennartz, S. and A. Bunde, 2011: Distribution of natural trends in long-term correlated records: A scaling approach, *Phys. Rev. E*, **84**, 021129, doi: 10.1103/PhysRevE.84.021129.

Lieberman, D., M. Jonas, Z. Nahorski and S. Nilsson (eds.), 2007*: Accounting for Climate Change. Uncertainty in Greenhouse Gas Inventories – Verification, Compliance, and Trading*. Springer, Dordrecht, Netherlands, pp. 159.

Lo, A.W., 1991: Long-term memory in stock market prices. *Econometrica*, **59**(5), 1279–1313, http://www.jstor.org/stable/2938368.

Lüdecke H.J., A. Hempelmann and C.O. Weiss, 2013: Multi-periodic climate dynamics: Spectral analysis of long-term instrumental and proxy temperature records. *Clim. Past*, **9**, 447–452, doi: 10.5194/cp-9-447-2013.

Majunmdar, S.N. and D. Dhar, 2001: Persistence in a stationary time series. *Phys. Rev. E*, **64**, 046123, doi: 10.1103/PhysRevE.64.046123.

Mudelsee, M., 2010: *Climate Time Series Analysis. Classical Statistical and Bootstrap Methods.* Springer, Dordrecht, Netherlands, pp. 474.

Ometto, J.P., R. Bun, M. Jonas and Z. Nahorski (eds.), 2015: *Greenhouse Gas Inventories: Expanding Our Uncertainty Perspective*. Springer, Dordrecht, Netherlands, pp. 240.

Palma, W., 2007: *Long-Memory Time Series. Theory and Methods.* John Wiley & Sons, Inc., Hoboken, New Jersey, United States of America, pp. 285.

Wikibooks, 2017: *Signals and Systems*. https://upload.wikimedia.org/wikipedia/commons/6/67/Signals_and_Systems.pdf (accessed 15 Nov. 2017).

Wikipedia, 2017: Complex system. https://en.wikipedia.org/wiki/Complex_system (accessed 15 Nov. 2017)

Wolberg, J., 2006: Data Analysis Using the Method of Least Squares. Extracting the Most Information from Experiments. Springer, Berlin, Germany, pp. 250.

Acronyms and Nomenclature

A	maximal EO aperture
C	concentration
E	emissions
EO	explainable outreach
ew	exponential weighting
GHG	greenhouse gas
IIASA	International Institute for Applied Systems Analysis
IID	independent and identically distributed
inConf	in-sample confidence band
IPCC	Intergovernmental Panel on Climate Change
L	EO length
lo	lower
m	slope
M	memory
O	observer
outConf	out-of-sample confidence band
R	resolution
t	time
up	upper
uw	unit weighting
δ	variable difference ($0 \le \delta \le \Delta$)
Δ	difference