Weather and Climate Extremes 20 (2018) 9–32

Contents lists available at ScienceDirect

Weather and Climate Extremes

journal homepage: www.elsevier.com/locate/wace

Upgrade of the HadGEM3-A based attribution system to high resolution and a new validation framework for probabilistic event attribution

Andrew Ciavarella [a,*], Nikos Christidis [a], Martin Andrews [a], Margriet Groenendijk [b], John Rostron [a], Mark Elkington [a], Claire Burke [a], Fraser C. Lott [a], Peter A. Stott [a,b]

[a] Met Office Hadley Centre, FitzRoy Road, Exeter, EX1 3PB, UK
[b] Department of Mathematics, University of Exeter, EX4 4QF, UK

A B S T R A C T

We present a substantial upgrade of the Met Office system for the probabilistic attribution of extreme weather and climate events with higher horizontal and vertical resolution (60 km mid-latitudes and 85 vertical levels), the latest Hadley Centre atmospheric and land model (ENDGame dynamics with GA6.0 science and JULES at GL6.0) as well as an updated forcings set. A new set of experiments designed for the evaluation and implementation of an operational attribution service are described which consist of pairs of multi-decadal stochastic physics ensembles continued on a season by season basis by large ensembles that are able to sample extreme atmospheric states possible in the recent past. Diagnostics from these experiments form the HadGEM3-A contribution to the international Climate of the 20th Century Plus (C20C+) project and were analysed under the European Climate and Weather Events: Interpretation and Attribution (EUCLEIA) event attribution project as well as contributing to the Climate Science for Service Partnership (CSSP)-China programme.

After discussing the framing issues surrounding questions that can be asked with our system we construct a novel approach to the evaluation of atmosphere-only ensembles intended for event attribution, in the process highlighting and clarifying the distinction between hindcast skill and model performance. A framework based around assessing model representation of predictable components and ensuring exchangeability of model and real world statistics leads to a form of detection and attribution to boundary condition forcing as a means of quantifying one degree of freedom of potential model error and allowing for the bias correction of event probabilities and resulting probability ratios. This method is then applied systematically across the globe to assess contributions from anthropogenic influence and specific boundary conditions to the changing probability of observed and record seasonal mean temperatures of four recent 3-month seasons from March 2016–February 2017.

1. Introduction

Event attribution is the emerging science of establishing when human influence on weather and climate events can be discerned and quantifying the changing likelihood of their occurrence (Stott et al., 2016). From conception event attribution has oriented itself toward various stakeholder groups in wider society (Allen, 2003) and the requirement for an operational service has been perceived for some time (Stott et al., 2013) with the intention of providing calibrated, objective information from a well validated modelling system on a regular basis. Prototyping of a system based on ensemble realisations of the UK Met Office physical climate model HadGEM3-A at seasonal forecast resolution began some years ago (Christidis et al., 2013a) and here we present a major upgrade of this system's dynamical core at higher horizontal and vertical resolution together with a novel means of validation and calibration together with case studies.

Event attribution grew out of formal climate change detection and attribution studies (Jones et al., 2013) in which causal influences resulting in climate forcings are separated into sets which are prescribed to a physical model to simulate worlds responding to isolated sets of influences. Typically, naturally occurring historical forcings (due to solar variability and volcanic eruptions) are separated out to produce realisations of a natural forcings-only world (NAT) while simulations including both naturally occurring and historical anthropogenic forcings form realisations of the all forcings (ALL) world. At root all event attribution studies concerning anthropogenic influence comprise a comparison of the likelihoods p_0 and p_1 respectively of an event occurring under these different scenarios. Whereas formal detection and attribution uses information integrated across the historical period to quantify the contribution of different influences to historically observed trends the science of event attribution quantifies the change in likelihood of transient climate features - *events* - which in principle may be defined to be of arbitrary complexity, duration and spatial extent.

The Met Office system for the attribution of extreme weather and

* Corresponding author.
E-mail address: andrew.ciavarella@metoffice.gov.uk (A. Ciavarella).

https://doi.org/10.1016/j.wace.2018.03.003
Received 24 August 2017; Received in revised form 9 March 2018; Accepted 15 March 2018
Available online 30 April 2018

climate events is based on the global atmospheric component of the Unified Model at a resolution already used operationally for seasonal forecasting (MacLachlan et al., 2015), running pairs of large ensemble ALL and NAT experiments simulating recent periods on a repeating seasonal basis. By providing the atmospheric model with observed sea surface temperature and sea ice boundary conditions the system can be used to ascertain estimates of the likelihood of events unfolding under the same conditions as those in which a real event occurred in the recent past, or as it could have occurred under similar conditions in the NAT world. As an operational activity event attribution could also be considered an essential form of seasonal monitoring of human influence on the climate in near real time.

The novelty of the HadGEM3-A system lies in it state of the art science run at a high resolution, only beaten by regional models nested within coarser global counterparts. The experiments we have run with the system are contributing to several international collaborative projects concerning event attribution and wider climate science. Development took place within EUCLEIA (EUropean CLimate and weather Events) which evaluated a range of the system's physical mechanisms (Vautard et al., 2018) relevant to extremes over Europe and produced detailed case studies exploring the capability of the HadGEM3-A based system to simulate extreme events of diverse types and spatiotemporal extent (Wilcox et al., 2017; Christiansen et al., 2017; Climate science for service partnership china). Authors are also making use of the experiments under C20C+ (C20C+ Detection and Attribution Project) and CSSP-China (Climate science for service partnership china) and a number of articles have already appeared or are in preparation which use the data (Christidis et al., 2016; Burke et al., 2016; Qian et al., 2018; Angélil et al., 2016, 2017; Dunn et al., 2017; Wehner et al., 2017; Eden et al., 2016), highlighting the value to the community of Met Office involvement before the operational system is even delivered.

Strong emphasis has been placed on model evaluation (Stott et al., 2013; Christidis et al., 2013a) because confident attribution statements can only be made when a model is able to reproduce the defining features of an event at the right time. The causal influences acting on any given event involve complex hierarchies of response to external forcings, teleconnections to important modes of variability and the immediate physical and dynamical processes leading to it, all of which should in principle be captured by the model when required. Nevertheless an attribution system is different from a forecasting system, in which any deviation from the real world is forecast uncertainty (before the event) and model error (after the event). In an attribution system the features of an event that must be predictable depend upon the conditioning of the attribution question (Otto et al., 2012), an issue referred to now as the framing issue.

Attribution ensembles sample a space of solutions sharing conditioning, first of all on equal external forcings: each member of the ensemble represents an equally likely outcome under the same forcing conditions. To assess the importance of a given causal influence on an event we construct probabilities from the ensemble which involves subsampling this solution space (typically by subsampling the ensemble itself) to further condition on solutions sharing special features such as the phase of some mode, but also on event criteria such as mean values over some spatiotemporal scale. An example could be an ensemble of all members sharing the same historical external forcing, a positive projection onto an atmospheric mode of variability and probabilities determined by partitioning this set into those sharing event criteria of monthly mean precipitation totals above or below some threshold. In event attribution we normally then construct probability ratios from the two forcings scenarios, ALL and NAT, which can further be thought of as conditioned solutions in a larger space of all possible forcing scenarios.

The point is that the relationship between the resulting probability ratios and the causal influences we are concerned with can be entirely dependent upon the selected conditions, the framing. An example of immediate relevance to the system described in this article is conditioning of our ensembles on observed historical boundary conditions: event likelihoods will contain a degree of predictability through time that would not be present in solutions of a coupled model on account of their shared boundary conditions. Aspects of this freedom in framing of the attribution question are being explored in parallel (Christidis et al., 2018) and we place an equal emphasis on clarity of the framing in this study.

Evaluation of an ensemble system's ability to produce accurate probabilities, involving assessment of the timing of changes from climatological values, broadly goes by the name of model "skill". To continue with the above example, in moving from an ensemble of coupled ocean-atmosphere simulations (sampling all oceanic states) to atmosphere-only simulations (conditioned to sample solutions which see a single realisation of the oceanic state) we will see the appearance of skill in the ensemble. The choice of using coupled or atmosphere-only simulations however is framing. It is obvious then that skill is not automatically a requirement of an attribution ensemble but that we must understand when skill is required both as a function of the degree of predictability inherent in the climate system and when that predictability is hidden or revealed through framing. Only when we know if skill is present when it is required can we make statements about model performance in terms of hindcast skill.

In light of the volume of work already being performed on the mechanistic side of the evaluation of the system we will focus our validation on addressing this question by introducing a means of assessing changes hindcast by the model on interannual time scales as distinct from those occurring on longer time scales. We limit the analysis to seasonal mean near-surface air temperatures as an example in which we expect there to be some degree of interannual predictability due to the boundary conditions imposed by the system.

In the following section we describe the system's technical and scientific setup, followed by the experiments designed for its validation and the production of large ensembles on a seasonal basis. In Section 3 we clarify the framing issues within our system and further discuss the relationship between an attribution system and ensemble forecast systems as well as describing the different ways we shall frame the exceedance probability ratios p_1/p_0 to be presented in Section 6. Section 4 sets out a novel validation framework in which we first assess the degree of predictability present in an ensemble and hence then assess model performance according to the presence of skill where it is required. In the same section we suggest a form of bias correction similar to that used in seasonal forecast systems but appropriate to event attribution which is concerned with pairs of ensembles.

Our results begin with the application of the validation presented in Section 5. In Section 6 we then select a handful of case studies from a systematic global analysis to which the validation and bias correction have been applied and with which we can assess the change in likelihood of events under anthropogenic influence, the probability ratio, through several different framings of the event attribution question. In particular we estimate the change in probability ratio due to the specific boundary conditions occurring over the last year compared to what that probability ratio would be subject only to long term changes occurring in the climate system. These case studies could be considered as brief rehearsals of systematic assessments output from an operational system. We end with a discussion.

2. System description

We conduct experiments consisting of pairs of ensembles that differ through the external climate forcings included, one with both natural and anthropogenic forcings present (ALL) and the other with only natural historical forcings, all others being held fixed at 1850 levels (NAT). Natural external forcings are, firstly, variability in total solar irradiance at the top of the atmosphere, and secondly volcanic activity represented through a latitudinal variation of stratospheric aerosol optical depth. Other external forcings provided are well-mixed green house gases (GHG) and zonal-mean ozone concentrations, aerosol emissions and land use change, which has been shown to affect the occurrence of daily

temperature extremes (Christidis et al., 2013b). The prescription of lower boundary conditions (the sea surface temperatures and sea ice coverage) also integrate historical external forcings as well as communicating historical, internally generated modes of ocean and coupled variability to the atmosphere.

The core atmospheric model, land model, initial conditions and method of ensemble generation remain identical between experiments. The attribution system can therefore be visualised as in Fig. 1 where the core atmosphere and land model which underlie a diversity of Met Office systems are supplemented by inputs specific to the event attribution system. Below we will describe these system components in some detail and the attribution experiments conducted thus far.

2.1. High resolution global atmosphere and land components

Scientific configurations of Met Office global coupled and atmospheric models are described by their Global Coupled (GC) and Global Atmospheric (GA) number. The new attribution system was developed using the GA6 atmospheric science package (Walters et al., 2016). GC2 and GA6 science are currently operational across Met Office Numerical Weather Prediction and climate systems.

Two significant upgrades to the previous HadGEM3-A based system are to the non-hydrostatic dynamical core of the model and the resolution. GA6 uses the ENDGame dynamical core (Wood et al., 2014) while the previous version used New Dynamics (Davies et al., 2005), bringing improvements in atmospheric dynamics that include synoptic scale features such as extra-topical storms. Resolution has increased from N96 L38 to N216 L85. This refers to a horizontal latitude/longitude grid that is 2N cells East-West by 1.5N cells North-South where $N = 216$, which gives $0.83^{\circ} \times 0.56^{\circ}$ angular resolution equivalent to around 60 km at mid-latitudes. There are 85 vertical levels: 50 tropospheric and 35 stratospheric. A realistic representation of stratosphere-troposphere interactions gives the Met Office system an advantage over those lacking a proper stratosphere, for e.g. when addressing European cold events whose likelihood is influenced by the strength of the stratospheric vortex.

The land surface and hydrology schemes are also upgraded from the previous system, which used the MOSES-II model, to JULES (Best et al., 2011; Clark et al., 2011) (Joint UK Land Environment Simulator) version 6.0, a community land surface model. This handles fluxes of heat, moisture and gases between the atmosphere and land, surface hydrology as well as deep soil processes through 4 sub-surface layers with thickness of 0.1 m, 0.25 m, 0.65 m and 2.0 m in descending order. JULES assigns fractions of 9 surface types (on "tiles") to each grid cell of which 5 are vegetation functional types with seasonally modulated leaf area and canopy height parameterizations. The non-plant types are: land water, land ice, bare soil and urban.

2.2. External forcings prescription

Horizontal boundary conditions at the bottom of the atmosphere are given by series of Sea Surface Temperatures (SST) and Sea Ice (SIC) fields. The ALL experiments take these from observed values of the HadISST1 dataset (Rayner et al., 2003) which provides gridded monthly values interpolated across regions of missing data. The NAT experiments are provided with SST and SIC fields equal to those provided for ALL but from which an estimate of the changes due to anthropogenic influence, ΔSST and ΔSIC, have been removed (Christidis et al., 2013a). ΔSST are calculated as the difference between estimates using coupled model simulations of the ALL and NAT scenarios. Our system currently uses CMIP5 (Taylor et al., 2012) multi-model mean fields for this purpose. ΔSIC are produced through an empirical relationship between historically observed polar SST and SIC. The precise implementation of this method that we have used is described in detail elsewhere (Stone and Pall, 2017).

These boundary conditions are prescribed as monthly data using the AMIP II method (Taylor et al., 2000), involving pre-processing to avoid reduction of variance in monthly and seasonal means that result when the model interpolates to daily values at run time.

Well-mixed GHG concentrations are prescribed as series of annual values obtained from the RCP Scenario Data Group (CMIP5 recommended data, 2013) which are historical values up to 2005 and following the RCP4.5 scenario (Meinshausen et al., 2011) beyond 2005. Five gases are prescribed: CO2, CH4, NOx, and two sets of fluorocarbon equivalents, CFC12 equivalent and HFC134-A equivalents. Estimates of these concentrations during the historical period have not changed from those supplied to the original HadGEM3-A system.

Emissions of aerosols, handled by the CLASSIC scheme, are prescribed as CMIP5 recommended monthly values. The list of aerosol types and specific radiative effects included within the model remain unchanged from the previous system but the data sets have been updated.

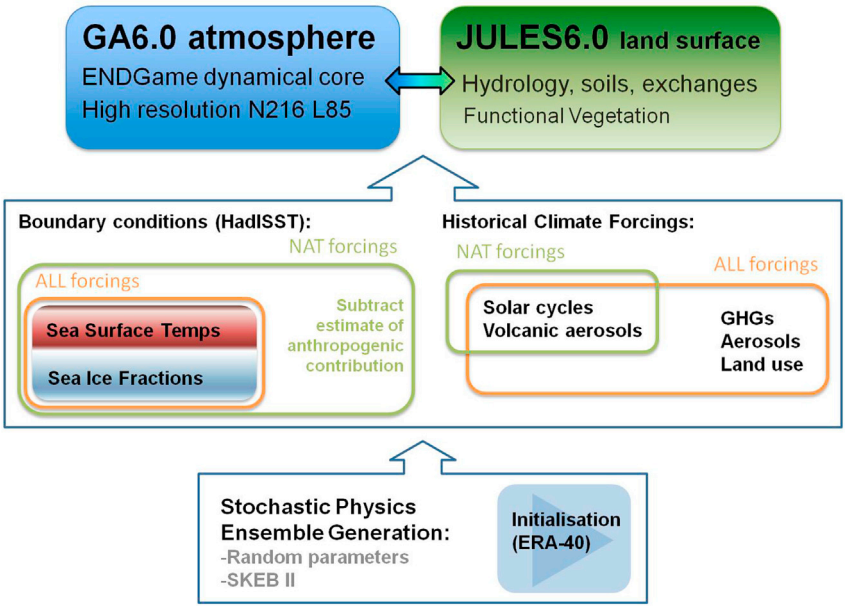


Fig. 1. The HadGEM3-A based event attribution system. The high resolution atmospheric dynamical core and land model are supplied with boundary conditions and forcings specific to this system. Ensembles are generated through stochastic physics. Details are given in the main text.

Anthropogenic precursor gases and aerosol emissions of the following types are included, with the indicated radiative effects in short wave (SW) and long wave (LW) radiation (see the references for source data). Sulphates: high and low level SO2 + surface dimethyl sulphide (Steven et al., 2001; Smith et al., 2004) Direct (SW + LW) + 1st (SW + LW) and 2nd indirect radiative effects. Soot (Bond et al., 2007): Direct (SW + LW) radiative effects. Organic Carbon Fossil Fuels (OCFF) (Bond et al., 2007) Direct (SW + LW) + 1st indirect (SW + LW) and 2nd indirect radiative effects. Biomass (Lamarque et al., 2010) Direct (SW + LW) + 1st (SW and/or LW) and 2nd indirect radiative effects.

Additionally, mineral dust and sea salt are modelled interactively at run time while biogenic aerosols are included via a 12-month climatology.

Ozone (Cionni et al., 2011) is prescribed as monthly 2-dimensional fields of zonal mean values across the 85 model levels. The values are historical up to 2005 and follow RCP4.5 thereafter.

Land use forcing is prescribed once per decade as annual fields of tile type fractions described above. Fig. 3 displays hemispheric mean fractions over the experimental period for both ALL and NAT experiments. The model interpolates in time between these fields and subjects five of them (the vegetation functional types) to monthly modulation within JULES. This forcing is much improved from the data supplied to the previous system. The source is the ISAM-HYDE dataset (Meiyappan and Jain, 2012), which is the HYDE3.1 dataset (Klein Goldewijk et al., 2011) harmonized with satellite-derived (MODIS) estimates. The ISAM-HYDE source data consists of fractions of 28 land surface types. Prior to run time these are mapped via 18 IGBP types to the 9 MOSES-II types used by the JULES model. The values prescribed at 2020 are a repeat of those at 2010 so (annually averaged) land use forcing is constant between these dates.

Natural forcings due to solar variability and aerosols from volcanic activity are included via branch modifications to the UM code itself, feeding time series data to the radiation scheme. Total solar irradiance (TSI) is partitioned into 6 short-wave spectral bands and supplied as monthly global mean values. TSI is historical up to 2009 after which an idealised cycle is used. The data (Lean, 1882) is that used to force the HadGEM2-ES CMIP5 historical simulations. Volcanic stratospheric aerosol optical depth (AOD) is supplied as monthly mean values for 4 equal area latitudinal bands and the data (Sato et al., 2006) is also that supplied to the HadGEM2-ES CMIP5 historical simulations. Global mean series of TSI and stratospheric AOD are displayed in Fig. 2.

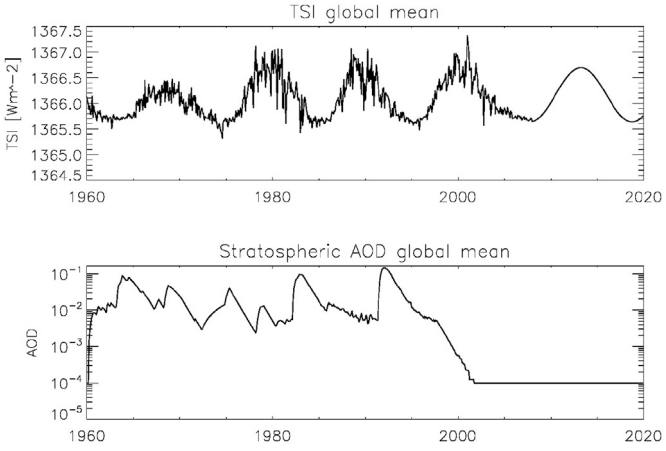


Fig. 2. Natural climate forcings prescribed to both the ALL and NAT experiments. Shown are global mean values of Total Solar Irradiance (TSI) and Stratospheric Aerosol Optical Depth (AOD) covering the period of the multi-decadal simulations (1960–2013) and their continuations up to the near-present. After 2009 TSI is an idealised solar cycle while 21st century AOD values reduce to a constant level.

2.3. Ensemble generation and experiments

The experiments we describe consist of pairs (ALL and NAT) of small ensembles of long multi-decadal simulations intended primarily for model validation, and large ensembles of short simulations used for attribution assessments. The large ensembles are continuations of the small number of multi-decadal simulations. Ensemble size is increased by producing batches of members branching from the end of a single multi-decadal simulation, hence sharing initial conditions but differing through their particular realisation of the stochastic physics. Details of the ensemble generation and structure follow below.

A set of diagnostic outputs including the C20C + requests and pre-agreed with EUCLEIA project partners were output as a mixture of monthly, daily, 6 hourly and 3 hourly mean values at a selection of different levels from surface to 17 layer 3D fields. These were subsequently converted from the native UM pp format to CF compliant NetCDF4 format and transferred to CEDA/ESGF (Earth system grid federation (ESGF) portal at the stfc centre for environmental data (CEDA)) from where they are available under license. There is a total of 36Tbytes of validation data across both experiments.

Stochastic physics ensembles are generated through the simultaneous operation of two schemes, Random Parameters (RP) and Stochastic Kinetic Energy Backscatter II (SKEB II), originally developed for and still used by the Met Office Global and Regional Ensemble Prediction System (MOGREPS) (Warren et al., 2011). Originally intended to explore forecast model uncertainty the scheme is used in our system to generate realisations of atmospheric states evolving under given forcing conditions occurring with likelihood equal to the deterministic solution.

RP varies the values of a number of physical parameters mainly related to parameterised convection within some fixed range. Values are resampled every 3 hours throughout the run such that each member samples the same set of values in order to avoid introducing a bias into the climate of a member with respect to the best estimate values. SKEB II estimates numerical dissipation of kinetic energy and reintroduces this back into the mean flow. The seeding of both stochastic schemes is controlled by the initial choice of single random number integer parameter uniquely (and repeatably) defining a stochastic physics member. Each member of a stochastic physics ensemble should hence be treated as an equally likely realisation of the same model.

A pair (ALL and NAT) of 15 member ensemble simulations, referred to as *historical* and *historicalNat*, were produced spanning the period December 1959–December 2013 (54 complete years) and principally used to validate the model but additionally required to form historical climatologies and anomalise variables in continuation experiments. All 30 runs shared identical initialisation of the atmospheric state from ERA-40 reanalysis at 0000Z on 1st December 1959, giving the experimental period one month's spin-down.

At 0000Z 1st December 2013 we branched 7 members from each simulation producing a pair of ensembles of size $(15 \times 7 =)105$ and referred to as *historicalShort* and *historicalNatShort*. At 1st January 2016 we branched a further 5 simulations from each simulation, producing a pair of ensembles of size $(105 \times 5 =)525$, referred to as *historicalExt* and *historicalNatExt* which are then continued quasi-operationally on a seasonal basis. The ensemble initialisation and continuation structure is summarised in Fig. 4 and in Table 1.

The data disseminated to the EUCLEIA and C20C + projects is presented following as closely as possible the CMIP5 file naming syntax (Taylor and Doutriaux, 2010). Labelling of ensemble members adopts the "r", "i" and "p" indexing scheme but with an important modification to accommodate the continuation structure of the ensembles, as follows. The CMIP5 convention is that "r" indicates equivalent realisations of the same model, which for coupled models would differ typically through time of initialisation from a control run. "i" indicates initialisation method and "p" would normally label the perturbed physics version and hence indicate the use of different physical models. Here we have instead made use of "r" to indicate where members share the same initial

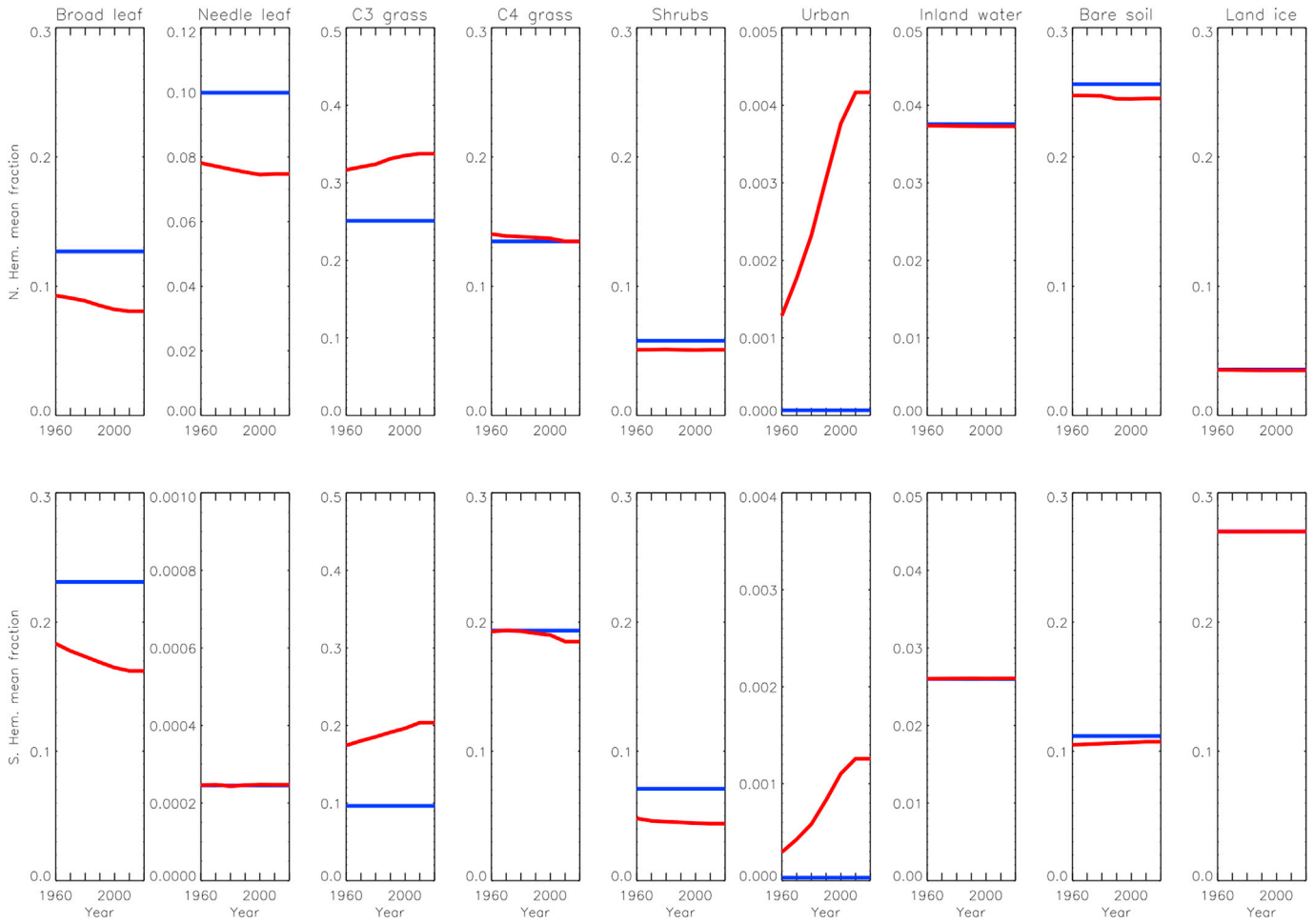


Fig. 3. Land usage as nine surface type fractions prescribed at decadal intervals (1960–2020) to the experiments (in red for historical ALL experiments, in blue for NAT experiments corresponding to 1850 values). Fractions shown are area averaged over the northern hemisphere (top row) and southern hemisphere (bottom row). Values are taken from ISAM-HYDE3.1 (Meiyappan and Jain, 2012) with values at 2020 being a repeat of those at 2010. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

conditions and the "p" distinguishes members with the same initial conditions but different stochastic physics seeds. As described above, all members of our ensemble should hence be treated as equally likely realisations of the same model, irrespective of the values of "r" and "p".

3. Framing

In its most basic form event attribution must answer a few questions: was this event "just the weather", if not was it something naturally occurring or man made? The science of event attribution makes sense of these questions, in the only possible way, through a probabilistic framework. The question being asked frames the study that should be conducted (albeit this logic is often reversed, sometimes by necessity). Amid the diversity of event attribution studies now being performed (Herring et al., 2014, 2015, 2016; Peterson et al., 2012, 2013) it is important to consider carefully the precise set of questions that a study asks (Otto et al., 2012).

Atmosphere-only models provided with boundary conditions corresponding to observed conditions are best positioned to answer questions of the form: "What is the likelihood of an event such as this conditioned on these particular boundary conditions". The HadGEM3-A based system generates ensembles sampling all possible atmospheric states consistent with these particular boundary conditions, Sea Surface Temperature (SST) and Sea Ice (SIC) patterns (referred to jointly from hereon simply as SST). Questions enforcing further conditions, such as on particular atmospheric circulation patterns, are also possible by sub-sampling the full ensemble. Conversely, conditioning on particular boundary conditions may be relaxed by combining slices of the simulations from different time periods resulting in event likelihoods comparable to those derived from uninitialised ensembles of coupled ocean-atmosphere simulations (strictly those that we trust sample historically observed oceanic variability).

In this study we will present results for three different framings of the attribution of regional seasonal mean near-surface air temperatures (see Fig. 5): 1. events conditioned on observed SST, using the most recent large ensembles (*historicalExt, historicalNatExt*) simulating a recent season to produce probability density estimates we term "2016/17 PDF", 2. events conditioned only on secular climate changes to date, hence including events occurring under all SST patterns observed over recent decades, "Secular PDF" and 3. events unconditioned on SST pattern or on secular changes (both using *historical* and *historicalNat*), termed "Recent PDF". By sub-selecting the set of all possible events each framing determines its own pair of distributions of temperatures from which we shall calculate probability ratios $R = p_1/p_0$ for the exceedance of observed seasonal mean temperatures. This will be discussed further in the final section after we have set up the validation framework in the following section.

Further aspects of the framing of event attribution questions could also be and have been considered with the HadGEM3-A based system through choices in the analysis. These concern the different climate

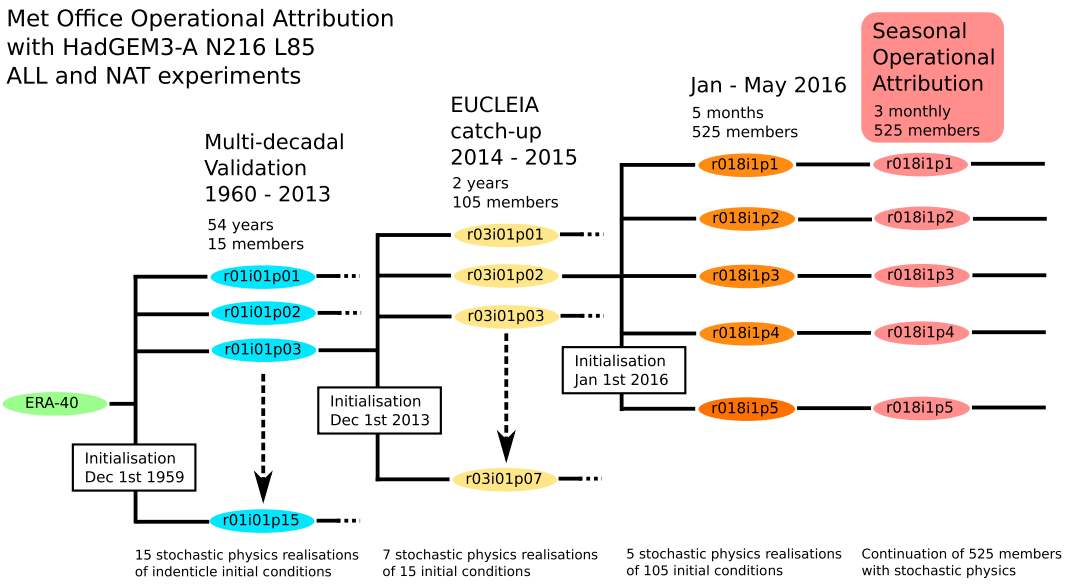


Fig. 4. Structure of the ensembles produced with the HadGEM3-A system, the resulting indexing scheme and experimental nomenclature (common to both ALL and NAT configurations). The 15 member multi-decadal simulations (1960–2013) were initialised from a single European Reanalysis ERA-40 field and are each distinguished only by a unique random number seed given to the stochastic physics scheme. All 15 members therefore share a common "r" index and differ by the value of the "p" index (the members should be treated as equivalent realisations despite the use of "p"; these are not perturbed physics ensembles). These 15 members provide 15 initial conditions for the 105 member ensembles (2014–2015) which therefore possess 15 "r" values, each with 7 "p" values. The 105 initial conditions arrived at by 2016 give rise to 525 member ensembles indexed by 105 "r" values, each with 5 "p" values.

Table 1
Naming of the three pairs of ensembles, experimental period, ensemble size and purpose. A large range of diagnostics from the *historical(Nat)* and *historical(Nat) Short* experiments are freely available in NetCDF4 format from ESGF/CEDA (Earth system grid federation (ESGF) portal at the stfc centre for environmental data (CEDA)) with the above naming by searching for the EUCLEIA project tag.

Experiment	Period	Ensemble size	Purpose
historical, *historicalNat*	1960–2013	15	Validation
historicalShort, *historicalNatShort*	2014–2015	105	EUCLEIA test cases
historicalExtxy, *historicalNatExtxy*	2016 onward	525	Operational attribution

variables used (Burke et al., 2016), the epoch (Christidis et al., 2014) and duration of an event (Christidis and Stott, 2015), its spatial extent (Angélil et al., 2017) or on particular atmospheric regimes (Christidis et al., 2016). The latter approach is important in establishing the role of influences not strongly forced by boundary conditions, and in the limit of this process the precise story lines (Trenberth et al., 2015; Knight et al., 2017) for the occurrence of an event. In this study we focus our attention on questions answered by the full ensemble but in principle there is no reason why the methodology cannot be applied to ensembles subject to further conditioning.

It should be understood that for every event z we describe we in fact refer to a set of events Z identified as those that cannot be distinguished by the event definition alone. So while we know $z \in Z$ we will also have $z' \in Z$ for some number of alternative events z'. The event set Z is a subset of all events S which are possible under the conditions of the study, and the likelihoods we state are (schematically) $P(Z) = |Z|/|S|$, the ratio of the sizes of the sets. The size of both sets of events depends on the study. For example events Z defined as seasonal mean values of a variable exceeding a threshold will describe a rather large set of possible events, only one of which, z, was actually realised by the real world. Furthermore if we condition the experiment on observed boundary conditions then S differs from an experiment where this conditioning is relaxed. Typically the likelihood $P(z) = |z|/|S|$ is vanishing. Unless we want to concern ourselves with sets of vanishing probability measure then we will always work on this understanding.

Correspondingly, for events of vanishing probability measure it is always trivially possible to attribute a necessary causal influence to a given set of causal factors (the event z would not have occurred without said causal influence, for example through sensitive dependence upon initial conditions). As soon as we consider a wider set of events Z the same causal factors may no longer have any detectable influence on the likelihood whatsoever, through low signal to noise or potentially through averaging over responses of opposite sign within sub-populations of Z. This apparent discrepancy in the attribution of causal influence highlights the importance of understanding the precise question that is asked and the associated null hypothesis (Trenberth et al., 2015). We may be at once confident that human influence is trivially all-pervasive in each of the specific events $z \in Z \subset S$ but that the influence on $P(Z)$ is nevertheless undetectable. It is also therefore a fair and informative null hypothesis that there is no human influence on the likelihood $P(Z)$ of 'an event' Z that is defined broadly enough.

So for example we know that if $z \in Z$ was some specific event that exceeded a high threshold of monthly total precipitation, and Z is defined ("framed") as the set of all possible events exceeding this threshold, we might be happy that in the absence of human influence our specific event z would not have occurred in the same way; $P(z)$ changes significantly. However it is often safe to take the null hypothesis that the probability of all such events $P(Z)$ has not changed significantly due to low signal to noise. It will often be this sort of question, regarding $P(Z)$, that is asked of an event attribution service. Users of attribution information will often simply want to know about the probability of similar events recurring. This does not misrepresent the role we already know human influence has in the climate system (Trenberth et al., 2015) but represents the requirement that the effects of this influence must be detected on each type of event.

We may highlight here the similarities and differences between the activities of seasonal forecasting and the seasonal hindcasting performed with the event attribution system. If we assume that a seasonal forecast has made an accurate forecast of SST patterns then initialised seasonal forecasts effectively sub-sample the space of states sampled by the attribution system, which produces essentially uninitialised seasonal

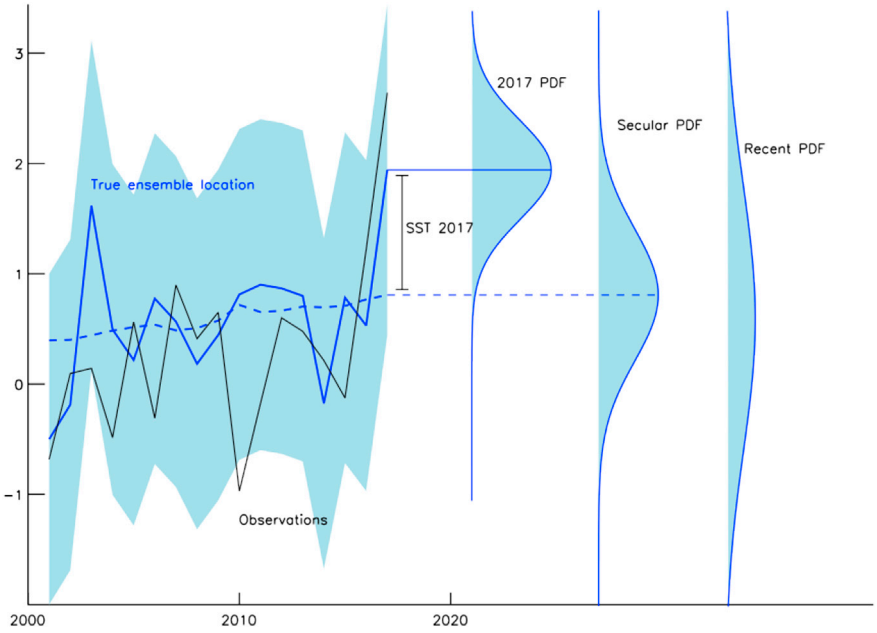


Fig. 5. Three alternative ways of framing an event leading to three alternative distributions to determine its likelihood of occurrence. SST pattern influence at a given year is defined through the deviation of the true ensemble mean about its long term mean, leading to the distribution labelled "2017 PDF". This is contrasted with the distribution "Secular PDF" formed from all residuals from the long term mean and representing the likelihood of events in 2017 that are unconditioned on specific SST patterns that occurred in this season. The third distribution "Recent PDF" is the climatology of a recent 30 year period, typically widened by a trend. In this illustration the likelihood of the most recent season being warm will have been enhanced by SST conditioning from the likelihood found from the secular change alone.

hindcasts (with a representation of perfect boundary conditions). From the perspective of the full, uninitialised ensemble any predictability arising from initialisation (or further conditioning, such as on circulation) is entirely intra-ensemble and will not be present when using the full ensemble. As a result the likelihood of some threshold being exceeded in each system will differ. Of course, this need not be the result of inconsistency or uncertainty in either approach but the result of asking different questions of the same data set. This is then just a form of subsampling as described above.

The desire of seasonal forecasting is to make a confident, accurate forecast. In lieu of regionally important SST influence this predictability must come through the atmospheric initialisation. Such intra-ensemble predictability may exist even in the absence of SST influence and takes the form of phases of important modes of atmospheric variability for example. If intra-ensemble predictability is not present we may in principle still possess a near perfect model while failing to make a confident or accurate seasonal forecast. In this circumstance event attribution is still possible as it requires only that we have a good model and make a good sample of the atmospheric states that are consistent with the observed boundary conditions. Put another way, in seasonal forecasting a lack of skill is always model error. In event attribution this clearly need not be the case and statements about the performance of models used to produce seasonal forecasts should not be taken at face value to apply to attribution hindcasts.

4. Model validation and bias correction

Most event attribution studies ultimately concern the estimation of the change with time of probability density functions over just one or a small number of climate variables averaged over some time period. A good study will also discuss the meteorological or climate mechanisms thought to have had a causal influence on the event to obtain a complete picture of the conditions that led to an extreme and an emphasis has been placed on this aspect of model evaluation in the EUCLEIA project (Vautard et al., 2018). We therefore refer the reader to the associated articles, including those related to C20C + in this special issue, for the assessment of the model's representation of dynamical and physical mechanisms.

In this study we shall focus on the evaluation of the full ensemble of historical seasonal mean near surface air temperatures. Our approach is an attempt to lay bare the source of interannual skill within the system, which enters through the boundary condition forcing. Further, we wish to clarify the relationship between model skill and performance in event attribution. By model "skill" we refer to the ability to predict the timing of variability within the climate system, which may be independent of model performance. By "performance" we refer to the degree of model error present; a model performing well is one that has no physical or dynamical differences from the real world discernible in the data to hand. Emphasis has been placed on the importance of SST patterns for regional model skill previously (Christidis and Stott, 2014). Here we stress that after removing secular changes (and neglecting rapidly varying externally forced responses, for e.g. to volcanic eruptions) then boundary condition forcing, essentially SST pattern changes, are responsible for *all* relevant model skill. Unless the ensemble is sub-sampled (through conditioning on specific circulation patterns for example) then there is simply no other reason for the distribution of climate variables at any given time to change from the climatological distribution.

Our validation procedure therefore splits the predictive elements in the model by time scale. Trends (all long term changes) in data are a misleading source of skill for event attribution purposes where we wish to know if the change in likelihood of an event was merely attributable to transient predictive factors, such as El Nino, or is the result of long term change. Secular climate changes (only approximately amounting to anthropogenic influence) are therefore removed prior to the assessment of hindcast skill. As part of the assessment we include spectra of the unmodified observations and simulations from which the consistency of long terms changes can of course be assessed.

We choose to perform the validation against observations from HadCRUT4(.5.0.0) which is a gridded global data set resulting from the blending of quality controlled and homogenised in situ land near surface air temperatures with sea surface temperatures (Morice et al., 2012). This choice was made to both increase the availability of data to coastal and oceanic regions and in the hope of enhancing comparability to the model's near surface air temperature diagnostic.

Below we set up the statistical framework used throughout the validation. Having established that the non-trivial predictive element (short time-scale predictability) in the full ensemble is driven by boundary conditions we would like to validate the system in an appropriate manner. By setting up the appropriate conception of validation we are led toward what is in essence a detection and attribution of SST influence. As well as allowing us to understand what the model is getting right we are naturally led to a procedure for bias correction of attribution experiments. Ultimately we must ask if the observed series can have reasonably

been generated by the model that generated the ensemble, which is the only circumstance in which event attribution statements can seriously be made.

Approaches to evaluation not dissimilar to ours have been discussed elsewhere recently (Sansom et al., 2016; Siegert et al., 2016) in the context of forecast evaluation but also using predictable components and the same principle of exchangeability between ensemble members and observations. Here we will follow a procedure appropriate for the assessment and calibration of attribution ensembles.

4.1. Validation framework

An event attribution study normally begins by selecting an averaging period (for e.g. seasonally) with instances labelled by t and a climate variable X_t averaged over some spatial region (that could even correspond to a model grid cell) prepared as closely as possible to an observational series Y. We shall immediately drop the index t to clear up the notation. Members of the ensemble, labelled by $a \in \{1, 2, \ldots, n_{\text{ens}}\}$, may be decomposed as

$$X_a = F + \varepsilon_a \tag{1}$$

where F is a response common to members and ε_a a member specific variation. F could in principle be calculated from an ensemble of infinite size simply as the ensemble mean, $F = \langle X_a \rangle,\ n_{\text{ens}} \to \infty$.

This common response contains both secular (essentially long term) changes f and quasi-periodic or stochastic (essentially short-term) changes μ and so we can further decompose members as

$$X_a = f + x_a, \quad x_a \equiv \mu + \varepsilon_a. \tag{2}$$

The definition of f (and so μ) will inevitably contain some arbitrariness but is such that we have centred variables x_a in which μ clearly represents the common response to SST. Just as $F = f + \mu$ may be thought of as a genuine feature of the model (the limit F exists whether we take it or not) so the series ε_a may be considered to be sampling some t-dependent parent population $\varepsilon_a \sim g$ which in principle can be constructed for $n_{\text{ens}} \to \infty$. The ε_a essentially represent variability generated internal to the atmosphere but of course responses to secular and quasi-periodic changes in SST patterns could remain through changes in the shape of each g. Together the locations μ and the details of g constitute the complete predictive component of the full ensemble $\{x_a\}$ and determine the likelihood of exceeding some threshold at t.

By contrast our observations Y are a single series. Nevertheless it is a physically well-posed question to ask what Y would have been if at some distant previous time a small perturbation to the initial conditions of the atmospheric state were made (the butterfly flaps). Conditional on sequences of events that leave the boundary conditions unchanged we therefore conceive an analogous non-stationary distribution from which the Y have been drawn such that

$$Y = f + y; \quad y = m + e, \tag{3}$$

m being the response to SST and $e \sim h$ the component generated internal to the atmosphere. We could have introduced extra parameters to describe the true secular component but here we opt to assume that this is equal to f because we shall be dealing with a variable that is dominated by boundary conditions. The assumption should itself be tested and it will enter into our analysis below.

The validation task becomes one of assessing the statements

$$\mu = m, \tag{4}$$

$$g = h\ (\forall\ t). \tag{5}$$

If (4) and (5) are true then event likelihoods calculated for the historical experiments (ALL) are dependable (for any threshold) as the model could be said to have generated the observations. Simply put (m, h) corresponds to a perfect model and (μ, g) to our imperfect model.

In practice we have an ensemble of finite size and some degree of model error. In the presence of model error in the response to SST we may have an error in the magnitude and timing of the common response μ as well as in the shape of g. As the sample ensemble mean $\langle x_a \rangle$ estimates the true mean μ we can write

$$\langle x_a \rangle = m + d + \nu \tag{6}$$

where ν is the sampling error and d is a total common error in response. We may split the total common error as $d \equiv d_m + \delta, d_m \equiv (\alpha - 1)m$ in which d_m is the component projected parallel to (i.e. scales with) m and δ is the orthogonal complement. Then we have

$$\langle x_a \rangle = \alpha m + \delta + \nu \tag{7}$$

so that α is the factor accounting for error in scale of response to SST in the true model mean.

We want to assess the statements (4) and (5) but with only a single observed series y we lack a means of directly estimating m and h (separating the signal from the noise). Instead we must use techniques that combine data from across t. This is in essence what has been attempted previously through the reliability diagram (Christidis et al., 2013a; Bellprat and Doblas-Reyes, 2016), which in this framework we can see plots estimates of the cumulative distribution functions, $H(Y \geq y_c|t)$ against $G(X \geq y_c|t)$, for some critical threshold y_c. The very same predictable components in our system that are assessed through the reliability diagram however are clearly related to a regression of the observations to the ensemble mean: from (7) we can write

$$\begin{aligned} &\langle x_a \rangle = \alpha y + \delta + \nu - \alpha e \\ &\Rightarrow y = \beta\{\langle x_a \rangle - \delta - \nu\} + e \end{aligned} \tag{8}$$

where $\beta = \alpha^{-1}$ and so we arrive at the form of a total least squares (TLS, Allen and Stott, 2003) + "errors in variables" (Huntingford et al., 2006) (EIV) regression of y onto $\langle x_a \rangle$. After the decomposition by time scales and removal of the secular changes we have made no approximations or undue assumptions, and in particular we did not arrive at (8) by assumptions akin to the linear composition of different physical responses in a fingerprinting study. This is an almost completely general treatment of the predictable component due to rapidly changing influences in an ensemble of uni-variate time series.

Performing the suggested regression involves finding the best fit parameters $\widehat{\beta}$, $\widehat{\delta}$ and ensuring that residuals $\widehat{e}$ pass a test regarding our assumptions as to the nature of h. Having removed low frequency variability in X_a, Y through f the variability remaining in x_a, y that is relevant to the event attribution question is found in the high frequency variations that can be examined through the high frequency region of the power spectra. The high frequency region is also the best sampled and the ensemble mean power spectrum shall be a fair estimate of the associated power spectral density there. If the power spectral density of residuals to the mean can be argued to be approximately constant at high frequency then the correlations between y and $\langle x_a \rangle$ measured by $\widehat{\beta}$ arise from data equivalent to pre-whitened data and $\widehat{\beta}$ can be considered to be close to the best linear unbiased estimator (BLUE) so the regression is also near optimal. To be properly optimal an analysis would require a form of control experiment that we do not possess but we shall return to this issue in section 7.

The values of $\widehat{\beta}$ (and if obtained, $\widehat{\delta}$) determine a description of model error in the predictable component of the ensemble system and could therefore be used in a direct bias correction of event attribution results, telling us how to map between realisations of the two statistical models: (m, h) and (μ, g). Unlike bias correction via reliability diagrams this method does not rely on the arbitrary choice of a critical threshold, which in any case is often unrepresentative of the more extreme thresholds we are interested in on account of finite sample sizes. Unlike with reliability

diagrams this method does not require us to bin data together to form the "observed" event probabilities, but the trade-off is that we make assumptions regarding g and h.

The problem of validating estimates of event probabilities in the ALL experiment therefore naturally involves a detection and attribution exercise on the response to boundary condition forcing.

If we restrict our attention to series in which the averaging period τ is long enough, monthly or seasonal time series for example, the atmospheric response to SST is instantaneous. As the model is supplied with observed SST there is also no relevant error in the timing of forcing. We would therefore expect vanishing contribution to δ from this sort of "timing error" when τ is large compared to the characteristic timescale of synoptic evolution. If we were using a coupled model, in which there is error in SST itself, or were producing forecasts, in which information from the initialisation is retained and the precise evolution of synoptic patterns mattered, then the same would not be true. This is an important (simplifying) difference in the validation of the event attribution system from that of a seasonal forecasting system.

This argument does not preclude all time dependent errors in response. We could consider error occurring at specific times, such as an incorrect response to some important mode of variability such as El Nino whose influence arrives through the boundary conditions. Moreover, any error in the secular response f of the ensemble to external forcing will appear here. It therefore makes sense to split δ up into timing (τ) and instantaneous (0) components, $\delta = \delta_\tau + \delta_0$. We would require additional evidence that $\delta_0 \approx 0$, but we can consider the following method: assume that $\delta_0 = 0$ ($\Rightarrow \delta = 0$ if $\delta_\tau = 0$) and conduct a regression to determine $\widehat{\beta}$. If we find that $\widehat{\beta}$ is consistent with 1 and that the residuals pass a test regarding acceptable assumptions on h then we have evidence that there is also no significant time dependent error in response to boundary condition forcing. If we find $\widehat{\beta}$ is inconsistent with 1 then we may have $\delta \neq 0$, or we may just have a significant scale response error; an EIV analysis or some other argument would have to be produced to determine the contributions, which we do not attempt here.

Here we conduct an ordinary least squares (OLS) regression (Allen and Tett, 1999) without an explicit optimisation step. The sampling error ν in the ensemble mean estimate of the SST signal arises solely due to atmospheric variability within a finite ensemble. ν therefore shares the same noise-covariance structure as ε, which as we are dealing with time series involves autocorrelation only. We cannot with this method address the associated well known bias in $\widehat{\beta}$ toward zero but be mindful that it may therefore overestimate the need to scale up the model estimate of the SST signal. The secular component f_t will be defined as the moving 15 year (box-car) mean of the ensemble mean $\langle X_{a,t}\rangle$ centred at t. We explore the sensitivity of the regression to this definition in Appendix B where we find that it is reliably second order compared to the uncertainty in the regression itself.

The error $\Delta\beta$ on $\widehat{\beta}$ is produced as follows. The regression coefficient can be seen as Student t-distributed about its best estimate when the regressand is seen as a random variable with residuals from the best estimate signal that are normal (Von Storch and Zwiers, 2001). This is the same logic implicit in the form of uncertainties used for regression coefficients (scaling factors) in optimal trend detection (Allen and Tett, 1999) where multiple regression results in coefficients whose sum of squares is F-distributed about their best estimates in signal-space. For a single signal pattern this relationship reduces to the t-distribution we use here. Our regressor $\langle x\rangle$ is an estimate of the SST pattern response which is centred (over 15 year periods) about the secular changes. It is customary to separately check the residuals $\widehat{e} = y - \widehat{y}$ of the observations from the fit for normality and below we describe a whiteness test on the residuals that we conduct along side our results. We therefore use

$$\Delta\beta = \left(1 + \frac{1}{n_{\text{ens}}}\right)^{\frac{1}{2}} \frac{s_E}{\sqrt{\text{SS}_{\langle x\rangle}}}\, t_{\frac{1+\tilde{p}}{2}} \tag{9}$$

where

$$s_E^2 = \frac{(y-\widehat{y})\cdot(y-\widehat{y})}{n_t - 2} \quad \text{and} \quad \text{SS}_{\langle x\rangle} = \left(\left\langle x\right\rangle - \overline{\langle x\rangle}\right)\cdot\left(\left\langle x\right\rangle - \overline{\langle x\rangle}\right). \tag{10}$$

$\widehat{y}$ is the best fit, an over-bar indicates the series time mean has been taken and a central dot represents the scalar product between time series, $a\cdot b = \sum_t a_t b_t$. As our regressor is a finite size ensemble mean we have also inflated the width by the appropriate n_{ens}-dependent factor (Allen and Tett, 1999) (an inflation of only 3% for $n_{\text{ens}} = 15$). We pick t-values $t_{(1+\tilde{p})/2}$ ($n_t - 2 = 52$ degrees of freedom) at a significance of $\tilde{p} = 0.9$ and so quote values of $\beta(p)$ at $p = 5\%$ ($\widehat{\beta} - \Delta\beta$) and $p = 95\%$ ($\widehat{\beta} + \Delta\beta$). When $\widehat{\beta} - \Delta\beta > 0$ we say that the SST influence is detected (i.e. β is inconsistent with zero, $\beta \not\sim 0$) while if $\widehat{\beta} - \Delta\beta \leq 0$ we say it is not detected (i.e. β is consistent with zero, $\beta \sim 0$).

For the residuals whiteness test we assign a p-value to the spectra of the residuals $\widehat{e}$ as follows. We obtain the fraction of frequencies ω where the power contributed is greater than a standard deviation from the mean power and compare this to a set of fractions and corresponding p-values expected from a white process of the same length. The null hypothesis H_0 is that the spectra could have been generated by this white process and a small p-value indicates we should reject this hypothesis, which is also therefore a failure of this method to portray the observed series as equivalent to a member of the ALL experiment. If an examination of the spectra of the residuals indicates that low frequency residuals were the cause then we may suspect that the test failed because of the assumption that Y and X_a share the same secular changes, f.

The calculation of the above involves obtaining power $\left|w(\omega)_{\widehat{e}}\right|^2$ at frequency ω from the Fourier decomposition $\widehat{e} = \sum_\omega c(\omega)_{\widehat{e}} e^{-i\omega t}$ for comparison with equivalent contributions $\left|w(\omega)_{r,a}\right|^2$ from ensemble residuals to the mean, $r_a = \sum_\omega c(\omega)_{r,a} e^{-i\omega t}$. For a white process the power is constant across ω and we estimate the corresponding constant from the mean power $\lambda = n_{\text{ens}}^{-1} n_f^{-1} \sum_a^{n_{\text{ens}}} \sum_\omega^{n_f} \left|w(\omega)_{r,a}\right|^2$. We then count the number of exceedances of a single standard deviation from this mean, $\left|\left|w(\omega)_{\widehat{e}}\right|^2 - \lambda\right| > s(\omega)$ where $s(\omega)$ is the standard deviation of $\left|w(\omega)_{r,a}\right|^2$. We allow for an overall shift of the spectrum of $\widehat{e}$ to have the same mean power as $\{r_a\}$ as in practice the widths may differ and in section 4.2 we discuss the corresponding bias adjustment. Individual member realisations of the power at ω will be approximately normally distributed so we consider power deviations from a candidate white process to exceed $\pm s(\omega)$ with a fixed probability of $p_\pm \approx 0.317$ and the number of exceedances n_{ex} will then be binomially distributed with parameter $p_\pm$, $n_{\text{ex}} \sim B(n_{\text{ex}}|n_f, p_\pm)$. Finally, the p-value we associate with residuals is just the probability of having greater than or equal to n_{ex} exceedances, $p = 1 - \sum_{n'=0}^{n_{\text{ex}}-1} B(n'|n_f, p_\pm)$.

We require a means of assessing whether the model mean $\langle x\rangle$ likely contains a signal to be detected in the first place. Intuitively, as long as the common true model response μ has a variance σ_μ^2 which is a significant fraction of the ensemble true variance σ_x^2 then we know that the SST signal has a large influence on probabilistic predictions made by the ensemble. We estimate this fraction by the ratio $\Re^2$ of the two sample variances,

$$\Re^2 = \frac{s^2_{\langle x\rangle}}{s_x^2}, \qquad s_x^2 = \left\langle s^2_{x,a}\right\rangle \tag{11}$$

i.e. the ratio of the variance in time of the ensemble mean to the ensemble mean variance in time of each member, which is estimating the true ratio

$$\rho^2 = \frac{\sigma_\mu^2}{\sigma_x^2}. \tag{12}$$

This quantity is defined similarly to the "predictable component" discussed in a seasonal forecasting context (Eade et al., 2014) so we retain the terminology. We want this value to be close to the unknown real world (RW) predictable component

$$\rho_{\mathrm{RW}}^2 = \frac{\sigma_m^2}{\sigma_y^2}. \tag{13}$$

The value of $\Re$ can be used as a summary statistic for the importance of SST forcing in the model in a region. Under the null hypothesis that $\mu = 0$ we know that $\langle x \rangle$ will still be subject to finite ensemble noise with $s^2_{\langle x \rangle} = \mathscr{E}(n_{\mathrm{ens}}^{-1} s_x^2)$ and so, approximately, $\Re^2 \in [n_{\mathrm{ens}}^{-1}, 1]$. Measured in the associated units of signal to noise we can therefore use the variable

$$\tilde{\Re^2} \equiv n_{\mathrm{ens}} \Re^2 \quad \Rightarrow \quad \tilde{\Re} \in [1, \sqrt{n_{\mathrm{ens}}}\,] \text{ (approximately)}. \tag{14}$$

In practice we must test the hypothesis that there is no model signal, $H_0 : \mu = 0$. We can find an appropriate critical value of the $\tilde{\Re}$ variable by assuming that $x_{a,t} \sim \mathscr{N}(0, \sigma_x)$ which implies that under H_0 we expect $\tilde{\Re^2} \sim F(n_t - 1, n_t(n_{\mathrm{ens}} - 1) - 1)$ with reasonable accuracy as the denominator has very many degrees of freedom independent of the numerator. Here the numerator possesses n_t degrees of freedom with one removed through the mean, while the denominator possesses $n_t n_{\mathrm{ens}}$ degrees of freedom from which we remove $n_t + 1$ through the n_t ensemble means and one overall mean. Approximating with $n_t(n_{\mathrm{ens}} - 1) - 1 = 755 \approx \infty$ the critical value $\tilde{\Re}_c^2$ lying at 95% of the distribution $F(53, 755) \approx F(53, \infty) = \chi^2(53)/53$ gives $\tilde{\Re}_c = 1.16$. If $\tilde{\Re} > \tilde{\Re}_c$ then $p < 5\%$ and we reject H_0 to say there is an SST signal in the model.

We are now in a position to discuss the various possibilities for combinations of model skill and model error. Fig. 6 categorises the basic possibilities, starting on the left hand side with three cases that can be distinguished based on the information available to us: a single observed series and a set of ensemble hindcasts. In case (i) the model has a signal ($\tilde{\Re} > \tilde{\Re}_c$) and this signal is detected in the observations (β inconsistent with 0). Here we would say the model has hindcast skill and is performing well: we can be happy that real world SST influence is to some degree correctly captured by the model. If the real world contained either no SST signal or a signal that was significantly different to the model we would not likely have found β inconsistent with 0.

In cases (ii) and (iii) SST influence is not detected (β consistent with 0). We have two possibilities: either (ii) the model has a signal ($\tilde{\Re} > \tilde{\Re}_c$) which was not found, or (iii) the model has no signal ($\tilde{\Re} < \tilde{\Re}_c$). Case (ii) may indicate model error in the form of SST influence in the model that is not present in the real world, sub-cases (ii)(a) and (b). However we may simply have been unlucky and a relatively weak signal has not been detected due to low signal to noise. Case (iii) where the model has no significant SST signal involves model error if the real world has an SST signal but may also be consistent with a good model if there is similarly no real world SST influence, as may be expected in continental interiors. In this case we would not expect the model to possess any hindcast skill but it may still be performing well.

It has previously been suggested that where a model possesses no skill (in the *timing* of variability) the model may still be useful for attribution as long as the spectral properties of variability are adequately represented (Christidis et al., 2013a). Here we have made this statement more precise. Lacking skill here means falling into cases (ii) and (iii) which are consistent with zero influence from boundary condition forcing (β consistent with 0) in which case we further ask if boundary condition forcing is strong in the model (case (ii), $\tilde{\Re} \gg \tilde{\Re}_c$) and if it is then we should consider there to be model error. If model boundary condition forcing is weak (case (iii), $\tilde{\Re} \approx \tilde{\Re}_c$) then as long we have no good reason to suspect an important missing response (for example a missing teleconnection to El Nino) then the original advice prevails and we should examine spectra. A step by step summary of the validation process is given in Fig. 7.

4.2. Bias correction

The above method of diagnosing one degree of freedom (per region) of the model error immediately suggests a bias correction that takes the ensemble mean and scales this about the secular changes (i.e. scales the SST pattern response by shifting the distribution) in order to meet the best estimate of this signal found in the observations. The validation method does not estimate details of the distributions h other than their

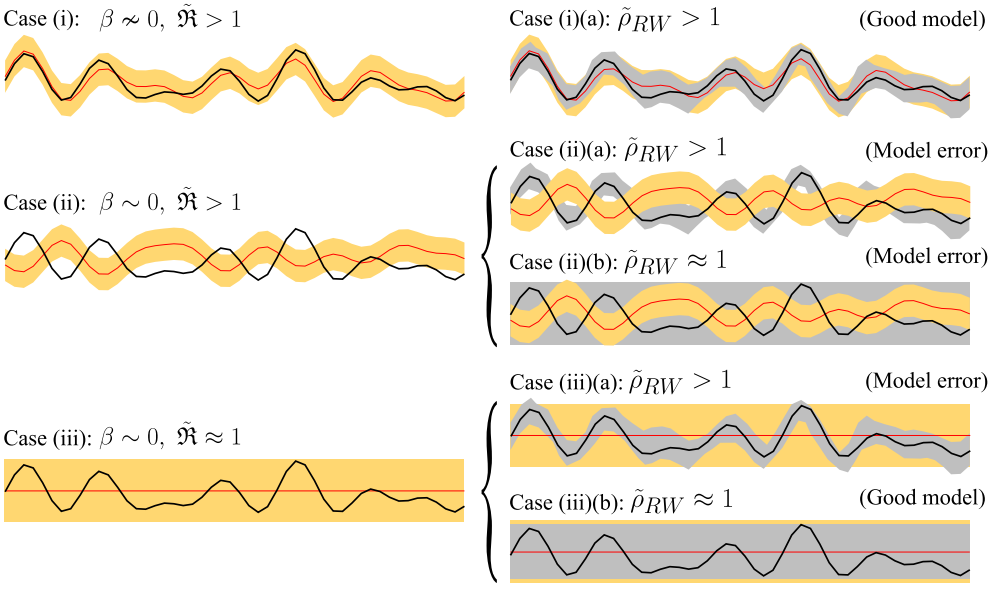
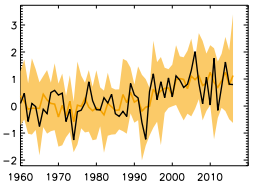


Fig. 6. The relationship between model and real world (RW) predictability due to SSTs as assessed through the value of the regression coefficient, β and predictable component $\tilde{\Re}$. In practice we may distinguish the three cases on the left hand side from the information available to us (a single observed series in black and ensemble in orange with mean in red). These correspond to five possibilities (extreme cases illustrated) on the right hand side distinguished by the value of the unknown real world predictability, $\tilde{\rho}_{\mathrm{RW}}$. Case (i): Detection of SST influence means β is inconsistent with zero ($\beta \nsim 0$) and can only happen in the presence of model predictability ($\tilde{\Re} > 1$) and likely only happens in the presence of real world predictability ($\rho_{\mathrm{RW}} > 1$). Cases (ii) and (iii): Undetected SST influence does not necessarily mean model error: weak signals ($\tilde{\Re} \approx 1$ or $\rho_{\mathrm{RW}} \approx 1$) or a complete lack of real world predictability may still be consistent with a good model. In such cases we should examine the spectra to assess variability not associated with predictable components. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

1) Raw data: observed, model member and ensemble mean series, $(Y,\ X_a,\ \langle X\rangle)$.

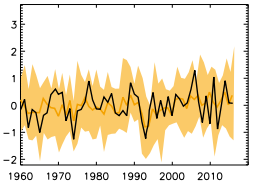
e.g. region mean time series of '61-'90 tas anomalies



2) Define and remove secular changes component.

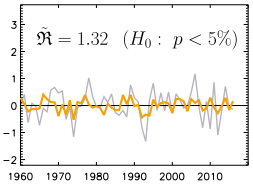
$(y,\ x_a,\ \langle x\rangle) = (Y,\ X_a,\ \langle X\rangle) - f$

e.g. rolling 15 year mean of ensemble mean $f \equiv \langle X\rangle^{15\text{year}}$



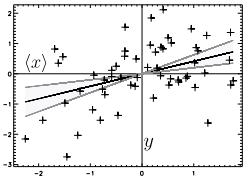
3) Obtain signal strength $\tilde{\Re}$ and assess significance against null hypothesis of no SST signal, $H_0 : \mu = 0$.

$$\tilde{\Re} = \sqrt{n_{\text{ens}}}\frac{s_{\langle x\rangle}}{s_x}$$



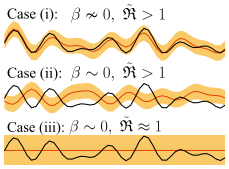
4) Obtain regression coefficient $\hat{\beta}$ and assess detection of SST signal in observations: $y = \hat{\beta}\langle x\rangle + \hat{e}$

Detection condition: $\beta > 0$ $(p > 5\%)$



5) Categorise into cases (i) - (iii) for presence of model SST signal and detection in observations.

Assess model performance, e.g. is skill found when it is expected?



6) Test observed residuals $\hat{e}$ for consistency with white process. Examine raw and residuals spectra for obvious inconsistencies of model with observed series.

Is adequate definition of the secular component f apparent at low frequency?

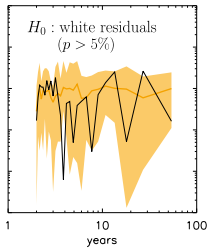


Fig. 7. A step by step schematic for the complete validation process. These steps allow us to consider model performance according to an assessment of the presence of predictable influences acting at short time scales whose magnitude can be as important or more important than secular changes. In this study predictability is due to boundary condition forcing, or SST and SIC pattern influence, but the method is quite generally applicable whenever the framing of the attribution question admits influences with predictive power over model distributions.

locations but rather just looks for consistency with the simplifying assumption that $g, h \sim \mathcal{N}\ \forall\ t$. Any difference in the widths of g versus h must therefore be addressed separately, similarly to the bias adjustment of seasonal forecasts (Eade et al., 2014; Weisheimer and Palmer, 2014), which we shall also consider here.

The bias correction of a single ensemble experiment is straight forward. By requiring that a given threshold y in the distribution h lie at the same percentile as a model threshold x lying within the distribution g implies the following condition on the residuals if we take them to be normally distributed but possibly of differing widths,

$$\frac{y - \widehat{m}}{s_{\widehat{e}}} = \frac{x - \langle x\rangle}{s_r}. \tag{15}$$

Here $s_{\widehat{e}}$ is the sample standard deviation of observed residuals $\widehat{e} = y - \widehat{m}$ and s_r is our estimate of the standard deviation of the model residuals, for which we use the ensemble mean $s_r = \langle s_{r,a}\rangle$ of the sample standard deviation of residuals of each member $r_a = x_a - \langle x\rangle$. Solving (15) for y leads to the bias correction mapping ensemble members $x_a \sim g$ into our estimate of the true model, $x'_a \sim h$,

$$x'_a = \widehat{m} + \frac{s_{\widehat{e}}}{s_r}(x_a - \langle x\rangle). \tag{16}$$

The x'_a are series that could have been produced by the process that produced y. This is similar to the correction suggested recently (Eade et al., 2014) for seasonal forecasting except that we rescale using the width of the ensemble s_r rather than the width on an individual ensemble member $s_{r,a}$, which is a more appropriate adjustment to map between distributions g and h; we should not expect the standard deviations of two exchangeable finite realisations of a random process to possess precisely equal variance.

However, attribution involves two ensemble experiments, ALL and

NAT, resulting in an ambiguity upon adjustment. For example, if the model has a positive variance bias in its residuals about the signal and we scale the width of both ensembles down about fixed locations as per (16) then we have produced two ensembles of the supposedly correct width, in the process changing the resulting likelihood ratios (a similar issue to that described recently (Bellprat and Doblas-Reyes, 2016)). In the tails of these distributions above a given threshold, for example, lie a different set of events in ALL than in NAT. See Fig. 8. Such an adjustment therefore inconsistently adjusts the likelihood of sets of similar sequences of events simulated under the ALL and NAT scenarios. Yet for event attribution purposes it is essential that we retain the same sampling of the intersection of events which can occur under an ALL forcings scenario but are also possible under the NAT scenario.

We therefore propose the following "inverse bias correction" which instead of mapping the ALL model ensemble into our best estimate of the real world now maps the observations into the ALL model world. From (15) we solve instead for x, hence mapping real world thresholds y into the equivalent model world thresholds, y'.

$$y' = \left\langle x \right\rangle + \frac{s_r}{s_{\widehat{e}}}(y - \widehat{m}). \tag{17}$$

We now have exchangeability between y' and x_a. Likelihoods calculated using the inverse corrected threshold to query the unadjusted ALL ensemble will be identical to those generated using the unadjusted observations to query the "forward corrected" ALL ensemble using (16). This way we preserve the internal physical consistency between sets of atmospheric states generated by the pair of experiments. To calculate likelihoods p_1 and p_0 we therefore query the density function estimates from the unadjusted ALL and NAT ensemble data with the "inverse corrected" threshold, y'.

This approach may also help address potential issues when directly comparing observational datasets and model data. For example the station density in HadCRUT4 can be low (in many locations a single station contributes to the cell values) while the interpretation of model data remains consistent between locations. By adopting this inverse bias correction the observational series at different locations all become exchangeable with the corresponding model data, which is always comparable between locations.

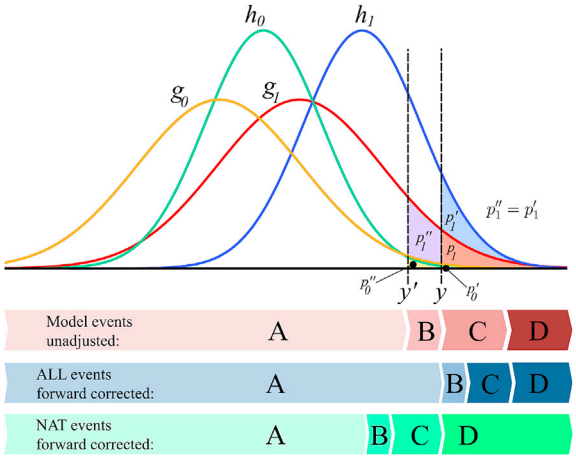


Fig. 8. The usual "forward" bias correction of model distributions inconsistently treats events in the ALL and NAT scenarios. Here NAT (0) and ALL (1) model distributions g_0, g_1 (cumulative distributions G_0, G_1) are transformed to h_0, h_1 (H_0, H_1) with a shift and reduction in width. A larger set of events ($B \cup C \cup D$) taking place in the ALL ensemble contribute to the resulting "forward" bias corrected probability $p'_1 = H_1(Y \geq y)$ than contribute to the "forward" corrected probability $p'_0 = H_0(Y \geq y)$ (just events D). This is a physical inconsistency that does not respect the attribution question, which concerns the change in probability of the same events. The inverse correction we suggest here amounts to simply adjusting the threshold $y \rightarrow y'$ such that the inverse corrected ALL probability $p_1'' = G_1(Y \geq y')$ is unchanged from the forward corrected probability p_1' while the inverse corrected NAT value $p''_0 = G_0(Y \geq y') \neq p'_0$ genuinely concerns the same set of events as the ALL value.

5. Validation of multi-decadal simulations

In this section we will apply the validation procedure of section 4.1 to seasonal mean near-surface air temperatures produced by the multi-decadal ALL experiment, *historical* (1960–2013). This is a variable in which we would expect to find significant regional interannual variations due to SST patterns that occur on top of the secular changes occurring due to external forcing and decadal scale variability and so it is a variable in which we have to be confident of changes in likelihood hindcast by the model. The procedure will allow us to be confident of where real world interannual SST responses are captured by the model and to begin to disentangle where lack of model skill could be model error versus where it could be a faithful presentation of real world lack of regional predictability in seasonal temperatures due to SST.

The analysis is conducted on model time series processed to be comparable to seasonal means of the HadCRUT4 observational data set. The processing of model data involves anomalising monthly mean absolute values to anomalies with respect to 1961–1990 (using each member's own climatology) at native model resolution and then regridding these to 5° × 5° resolution before taking seasonal means. Four seasons are defined as the three month means of December–February (DJF, labelled with the year in which January falls), March–May (MAM), June–August (JJA) and September–November (SON). Each cell then provides an observed series and model ensemble to which we apply the procedure. Observed series for a season are discarded where any of the monthly data over 1960–2013 are missing; we are therefore unable to apply the full validation in these cells, although we may still examine the strength of the model SST signal.

As a further step in the validation procedure we check that the values of $\widehat{\beta}$ we obtain are not sensitive to the definition of f, the secular change in X, by repeating the regression with different values of the smoothing period. This is discussed in Appendix B where the sensitivity to f is found to be sub-leading compared to the fit uncertainty described above. This confirms both that the skill we identify originates in interannual predictability from SST patterns and that the secular changes f found in the model mean are normally very close to those inferred to be present in the observed series.

In Figs. 9–11 we categorise each cell into the three cases of section 4.1 and Fig. 6 season by season and we shall discuss the results below.

First, the left hand column of Fig. 9 displays best estimate values of $\widehat{\beta}$ wherever SST is detected (β inconsistent with 0) and wherever we have complete observational coverage (this cuts out much of the African interior and southern ocean). As may be expected from near surface air temperatures in a model driven by observed sea surface temperatures and sea ice concentrations the fingerprint of interannual SST is detected and associated with values of $\widehat{\beta}$ close to one over almost all of the ocean surface (where observations are available). As could be anticipated SST fails to be detected over continental interiors where its influence on seasonal temperature variations is not dominant and much of these regions will fall into categories (ii) and (iii), discussed below. These regions are surrounded by regions of decreasing values of $\widehat{\beta}$ that may be associated with the known bias of $\widehat{\beta}$ toward 0 in the OLS method that will be present when signal to noise is low. We will discuss what to say about model performance in these regions below.

Significantly, interannual SST influence continues to be detected over most of the European land region in all seasons. Furthermore the best estimate values are often close to one indicating a near perfect representation of interannual shifts in the location of seasonal temperature distributions due to changing SST patterns.

The middle column shows SST signal strength. Depicted are the values of $\tilde{\Re}$ (defined in equation (14)) where $\tilde{\Re} > \tilde{\Re}_c$, i.e. where we can be

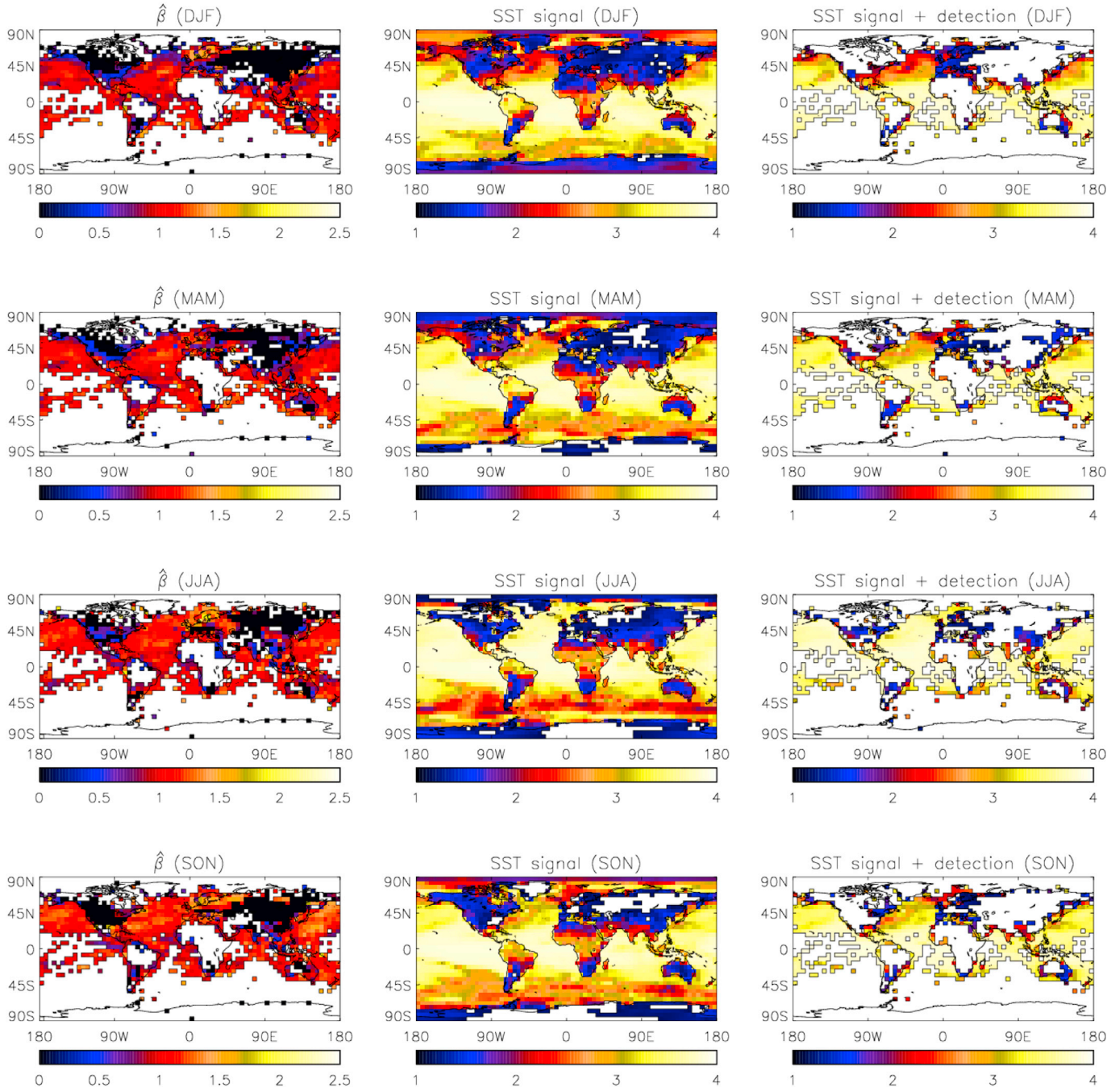


Fig. 9. Strength of SST signal in ensemble location and detection in observations. Each row is a season. Left hand panels: coloured for value of regression coefficient $\hat{\beta}$ where SST signal is detected in the observations (coloured black where not detected and left white for missing observational data). Middle: strength $\tilde{\Re}$ of SST location signal as represented by the ensemble mean. Shown with colour are all values of $\tilde{\Re} > \tilde{\Re}_c$ ($H_0 : \mu = 0, p < 5\%$) which represent regions where conditioning on SST will provide new information (almost everywhere). Right hand panels: coloured with SST signal strength where there was observational coverage, where $\tilde{\Re} > \tilde{\Re}_c$ and signal is also detected at 95% confidence, corresponding to case (i) in Fig. 6. This is where the model produces SST conditioned predictions which are also historically validated by the available observations. SST signal is largely present and detected over the oceans. Over continental interiors SST signal is weak and hence largely undetected. Much of Europe retains a detectable and historically validated influence from SSTs in all seasons. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

confident that the ensemble mean contains interannual variations on top of secular changes that are not likely to be due to finite ensemble random fluctuations. These are regions where conditioning an ensemble on historically observed SST provides new information compared to an ensemble unconditioned on SST, for example as would be generated by a large ensemble from a coupled model. We see that SST signal is significant almost everywhere at all times. The ensemble over the oceans contains a very strong SST signal, which nevertheless decreases in magnitude over the north Atlantic and Southern Ocean storm tracks. Tropical land regions retain a strong signal while temperate and polar land regions retain only a weak signal, albeit still significant.

The right hand side panels of Fig. 9 indicate where we have both a significant SST signal and detection, and hence cells that fall into our case (i). These are the regions where we can be confident that distributions of

seasonal mean near-surface air temperatures in the HadGEM3-A system are subject to interannual variations in location on top of secular changes that are to some degree capturing real world responses. These are cells where we should have skill and do have skill. European land regions are notably well modelled, particularly in DJF and MAM, along with many parts of the United States, Australian coastal regions and eastern Asia in JJA.

An example of a case (i) region is given in the top row of Fig. 12, representing England and Wales in SON. The SST signal (ensemble mean minus the secular component) is significant (central panel), though actually has reasonably low strength, and is both detected and consistent with $\beta = 1$ (scatter plot, right hand panel), suggesting a perfect timing and magnitude of response in the location of the distribution to interannual SST patterns.

Large areas of the continental interiors where SST is not detected fall into case (ii) where we have a significant SST signal which failed to be detected (note that much of the African interior has missing observational data where the analysis could not be conducted). The SST signal strength for these regions is displayed in Fig. 10. It is immediately apparent that most of these regions are associated with weak signals so that the failure to detect SST is more likely associated with low signal to noise as opposed to erroneous model responses to SST. Exceptions are regions of Alaska in DJF and MAM and areas surrounding the subcontinent in all seasons where signal strength is high and yet it is not detected. It seems that these could be regions where we can point to a lack of skill that is genuine error in the model response.

An example of a case (ii) region is given in the middle row of Fig. 12, approximately representing an area of the United States containing Washington DC and part of Maryland and Virgina in MAM. Casual inspection of the timeseries (left hand panel) would not in itself indicate any problems but the region contains a relatively strong SST signal which is nonetheless not detected.

Finally, regions that fall into case (iii) are displayed in Fig. 11. This is where the model presents no significant SST signal and it is not detected. There are a small number of such cells globally speaking, but they form large contiguous regions in northern Asia in MAM and SON. These are regions where the model possesses no skill and is not expected to. As long as there is not a real world response that the model has completely missed then these regions may still be consistent with a good model. Most of the area that fell strictly into case (ii) should be treated as case (iii). An example of a case (iii) falling within Europe is a cell representing a region of central France in JJA, bottom row of Fig. 12.

To summarise, the oceans and large areas of the land surface, through the seasons, immediately reveal real world skill in the interannual timing and response of near-surface air temperatures to boundary condition forcing that is captured by the model (case (i) regions). Much of the rest of the land surface possesses very low signal to noise in changes in the location of the ensemble on an interannual basis (cases (ii) and (iii)) and so all that would be required for attribution here, statistically speaking, is that the region exhibits fair variability in spectral terms. In the following section we shall examine the power spectra as a matter of course. We therefore have plenty that would give us confidence in the changes in likelihood produced by the ensemble on interannual time scales.

It is recommended elsewhere that a large emphasis be placed on the generically low skill of seasonal forecast models (Bellprat and Doblas-Reyes, 2016) used in event attribution. By contrast here we show that while forecast skill may be generically low in seasonal forecast ensembles this need not imply that the same models producing attribution

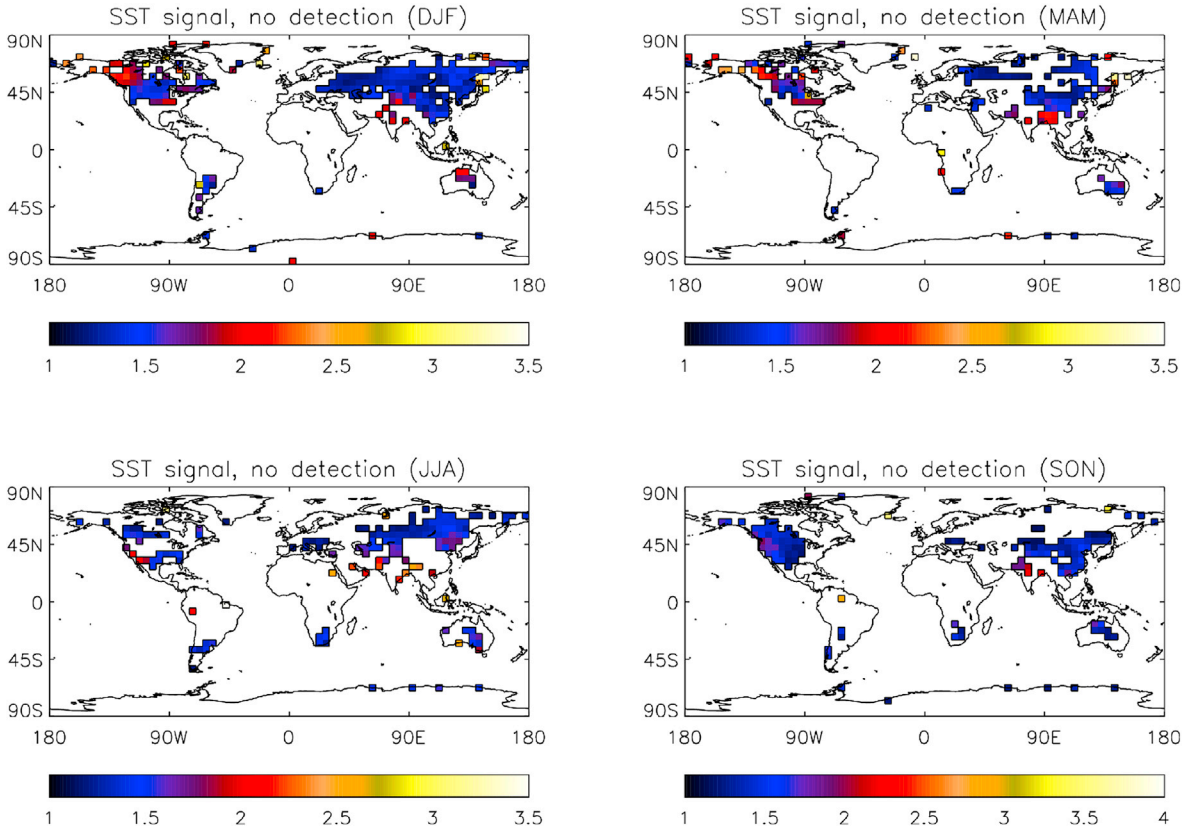


Fig. 10. Regions corresponding to case (ii) of Fig. 6 coloured with the value of $\tilde{\Re}$ where the model exhibits predictability ($\tilde{\Re} > \tilde{\Re}_c \Leftrightarrow H_0 : p < 5\%$) but SST influence is not detected ($\beta \sim 0$). In these regions there exists the possibility of model error as the model is making hindcasts departing from climatology but which are not significantly correlated with observations. However, across the seasons these regions are mainly those of relatively low model predictability and so failure to detect weak signals seems more likely. There are a small number of regions suggestive of genuine model error. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

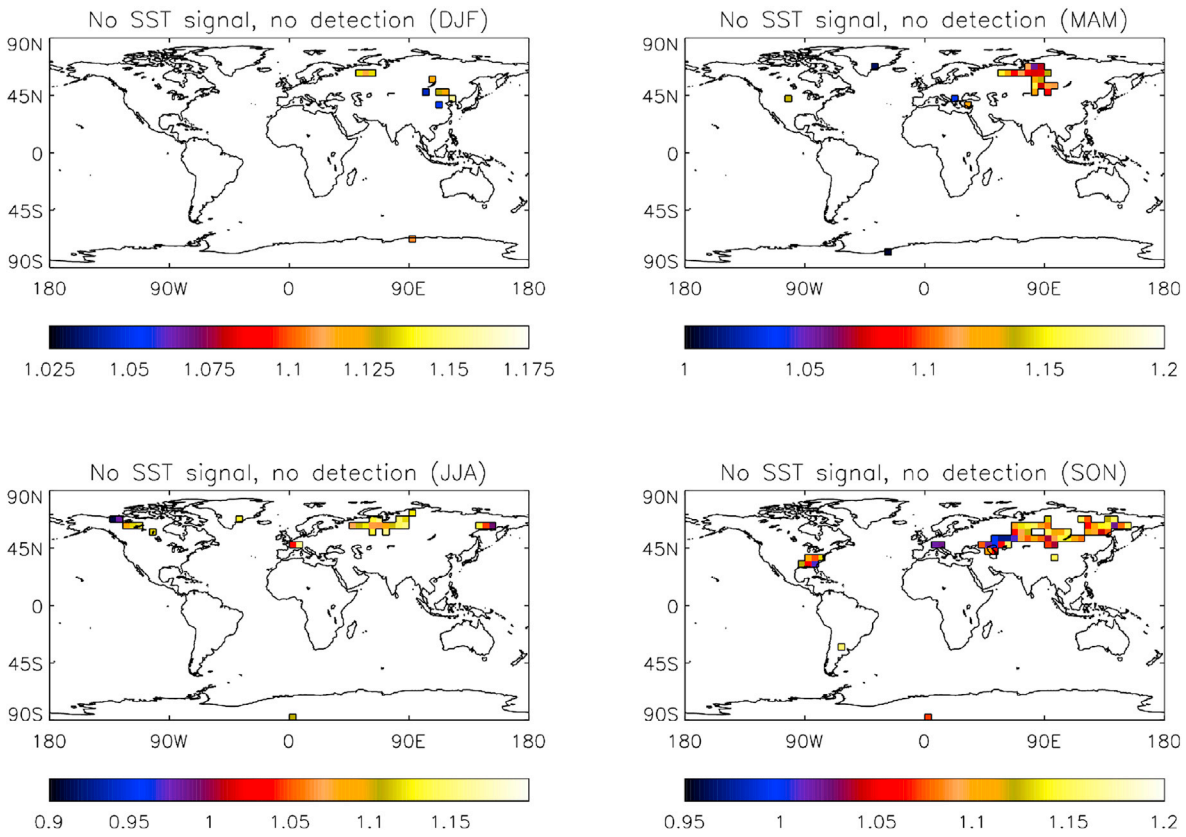


Fig. 11. Regions corresponding to case (iii) of Fig. 6 coloured with the value of $\tilde{\Re}$ where the model does not exhibit predictability ($\tilde{\Re} < \tilde{\Re}_c \Leftrightarrow H_0 : p > 5\%$) and SST influence is not detected ($\beta \sim 0$). In principle there could be model error but in practice it will be difficult to decide which is the case without new information and most likely there is simply very little SST influence in these regions. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

ensembles are not performing well, and may in fact be consistent with the same model capturing the complete predictive component in an attribution ensemble and even producing series consistent with a perfect model. The expectation of low skill has been associated, for example, with the low correlation of ensemble means and observed series (Weigel et al., 2008), $r(\langle x \rangle, y) = \mathcal{O}(0.1)$. However, as we have first argued and now demonstrated can be achieved, an attribution ensemble requires not perfect correlation but a response to the predictive components that is near perfect ($\widehat{\beta} \approx 1 \not\Rightarrow r(\langle x \rangle, y) \approx 1$). This is discussed further in Appendix A.

6. Case studies

In this section we give some examples of what could form part of an operational, systematic attribution assessment by considering four recent seasons from MAM 2016 to DJF 2017 which have been simulated by the large ensemble continuation experiments *historical(Nat)Ext* described in section 2.3. Operational analyses would be conducted every season at all possible locations irrespective of whether the most recent season constituted an extreme event, addressing selection bias as well as providing the possibility of a service with global reach.

Fig. 13 shows all 5° grid cells which have a top or bottom 5 record (for the period since 1960) in the most recent season for which we have run the system. It is obvious that a general warming must be taking place from the relative frequency of warm compared to cold records but it also demonstrates the need for the ability to perform systematic analyses of the regional contribution of secular climate changes versus interannual SST anomalies due to internal oceanic variability.

Below we will examine five essentially random locations, including examples of both warm and cold extremes appearing in Fig. 13, where we are able to quantify the different contributions to the probability ratio of secular climate changes and interannual variability forced by SST patterns. At each cell the validation procedure described in sections 4.1 and 5 has been applied followed by the bias correction of section 4.2 so that the likelihoods represent our best estimates of those that would result from a model with perfect timing and magnitude of response to SST. We focus on case (i) regions (influence of SST is present in the model and detected in historical observations) as examples of regions where use on an atmosphere-only model is adding value. We do not look for or consider impacts that may have been associated with these events.

For all cells for which we are able to conduct the validation (due to availability of observations) we use the ensembles to produce estimates of the change in probability of exceeding the seasonal mean temperature observed in the most recent simulated season subject to framing by the three different sets of conditions described in section 3:

1. 2016/17 seasonal SST: we construct density functions "2016/17 PDF" from the ensemble values taken strictly from the most recent season of interest, consisting of 525 members per experiment.
2. secular changes: we construct a density "Secular PDF" from residuals to the secular change over the full validation period which are centred at the value of the secular component in the most recent season, consisting therefore of 810 values per experiment.
3. climatology of a recent 30 year period (1984–2013): we construct a density "Recent PDF" from all values arising over the period 1984–2013, which is 450 values per experiment.

In other words condition 2. relaxes the conditioning on specific SSTs found in condition 1. mimicking the output of a coupled model to the extent that the secular component f is as approximation of the

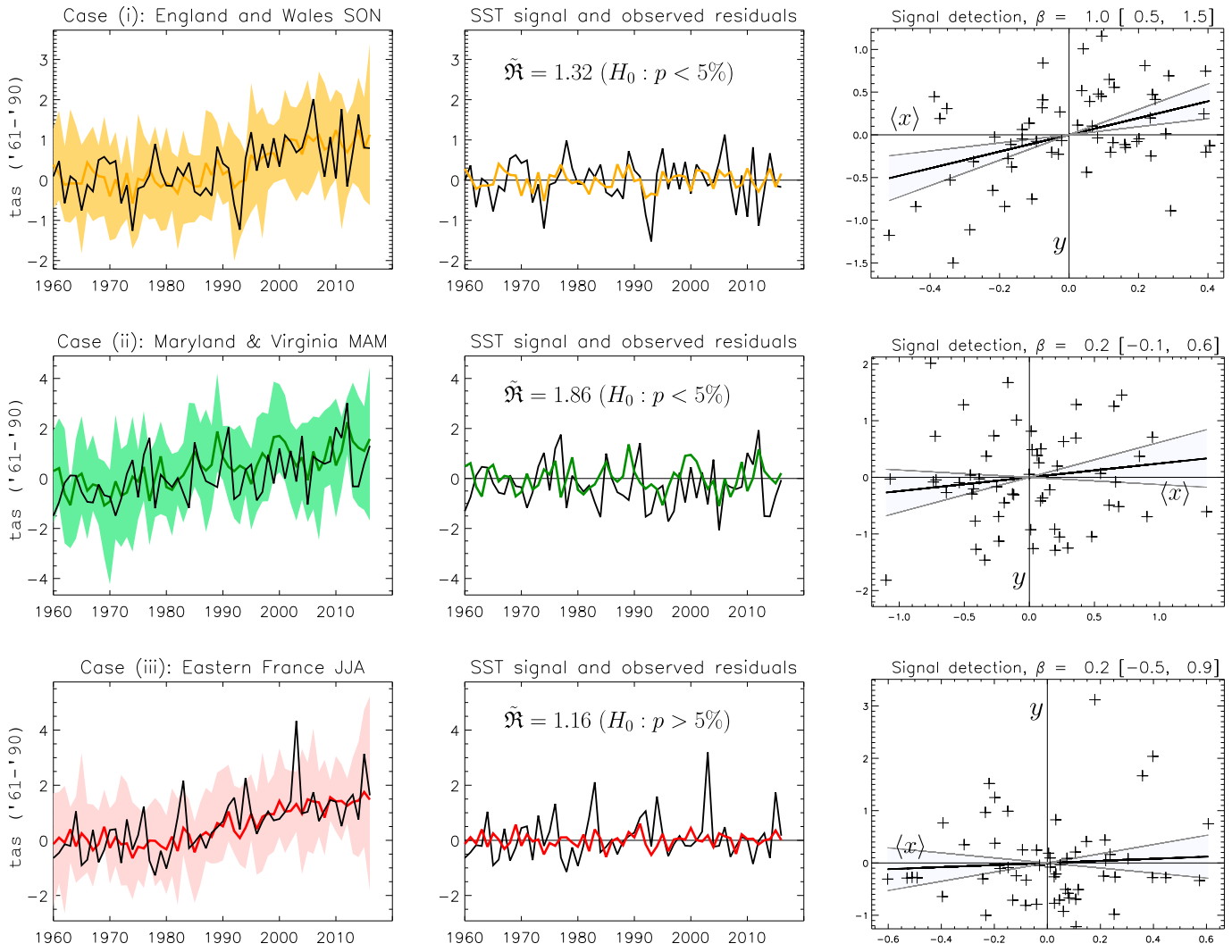


Fig. 12. Examples of seasonal timeseries falling into cases (i) - (iii) in rows. Left column: model envelope of X_a (filled colour), ensemble mean $\langle X \rangle$ (coloured line) and observed (black) time series Y. Middle column: SST signal $\langle x \rangle$ itself (coloured line), observations minus secular changes y (black) and an indication of signal strength $\tilde{\Re}$ and significance. Right hand panels are scatter plots of y versus $\langle x \rangle$ with linear relationships (gradients in the range $\hat{\beta} \pm \Delta\beta$) depicting interannual SST influence. Top row: Autumnal mean temperatures in England and Wales possess SST influence on the location of temperature distributions that is consistent with a perfect model (β consistent with 1). SST is not detected in either of cases (ii) and (iii). In case (ii), middle row, a significant SST signal is nevertheless present in the model indicating the possibility of model error. In case (iii), bottom row, the model does not contain a significant SST signal to detect so that as long as spectral properties of this region are well represented we would have no reason to suspect significant model error. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

anthropogenic contribution. In practice the secular component will contain a contribution from decadal scale oceanic internal variability which would hence vary between equivalent coupled model realisations. Condition 3. can be seen as the likelihood based on our recent experience, data which will contain trends.

Figs. 14–18 summarise the analysis and associated validation for each of our five examples. The top rows display time series of seasonal mean '61-'90 anomalies from 1960 - present of the observations (black, large dot at most recent observed value) superimposed on envelopes of the minimum and maximum values from the ALL experiment, the ensemble mean (solid coloured line) and dashed series representing the secular changes for both ALL (coloured) and NAT (grey). Next to these are the three pairs of density estimates just described (ALL are filled, NAT are depicted as outlines) together with the inversely bias adjusted observational threshold (black dashed line) used to calculate p_0 and p_1 and to their right are the three resulting pairs of return times (ALL red, NAT blue).

Our densities are estimated using the kernel method (Silverman, 1986), leading to probability distributions that will be robust where the density is estimated from many values so we quote values of the return times $\tau_0 = p_0^{-1}, \tau_1 = p_1^{-1}$ (and hence probability ratio $R = \tau_0/\tau_1$) with confidence if the values are not found to be above that represented by only a small handful of members. This upper bound is depicted as a dotted line at the return time corresponding to the 99th percentile of our 2016/17 distribution, exceeded by around 5 members, and as a dashed line at the return time typically exceeded by a single member.

Uncertainty bounds (vertical bars) on the return times are obtained by varying the inversely bias corrected observational threshold using the 5% and 95% confidence interval values of β. For the 2016/17 SST conditioned return times we have also calculated two further pairs of realisations produced by exploring sensitivity to the kernel bandwidth parameter η used in constructing the density functions. Alternate density functions are created with values of η either side of the optimal value η_{opt}, $\eta = 0.5\eta_{\text{opt}}$ and $\eta = 1.5\eta_{\text{opt}}$. These return times are depicted by dashed lines to the right of those obtained with $\eta = \eta_{\text{opt}}$.

On the bottom row the far left displays min/max envelopes of

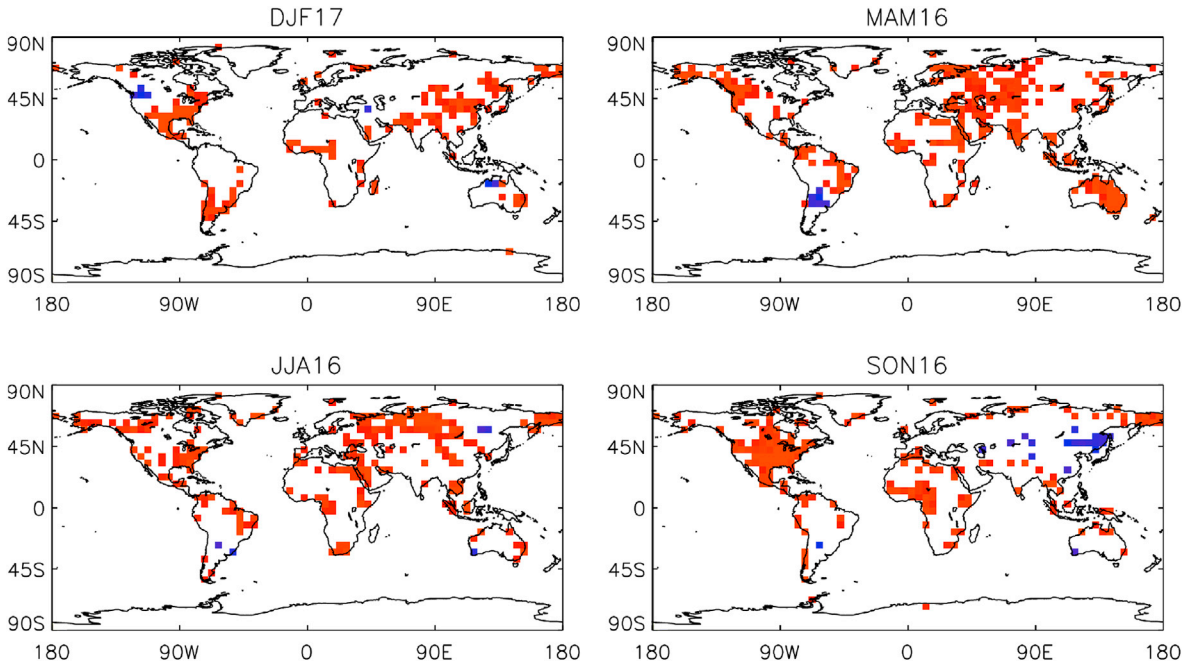


Fig. 13. HadCRUT4 cells (over land) with seasonal means in which the most recent simulated season is a top 5 or bottom 5 record over the period 1960–2017. Top 5 records are coloured red and bottom 5 blue. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

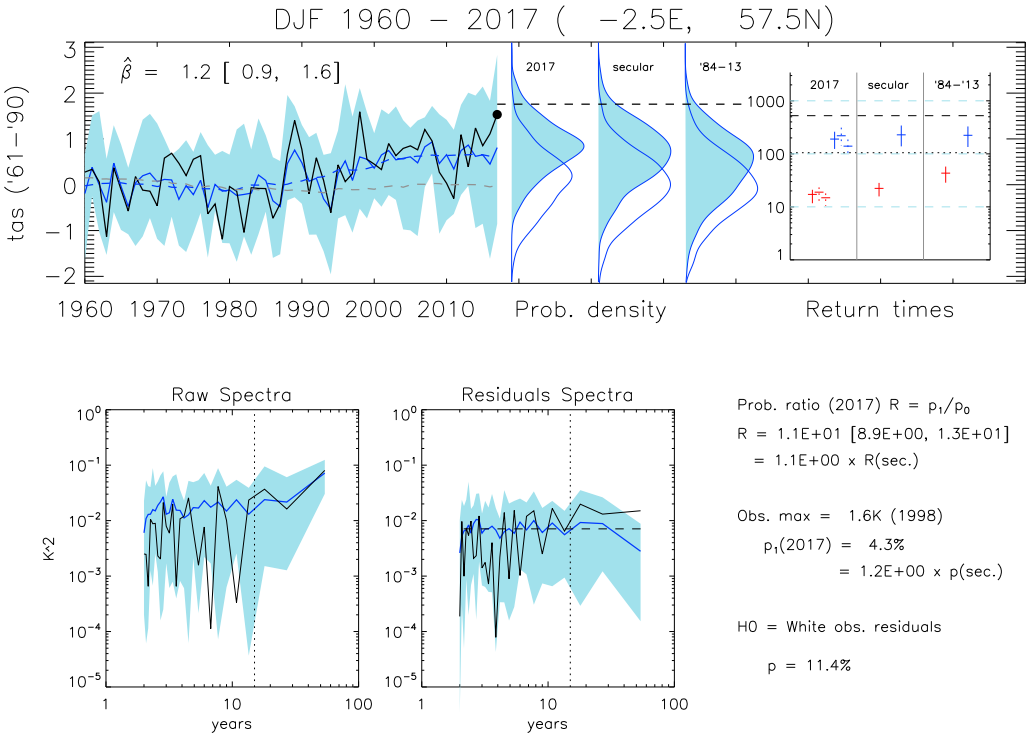


Fig. 14. Warm event primarily attributable to secular climate changes: December 2016–February 2017 in Scotland was the 2nd warmest winter since 1960 in the HadCRUT4 dataset. This event was made at least 9 times more likely by anthropogenic influence under 2016/17 SST conditions, which is only a 10% enhancement from the probability ratio obtained by only conditioning on secular changes, indicating secular climate change was primarily responsible for increasing the likelihood of this event. For legend see main text.

ensemble member power spectra (1960–2013) and their ensemble mean (solid coloured line), overlaid with that of the observations (black). A vertical dotted line separates variability below and above the 15-year time scale. In the middle is the same for member residuals r_a to the ensemble mean overlaid with observed residuals $\hat{e}$ to the best estimate signal $\hat{m}$. This is used to hypothesis test for the whiteness of the residuals of the observations, $\hat{e}$, supporting the regression and the form of $\Delta\beta$ we have used. The ensemble mean power spectrum can be used as a quick check on the whiteness of model noise about the SST signal and hence whether the regression analysis is near optimal. Test results are given bottom right, below details of probability values and ratios. Further details will be given in the first example.

First we consider a warm event that can be attributed primarily to secular climate changes. Fig. 14 summarises the analysis for the winter December 2016 to February 2017 in Scotland (5° × 5° region centred at 2.5°E, 57.5°N). This was the second warmest value in the period 1960–2017 in the HadCRUT4.5 dataset. It was around a 1 in 200 year event in the NAT world conditioned on DJF17 SST patterns but was increased in likelihood by factor of 11 (8.9–13 by exploring the minimum

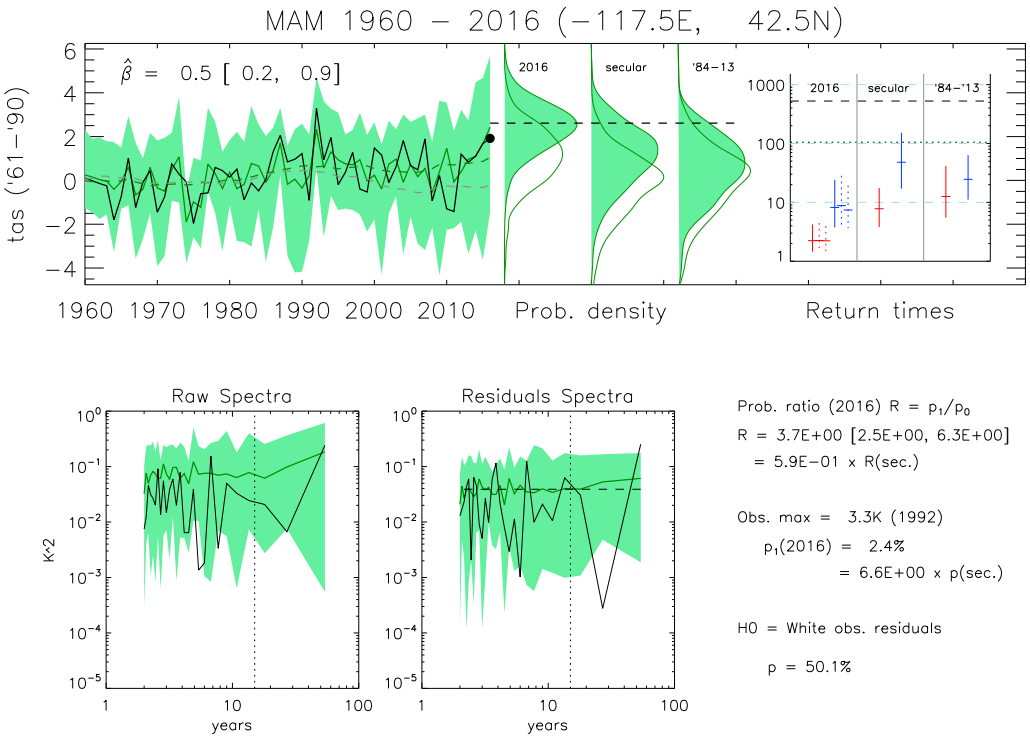


Fig. 15. Warm event with large SST pattern contribution: March–May 2016 in northern Nevada was the 4th warmest spring since 1960 in the HadCRUT4 dataset. In addition to an increase in the likelihood of such an event due to secular changes the response to specific SST patterns further amplified the likelihood of exceeding the observed threshold. For legend see main text.

and maximum probability ratios encountered in exploring both uncertainty in β and sensitivity to the kernel bandwidth η) an increase of 10% from the secular probability ratio. The density functions are virtually unchanged in location by SST conditioning, though the shapes do change, perhaps accounting for the small change in probability ratio with conditioning. Densities calculated from model values occurring over three recent decades indicates no significant change in the NAT return time from that found under secular and SST conditioning while the ALL return time is the highest of the three. This is the result of the presence of large variability about a trend. Observed residuals pass the whiteness test indicating that the assumptions of the analysis are defensible. In addition the raw spectra indicate a fair representation of variability across timescales. The conclusion is that the recent warm winter in Scotland was primarily the result of secular climate changes with only a small contribution from the specific oceanic state prevailing between December and February.

Fig. 15 shows that in the spring season March to May 2016 in a region corresponding to Northern Nevada and ranking 4th in the series since 1960, a clear shift in the likelihood of warm extremes took place due to the specific SST patterns occurring in this season. Under both ALL and NAT forcing scenarios this event was increased in likelihood by $\mathcal{O}(10)$. The probability ratio conditioned on SST, 3.7 (2.5–6.3) actually decreased from the secular estimate by around 40%, nevertheless indicating an increase in likelihood due to anthropogenic influence. The best estimate of the event return time was reduced by 75% from around 8 to 2 years by SST conditioning making it much more likely under these boundary conditions. This was an event first made more likely by secular climate change and then further amplified in likelihood by the regional response to SST patterns.

The austral winter of June to August 2016 in the far south west of Australia (the region around Perth), Fig. 16, was the 3rd coldest in the HadCRUT4 series and the ALL-forcings return times for an event at least this cold under 2016 and secular framings are around 100 years (and not distinguishable). NAT return times are closer to 10 years, indicating that this would be a cool event even in the absence of anthropogenic forcings. The probability ratio under JJA16 conditions is estimated at $R = 4.2(3.6 \text{ to } 6.8) \times 10^{-2}$ which is 30% of the probability ratio under secular conditioning. Nevertheless there remained a best estimate chance of breaking the regional *warm* record (set in 1983) of 7.6%, which was a reduction by 50% from the chance under secular conditioning, corresponding to the small SST enhancement in the likelihood of a cool event that season, as can be seen from a negative shift in the location of the density functions.

The cool spring of 2016 in Ireland, Fig. 17, while not a cold extreme, shows a strong enhancement in the likelihood of a cool season due to SST conditions, up to almost the 2 year return level. Despite a clear secular warming through anthropogenic forcing we find that SST conditioning meant there was only $\mathcal{O}(1\%)$ chance of breaking the spring warm record set recently (in 2007) which is a reduction of around 99% due to the shift by SST patterns.

Finally we consider a distinct non-event. The season from September to October 2016 in the region of southern Portugal and south western Spain, Fig. 18, has exhibited a strong warming trend in recent decades and the regional HadCRUT4 record of +1.5K was set in 2006 and almost equalled in 2009 and 2014. The seasonal mean of September to October 2016 meanwhile was entirely unexceptional, sitting around the median of values over the three decades 1984 to 2013. Nevertheless there was around a 1 in 4 chance of breaking the warm record in this season, a figure not much changed from the likelihood found under all SST conditions. The conclusion is that this region saw a fluctuation below a warming trend that is only likely to further increase the chance of hot extremes in the region in the near future and we would expect this record to be equalled or broken well within the next decade under continued secular warming.

7. Discussion

We have presented the development of the Met Office system for event attribution using a state of the art high resolution global climate model and outlined a suggestion for one part of an operational analysis.

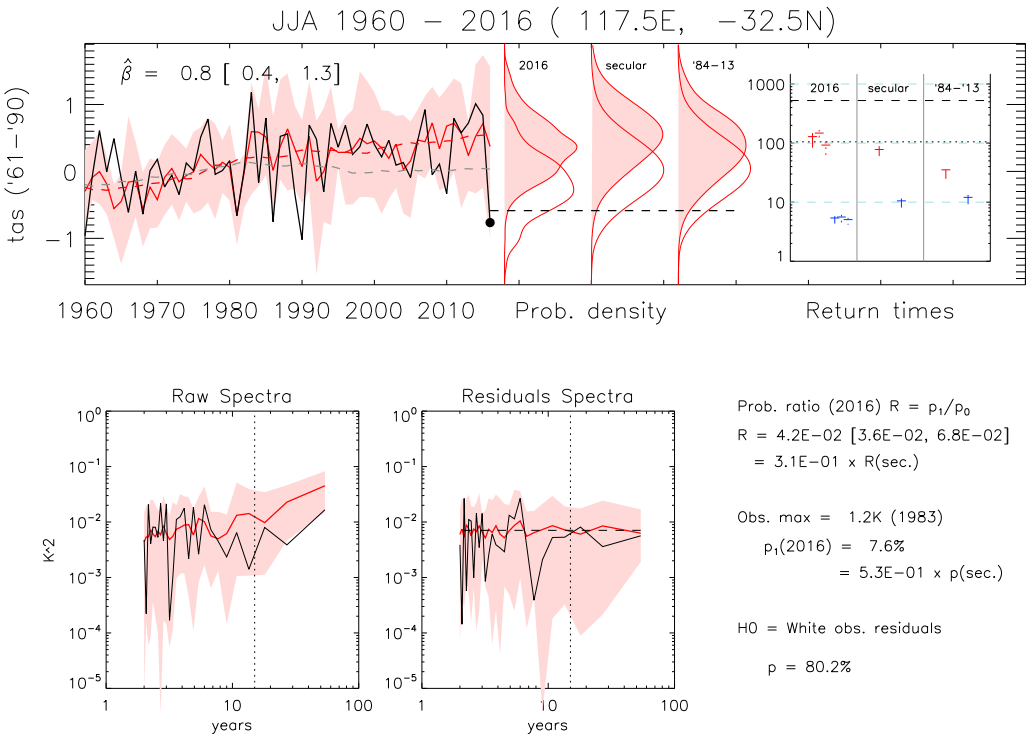


Fig. 16. Cold event with smaller SST pattern contribution: June–August 2016 in the far south-west of Australia (Perth) was the 3rd coldest austral winter since 1960 in the HadCRUT4 dataset. SST patterns had only a small influence on the likelihood of achieving these low seasonal mean temperatures with the event remaining around the 100 year return level. For legend see main text.

By framing the attribution question in different ways the analysis allows us to estimate the relative changes in the probability ratio of obtaining seasonal mean temperatures observed over the period March 2016–February 2017 due to secular climate changes versus specific seasonal SST conditions. The probability ratios we calculate are presented along side the results of a validation and after an appropriate bias correction. The aim was to be in a position to make calibrated statements regarding the changes in likelihood of extremes under circumstances when we can be confident that the model is representing influences when and where they should be present.

Event attribution concerns the causes of events which places predictability, and model skill, never far from the centre of the discussion. Not every framing of the event attribution question admits predictability into the equation however. By conditioning a study on realisations of the climate that hides predictable influences we may form an ensemble of perfect realisations that nevertheless lack any skill. We have therefore attempted to clarify the distinction between model skill and model performance in probabilistic hindcasts with attribution ensembles. In an attribution context, unlike in a forecasting context, the assessment of skill is not equivalent to the assessment of model performance. We therefore suggested a methodological approach to assess where we expect model skill to be important, as judged by the historical record.

First we assessed when a signal from the boundary conditions could be said to be present through the use of the predictable component in the ALL forcings historical ensemble. Where this predictable component is significant is where we require that the ensemble should also be skillful. Next we perform what amounts to a detection and attribution to boundary condition forcing where detection of boundary condition influence indicates that skill from sea surface temperature and sea ice patterns is present and that associated interannual changes in likelihoods should be judged to be represented well. We have demonstrated that a model can possess a near perfect representation of the relevant causal factors affecting likelihoods when skill as usually understood in a seasonal forecasting context is low. Intuition from seasonal forecasting should not be extended to event attribution without careful consideration.

We found that for seasonal mean near surface air temperatures over most of the oceans and European land areas the HadGEM3-A based system falls into the first of our three cases where the predictable component in the form of inter-seasonal changes in the ensemble mean is capturing real world predictability entering through boundary conditions. Here the system may be said to possess skill and be performing well. In the continental interiors we find large areas, falling into the second and third of our cases, where boundary condition forcing in the model is weak and where, in the absence of important missing responses, we would then require spectral properties of variability to be well represented.

We are able to identify a small fraction of regions, falling into the second of our cases, where it seems possible to say there could be model error in the form of strong signals for dependence of seasonal mean temperatures on boundary conditions which however are not detected in the historical record. It would be interesting to understand what is going on dynamically in these regions.

Of course, we also emphasise that in practice we require confidence in the physics and dynamics taking place within ensemble members in addition to statistical properties of time series and we have referred the reader to a large body of ongoing work in this direction.

There is plenty of scope for the extension and improvement of the methodology we have sketched out here. Firstly we could extend the analysis to different variables, shorter time scales and smaller spatial scales. The potential here is limited by the availability of quality observational data sets however and the HadGEM3-A system itself is forced with monthly SST data so that skill should not be expected to be present at temporal scales far below this. Nevertheless we may consider analyses at even the shortest temporal scales under an appropriate framing of the event attribution question.

We could consider methodological improvements. We have not made use of extreme value theory here but instead indicated where we would not expect small values of p_0, p_1 resulting from the kernel density method to be robust to finite ensemble effects. We could use TLS in place of the

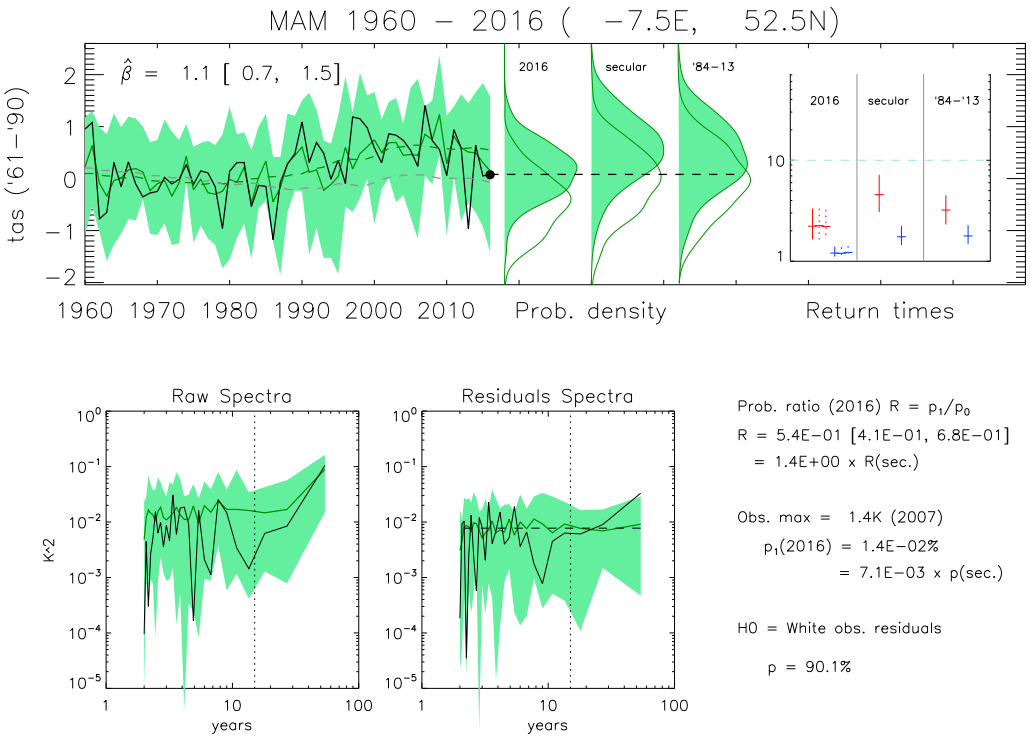


Fig. 17. Cool event with a larger SST pattern contribution: while March–May 2016 in Ireland was not a cold extreme in the HadCRUT4 dataset back to 1960 it was cooler than most years in the previous three decades. Return times are shown for events at least as cold as MAM 2016 The odds of breaking the seasonal record in 2016, denoted $p_1(2016)$ above, was reduced by around 99% by the response to specific SST patterns in this season. For legend see main text.

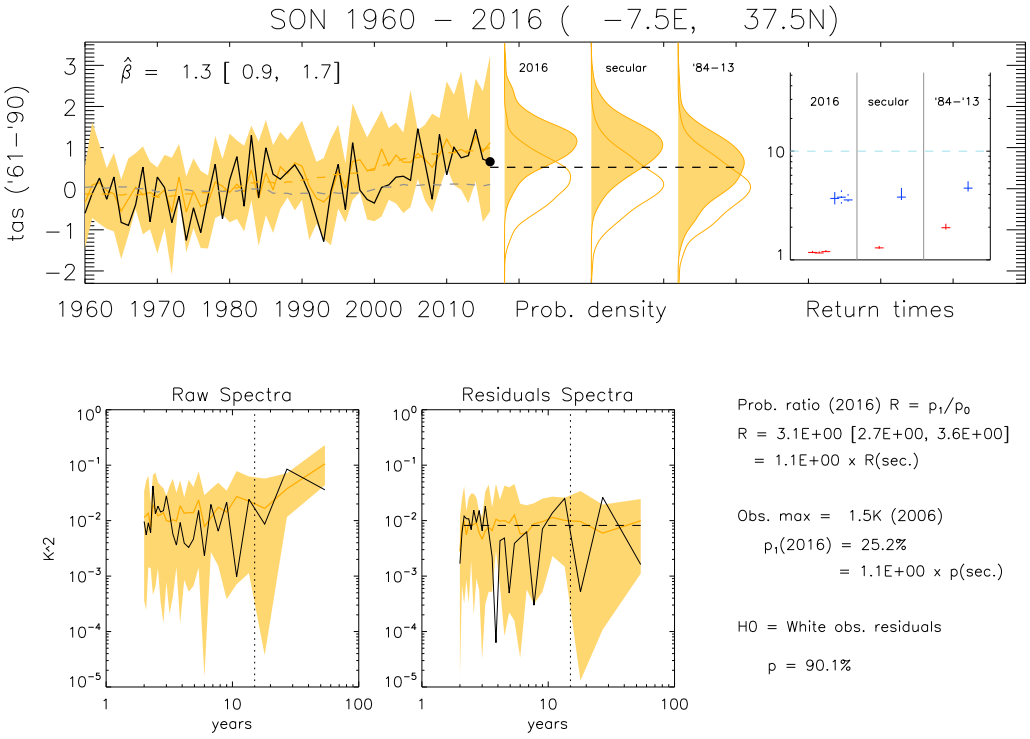


Fig. 18. Non-event attribution of an unexceptional season in southern Portugal. In Autumn (2016) there was around a 25% chance of breaking the regional 1960–2017 record of +1.5K set in 2006, a small enhancement by SST patterns of around 10% on top of the secular changes. For legend see main text.

OLS regression method to address the likely bias in $\widehat{\beta}$ toward zero in regions of low signal to noise. We have assumed the same secular component f to be present in both the model series X_a and observed series Y, which we justified here on account of experience with seasonal mean temperatures in our atmosphere-only model, and judging by the passing of tests on the residuals to our regression together with the sensitivity analysis in Appendix B this seems to have been defensible. The same justification could not be expected to be given for all variables (Dunn

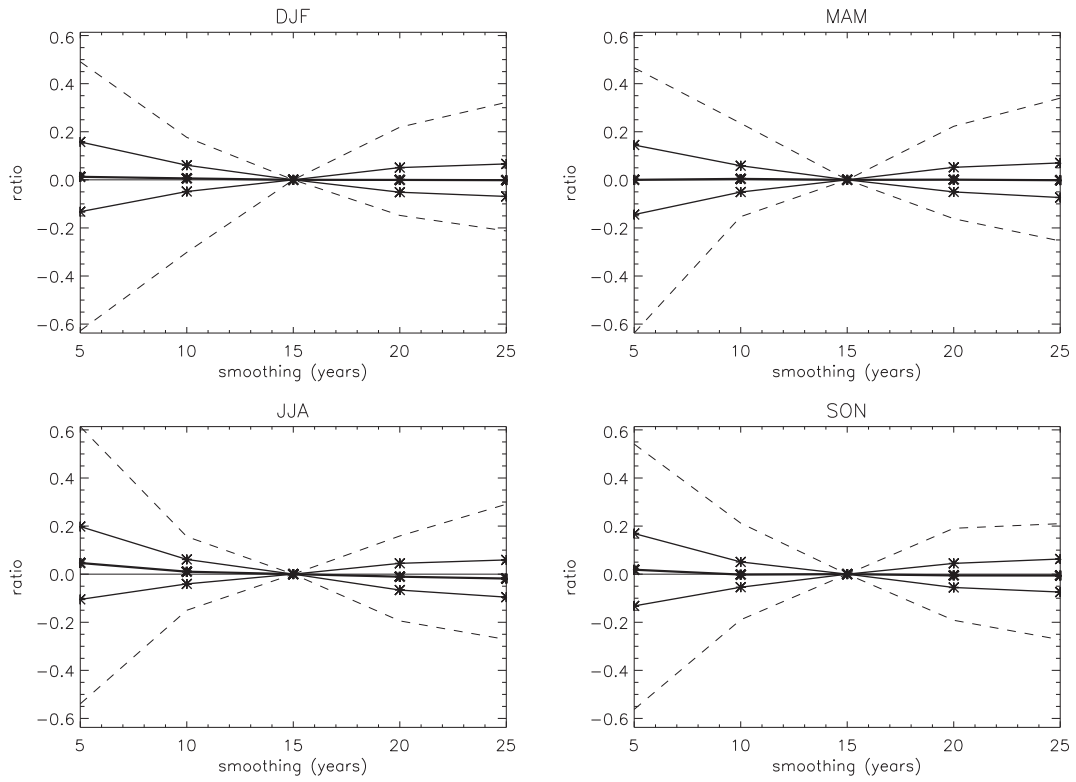


Fig. 19. Sensitivity of the value of $\widehat{\beta}$ to definition of the secular component f as the smoothing period is changed from 5 years up to 25 years. Plotted is the change in $\widehat{\beta}$ from that obtained using 15 year periods as a ratio to the value $\Delta\beta$ used as the uncertainty from the regression itself. The central solid line is the global mean of the ratio across all $5^\circ \times 5^\circ$ cells, the solid line either side is a standard deviation of the ratio and the dashed lines are the minimum and maximum cell values. The minimum and maximum cell values occur only very rarely and the sensitivity to the definition of f is therefore seen to be a second order effect compared to the uncertainty in the regression itself.

et al., 2017).

A simple extension, for example, could be to allow a one parameter modification of the secular component between the model and observations by a linear trend. The real importance of the secular component in our validation is to separate out predictive influences by timescale so that we can assess the presence and magnitude of short term influences. This is a matter of sorting signal (interannual changes in the location of the distribution) from noise (the centred distribution of states consistent with our framing) where the split between signal and noise is determined by the framing. In practice, as long as the regression results in believable residuals then we can defend our split between signal and noise.

We note that observed versus modelled seasonal mean near-surface air temperature trends in HadGEM3-A have been discussed elsewhere (Vautard et al., 2018) where it is claimed that the model possesses biases in full period (1960–2013) linear trends in parts of the European region. However, what the analysis really assessed was only where observed trends were ranked within those generated by ensemble members and did not ask whether the spatial pattern of rankings is itself significant. We have conducted an analogous perfect model analysis (offline) showing that the pattern of observed rankings is typical of individual ensemble members and so the conclusion that there exist biases here is unwarranted.

Less trivially we could modify the validation procedure to assess further degrees of freedom of the model error. The complete statistical description of possible model error we set out in section 4.1 involves a time series $\delta(t)$ of errors in the location of the distributions as well as all further moments of the distributions $g(t)$. In addition we used a finite ensemble mean $\langle x \rangle$ as our estimate of the boundary condition signal. In optimal fingerprinting studies there exist means of constructing noise free signals (Allen and Stott, 2003) and δ-type errors (Huntingford et al., 2006) but these involve additional control experiments that we do not currently possess in this context. An appropriate control experiment in our context would involve simulations forced with boundary conditions representing only secular changes, against which the relevant noise covariances could be obtained and a proper optimisation procedure conducted.

We conducted our analysis on the assumption that the orthogonal model error $\delta = 0$. As suggested in section 4.1 we could extend the analysis by looking for case (ii) regions where the model appears to contain a strong signal for regional response to specific SST but where this is not detected as evidence that we should consider $\delta \neq 0$.

The largest unexplored sensitivity of the results we present is likely to come through construction of the NAT boundary conditions (Christidis et al., 2018). Currently we remove estimates of anthropogenic influence on sea surface temperatures and sea ice concentrations using a multi-model mean of estimates (Stone and Pall, 2017). In the future we will consider including alternate realisations of the NAT boundary conditions as a matter of course.

We did not explore observational uncertainty or sensitivity to the choice of observational product. Observations enter our analysis through the validation, through the bias correction and then through thresholds. For seasonal mean near surface air temperatures at a spatial scale of $5^\circ \times 5^\circ$ we do not expect significant contributions to the uncertainty in β on account of this. In general, uncertainty in the observed threshold itself translates into use of a different event definition and resulting probability ratios, which is akin to a change of framing. Meanwhile, the bias correction procedure sees the historical observational series (to determine how to rescale both the SST signal about the secular component and the remaining variability about the SST signal) but by construction the adjustment we adopt would help to remove sensitivity by regionally enforcing comparability between model and observed series. It can be considered an advantage of the inverse correction approach that when

potential discrepancies between observational products indicate that we are uncertain of our 'truth' then we at least have a consistent interpretation of the adjusted observational thresholds through the mapping to model events. This is in addition to the advantage of a consistent treatment of events between ALL and NAT ensembles that the inverse correction tries to establish for the attribution question.

Finally, the system currently relies on extensions of historical climate forcings beyond 2005 (to 2010 depending) by the RCP 4.5 scenario. If anthropogenic forcing is indeed significantly different from this scenario then upon further upgrades of the system we will need to consider whether to perform extensions following alternate concentration pathways. The dominant forcing in our system however comes through the prescription of observed boundary conditions (as well as CMIP5 multi-model mean estimates of the changes due to net anthropogenic influence) so that, in the near term at least, a change such as RCP 4.5 to RCP 8.5 should not make a dramatic difference.

Competing financial interests

The authors declare no competing financial interests.

Acknowledgements

This work was supported by the Joint UK BEIS/Defra Met Office Hadley Centre Climate Programme (GA01101), the EUCLEIA project funded by the European Unions Seventh Framework Programme [FP7/20072013] under grant agreement no. 607085, the UK-China Research and Innovation Partnership Fund through the Met Office Climate Science for Service Partnership (CSSP) China as part of the Newton Fund, and the EUPHEME project which is part of ERA4CS, an ERA-NET initiated by JPI Climate, co-funded by the European Union (Grant 690462). The authors would also like to thank the following for additional advice and technical help: Jeff Knight, Warren Tennant, Dan Copsey, Jamie Kettleborough, Chris D Jones, Ag Stephens and the FCM, Climate Science IT Applications and Climate Research Unified Model teams, especially Erica Neininger and Stephen Haddad.

Appendix A. Bellprat & Doblas-Reyes model

The Bellprat & Doblas-Reyes (BDR) model (Bellprat and Doblas-Reyes, 2016) is a statistical model similar to that laid out in section 4.1 above and is intended to represent seasonal hindcasts produced using the atmosphere-only approach we use (referred to below as the ACE (Christidis et al., 2013a; Pall et al., 2011) approach), and specifically the Met Office HadGEM3-A based system. The BDR model is based on a model due to Weigel et al. (Weigel et al., 2008) by creating a pair of ensembles with and without secular trends. The central result of the BDR model is the over-estimation of the fraction of attributable risk (Allen, 2003) (FAR) in the presence of unreliability. Below we shall demonstrate why the BDR model is not a good model of ensembles produced using the ACE approach and how the treatment of the over-estimation of the FAR is potentially even peculiar to the BDR model itself, rather then being a general feature of attribution with seasonal hindcast ensembles.

Expressed in the notation of section 4.1 the BDR model represents the observed series Y as

$$Y = f + y \tag{18}$$

where f is a secular component (in their case a linear trend) and y is a stochastic component modelled as

$$y \sim \mathcal{N}(0, \sigma_y). \tag{19}$$

The statistical model for the ensemble of simulations is then constructed as

$$X_a^{\mathrm{BDR}} = \alpha^{\mathrm{BDR}} y + f + d^{\mathrm{BDR}} + \varepsilon_a^{\mathrm{BDR}}, \qquad \varepsilon_a^{\mathrm{BDR}} \sim \mathcal{N}(0, \sigma_{\varepsilon^{\mathrm{BDR}}}). \tag{20}$$

As such the ensemble members are represented by stochastic variations $\varepsilon_a^{\mathrm{BDR}}$ superimposed on some portion α^{BDR} of the observed variations y about a common trend f (with common model response error d^{BDR}). The portion α^{BDR} is kept small and fixed (≈ 0.1) to represent the low skill normally observed in seasonal hindcasts. α^{BDR} is actually defined (Weigel et al., 2008) as the ensemble mean correlation coefficient between the observations and ensemble members, which explains why it often takes this low value. The component y could be split as $y = m + e$ as we do in section 4.1 but BDR find no need to do so because the observations are not seen as fluctuations about a predictable component, m. Instead they consider the ensemble to be a forecasters ensemble in which the truth y is hoped to sit near the best estimate (centre) of an ensemble generated by a good model; this is what occurs in their model if we take $\alpha^{\mathrm{BDR}} = 1,\ d^{\mathrm{BDR}} = 0$.

In the BDR model the observed fluctuations are therefore taken as a predictive component of the model ensemble along with the common trend (and model error). If the ensemble were of infinite size then we would have

$$\lim_{n_{\mathrm{ens}} \to \infty} \left\langle X_a^{\mathrm{BDR}} \right\rangle = \alpha^{\mathrm{BDR}} y + f \quad \Leftrightarrow \quad \lim_{n_{\mathrm{ens}} \to \infty} \left\langle x_a^{\mathrm{BDR}} \right\rangle = \alpha^{\mathrm{BDR}} (m + e). \tag{21}$$

This is contrary to the expectation we should have of ensembles produced with the ACE approach, including the HadGEM3-A based system, where the observations should look like a single ensemble member. The infinite size limit of the ACE approach gives

$$\lim_{n_{\mathrm{ens}} \to \infty} \left\langle X_a \right\rangle = \mu + f \quad \Leftrightarrow \quad \lim_{n_{\mathrm{ens}} \to \infty} \left\langle x_a \right\rangle = \alpha m + \delta, \tag{22}$$

which says that the predictive factor in the ensemble location is the signal component.

This is not a matter of the interpretation of the ensemble but of the generation of the ensemble itself. The ensembles generated with the HadGEM3-A system produce realisations that have essentially forgotten their initialisation and should not be interpreted as hindcasts in the same manner as a seasonal forecast in which initialisation (in principle) plays an important role, such as for lagged initisliation ensembles of the sort that can be used for

seasonal forecast validation.

The central claim of over-estimated FAR in the presence of low reliability arises due to a constraint on the total variance imposed in the BDR model. The width $\sigma_{\varepsilon^{\mathrm{BDR}}}$ of the model ensemble about the model mean $\alpha^{\mathrm{BDR}}y + f + d^{\mathrm{BDR}}$ is constrained such that the total width of the ensemble timeseries σ_X and of the observed series σ_Y match, thus

$$\sigma_{\varepsilon^{\mathrm{BDR}}} = \sqrt{\sigma_Y^2 - (\alpha^{\mathrm{BDR}})^2\sigma_y^2 - \sigma_d^2}. \tag{23}$$

The effect is that the width $\sigma_{\varepsilon^{\mathrm{BDR}}}$ must become smaller in the presence of larger model error σ_d, narrowing the model distributions and leading naturally to decreased values of the tail probabilities which translates into larger values of the FAR. The correct statistical model of ACE ensembles will contain a different relationship between model error and skill. In this study we have adopted a bias correction which scales the width of distributions and so issues similar to those highlighted by BDR will probably apply in the presence of model error, but given the differences in construction of the statistical models we should be cautious about applying the results of the BDR analysis to attribution using ensembles such as ours. Note that it is entirely possible that a modelled region has zero skill but is nevertheless a statistically perfect representation of the real world, resulting in no bias in the FAR.

Appendix B. Sensitivity to definition of secular component

The secular component is defined as the 15 year (tapering at ends to 8 year) smoothed ensemble mean series at a cell. Sensitivity of our analyses to the secular component would enter through the value of the regression coefficients used to detect the interannual variations of the ensemble mean in the observations and which are involved in the subsequent bias correction.

We explore the sensitivity to this definition by recalculating the best estimate regression coefficient $\widehat{\beta}$ at each cell in every season for a number of different smoothing periods from 5 years up to 25 years. Our analyses are considered to be insensitive to the definition of the secular component if the resulting deviation of $\widehat{\beta}$ is small compared to the fit uncertainty $\Delta\beta$ already included. We therefore examine the ratio of deviations in $\widehat{\beta}$ to the 5%–95% range of $\Delta\beta$ at each cell. Fig. 19 shows the area weighted mean of this ratio, as well as one standard deviation either side of this mean and the minimum and maximum field values.

In all four seasons we find that the ratio of smoothing sensitivity in $\widehat{\beta}$ to that of the fit is almost always very small. This is because the regression analysis primarily picks up on interannual variations found in both the ensemble mean and observations that remain untouched by the removal of secular components with smoothing scales much in excess of 1, i.e. the model has real interannual skill. The ratio increases either side of 15 years, particularly for 5 year smoothing where we begin to remove a little of the interannual variation that is responsible for model skill. A visual inspection of the ratio fields confirms that the minimum and maximum field values are very uncommon and that the smoothing sensitivity is reliably small compared to fit uncertainty.

References

Allen, Myles, 2003. Liability for climate change. Nature 421 (6926), 891–892.

Allen, M.R., Stott, P.A., 2003. Estimating signal amplitudes in optimal fingerprinting, part i: Theory. Clim. Dynam. 21 (5), 477–491.

Allen, M.R., Tett, S.F.B., 1999. Checking for model consistency in optimal fingerprinting. Clim. Dynam. 15 (6), 419–434.

Angélil, Oliver, Perkins-Kirkpatrick, Sarah, Alexander, Lisa V., Stone, Dáithí, Donat, Markus G., Wehner, Michael, Shiogama, Hideo, Ciavarella, Andrew, Christidis, Nikolaos, 2016. Comparing regional precipitation and temperature extremes in climate model and reanalysis products. Weather and Climate Extremes 13, 35–43.

Angélil, Oliver, Stone, Daíthí, Perkins-Kirkpatrick, Sarah, Alexander, Lisa V., Wehner, Michael, Shiogama, Hideo, Wolski, Piotr, Ciavarella, Andrew, Christidis, Nikolaos, 2017. On the nonlinearity of spatial scales in extreme weather attribution statements. Clim. Dynam. 1–14.

Bellprat, Omar, Doblas-Reyes, Francisco, 2016. Attribution of extreme weather and climate events overestimated by unreliable climate simulations. Geophys. Res. Lett. 43 (5), 2158–2164.

Best, M.J., Pryor, M., Clark, D.B., Rooney, G.G., Essery, R., Ménard, C.B., Edwards, J.M., Hendry, M.A., Porson, A., Gedney, N., et al., 2011. The joint UK land environment simulator (JULES), model description–part 1: energy and water fluxes. Geosci. Model Dev. (GMD) 4 (3), 677–699.

Bond, Tami C., Bhardwaj, Ekta, Dong, Rong, Jogani, Rahil, Jung, Soonkyu, Roden, Christoph, Streets, David G., Trautmann, Nina M., 2007. Historical emissions of black and organic carbon aerosol from energy-related combustion, 1850–2000. Global Biogeochem. Cycles 21 (2).

Burke, C., Stott, P., Sun, Y., Ciavarella, A., 2016. Attribution of extreme rainfall in southeast China during May 2015. Bull. Am. Meteorol. Soc. 97 (12), S92–S96.

C20C+ Detection and Attribution Project. http://portal.nersc.gov/c20c/.

Christiansen, B., Christidis, N., Ciavarella, A., Alvarez-Castro, C., Bellprat, O., Colfescu, I., Cowan, T., Doblas-Reyes, F., Eden, J., Hempelmann, N., Klehmet, K., Lott, F., Nangini, C., van Oldenborgh, G.J., Orth, R., Radanovics, S., Stott, P., Tett, S., Vautard, R., Wilcox, L., Yiou, P., 2017. Was the Cold European Winter 2009-2010 Modified by Anthropogenic Climate Change? an Attribution Study (In preparation).

Christidis, Nikolaos, Stott, Peter A., 2014. Change in the odds of warm years and seasons due to anthropogenic influence on the climate. J. Clim. 27 (7), 2607–2621.

Christidis, Nikolaos, Stott, Peter A., 2015. Extreme rainfall in the United Kingdom during winter 2013/14: the role of atmospheric circulation and climate change. Bull. Am. Meteorol. Soc. 96 (12), S46–S50.

Christidis, Nikolaos, Stott, Peter A., Scaife, Adam A., Arribas, Alberto, Jones, Gareth S., Copsey, Dan, Knight, Jeff R., Tennant, Warren J., 2013. A new hadgem3-a-based system for attribution of weather-and climate-related extreme events. J. Clim. 26 (9), 2756–2783.

Christidis, Nikolaos, Stott, Peter A., Hegerl, Gabriele C., Betts, Richard A., 2013. The role of land use change in the recent warming of daily extreme temperatures. Geophys. Res. Lett. 40 (3), 589–594.

Christidis, Nikolaos, Jones, Gareth S., Stott, Peter A., 2015. Dramatically increasing chance of extremely hot summers since the 2003 european heatwave. Nat. Clim. Change 5 (1), 46.

Christidis, Nikolaos, McCarthy, Mark, Ciavarella, Andrew, Stott, Peter A., 2016. Human contribution to the record sunshine of winter 2014/15 in the United Kingdom. Bull. Am. Meteorol. Soc. 97 (12), S47–S50.

Christidis, Nikolaos, Ciavarella, Andrew, Stott, Peter A., 2018. Different ways of framing event attribution questions: the example of warm and wet winters in the UK similar to 2015/16. J. Clim. 2018.

Cionni, Irene, Eyring, Veronika, Lamarque, Jean-Francois, Randel, W.J., Stevenson, D.S., Wu, F., Bodeker, G.E., Shepherd, T.G., Shindell, D.T., Waugh, D.W., 2011. Ozone database in support of CMIP5 simulations: results and corresponding radiative forcing. Atmos. Chem. Phys. 11 (21), 11267–11292.

Clark, D.B., Mercado, L.M., Sitch, S., Jones, C.D., Gedney, N., Best, M.J., Pryor, M., Rooney, G.G., Essery, R.L.H., Blyth, E., et al., 2011. The joint UK land environment simulator (JULES), model description–part 2: carbon fluxes and vegetation dynamics. Geosci. Model Dev. (GMD) 4 (3), 701–722.

Climate science for service partnership china. http://www.metoffice.gov.uk/research/collaboration/cssp-china.

CMIP5 recommended data, accessed 18/09/13. http://www.pik-potsdam.de/~mmalte/rcps/, 2013.

Davies, T., Cullen, M.J.P., Malcolm, A.J., Mawson, M.H., Staniforth, A., White, A.A., Wood, N., 2005. A new dynamical core for the met office's global and regional modelling of the atmosphere. Q. J. R. Meteorol. Soc. 131 (608), 1759–1782.

Dunn, R.J.H., Willett, K.M., Ciavarella, A., Stott, P.A., 2017. Comparison of land surface humidity between observations and CMIP5 models. Earth System Dynamics 8 (3), 719–747.

Eade, Rosie, Smith, Doug, Scaife, Adam, Wallace, Emily, Dunstone, Nick, Hermanson, Leon, Robinson, Niall, 2014. Do seasonal-to-decadal climate predictions underestimate the predictability of the real world? Geophys. Res. Lett. 41 (15), 5620–5628.

Earth system grid federation (ESGF) portal at the stfc centre for environmental data (CEDA). https://esgf-index1.ceda.ac.uk/projects/esgf-ceda/.

Eden, Jonathan M., Wolter, Klaus, Otto, Friederike EL., van Oldenborgh, Geert Jan, 2016. Multi-method attribution analysis of extreme precipitation in Boulder, Colorado. Environ. Res. Lett. 11 (12), 124009.

EUropean CLimate and weather Events: Interpretation and Attribution (EUCLEIA), howpublished=http://eucleia.eu/.

Herring Stephanie, C., Hoerling Martin, P., Peterson Thomas, C., Stott Peter, A., 2014. Explaining extreme events of 2013 from a climate perspective. Bull. Am. Meteorol. Soc. 95 (9), S1–S104.
Herring Stephanie, C., Hoerling Martin, P., Kossin James, P., Peterson Thomas, C., Stott Peter, A., 2015. Explaining extreme events of 2014 from a climate perspective. Bull. Am. Meteorol. Soc. 96 (12), S1–S172.
Herring Stephanie, C., Hoell, A., Hoerling Martin, P., Kossin James, P., Schreck, C.J., Stott Peter, A., 2016. Explaining extreme events of 2015 from a climate perspective. Bull. Am. Meteorol. Soc. 97 (12), S1–S145.
Huntingford, Chris, Stott, Peter A., Allen, Myles R., Hugo Lambert, F., 2006. Incorporating model uncertainty into attribution of observed temperature change. Geophys. Res. Lett. 33 (5).
Jones, Gareth S., Stott, Peter A., Christidis, Nikolaos, 2013. Attribution of observed historical near–surface temperature variations to anthropogenic and natural causes using CMIP5 simulations. J. Geophys. Res.: Atmospheres 118 (10), 4001–4024.
Klein Goldewijk, Kees, Beusen, Arthur, Van Drecht, Gerard, De Vos, Martine, 2011. The HYDE 3.1 spatially explicit database of human-induced global land-use change over the past 12,000 years. Global Ecol. Biogeogr. 20 (1), 73–86.
Knight, Jeff R., Maidens, Anna, Watson, Peter A.G., Andrews, Martin, Belcher, Stephen, Brunet, Gilbert, Fereday, David, Folland, Chris K., Scaife, Adam A., Slingo, Julia, 2017. Global meteorological influences on the record UK rainfall of winter 2013-14. Environ. Res. Lett. 12 (7).
Lamarque, J.-F., Bond, Tami C., Eyring, Veronika, Granier, Claire, Heil, Angelika, Klimont, Z., Lee, D., Liousse, Catherine, Mieville, Aude, Owen, Bethan, et al., 2010. Historical (1850–2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: methodology and application. Atmos. Chem. Phys. 10 (15), 7017–7039.
J Lean. Calculations of solar irradiance: monthly means from 1882 to 2008, annual means from 1610 to 2008. 2009. Treatment described in Jones, C. D., et al., 2011, "The HadGEM2-ES implementation of CMIP5 centennial simulations." Geosci. Model Development 4.3: 543-570.
MacLachlan, C., Arribas, A., Peterson, K.A., Maidens, A., Fereday, D., Scaife, A.A., Gordon, M., Vellinga, M., Williams, A., Comer, R.E., et al., 2015. Global seasonal forecast system version 5 (glosea5): a high-resolution seasonal forecast system. Q. J. R. Meteorol. Soc. 141 (689), 1072–1084.
Meinshausen, Malte, Smith, Steven J., Calvin, K., Daniel, John S., Kainuma, M.L.T., Lamarque, J.F., Matsumoto, K., Montzka, S.A., Raper, S.C.B., Riahi, K., et al., 2011. The rcp greenhouse gas concentrations and their extensions from 1765 to 2300. Climatic Change 109 (1–2), 213–241.
Meiyappan, Prasanth, Jain, Atul K., 2012. Three distinct global estimates of historical land-cover change and land-use conversions for over 200 years. Front. Earth Sci. 6 (2), 122–139.
Morice Colin, P., Kennedy John, J., Rayner Nick, A., Jones Phil, D., 2012. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 data set. J. Geophys. Res. Atmos. 117 (D8), 1984–2012.
Otto, Friederike EL., Massey, N., Oldenborgh, G.J., Jones, R.G., Allen, M.R., 2012. Reconciling two approaches to attribution of the 2010 Russian heat wave. Geophys. Res. Lett. 39 (4).
Pall, Pardeep, Aina, Tolu, Stone, Daithi A., Stott, Peter A., Nozawa, Toru, Hilberts, Arno GJ., Lohmann, Dag, Allen, Myles R., 2011. Anthropogenic greenhouse gas contribution to flood risk in england and wales in autumn 2000. Nature 470 (7334), 382.
Peterson Thomas, C., Stott Peter, A., Herring Stephanie, C., 2012. Explaining extreme events of 2011 from a climate perspective. Bull. Am. Meteorol. Soc. 93 (7), 1041–1067.
Peterson Thomas, C., Hoerling Martin, P., Stott Peter, A., Herring Stephanie, C., 2013. Explaining extreme events of 2012 from a climate perspective. Bull. Am. Meteorol. Soc. 94 (9), S1–S74.
Qian, C., Wang, J., Dong, S., Hong, Y., Burke, C., Ciavarella, A., Dong, B., Freychet, N., Lott, F.C., Tett, S.F.B., 2018. Human influence on the record-breaking cold event in january of 2016 in eastern China. Bull. Am. Meteorol. Soc. 99 (1), S118–S122.
Rayner, N.A., Parker, De E., Horton, E.B., Folland, C.K., Alexander, L.V., Rowell, D.P., Kent, E.C., Kaplan, A., 2003. Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. J. Geophys. Res.: Atmospheres 108 (D14).
Sansom, Philip G., Ferro, Christopher AT., Stephenson, David B., Goddard, Lisa, Mason, Simon J., 2016. Best practices for postprocessing ensemble climate forecasts. Part i: selecting appropriate recalibration methods. J. Clim. 29 (20), 7247–7264.
Sato, M., Lacis, A., Hansen, J., Thomason, L., 2006. 1993. Extended post-2000 as described in Stott, Peter A., et al. "Transient climate simulations with the HadGEM1 climate model: causes of past warming and future climate change. J. Clim. 2763–2782, 19.12. http://data.giss.nasa.gov/modelforce/strataer.
Siegert, Stefan, Stephenson, David B., Sansom, Philip G., Scaife, Adam A., Eade, Rosie, Arribas, Alberto, 2016. A bayesian framework for verification and recalibration of ensemble forecasts: how uncertain is nao predictability? J. Clim. 29 (3), 995–1012.
Silverman, Bernard W., 1986. Density Estimation for Statistics and Data Analysis, vol. 26. CRC press.
Smith, S.J., Conception, E., Andres, R., Lurz, J., 2004. Historical Sulfur Dioxide Emissions 1850-2000: Methods and Results.
Smith, Steven J., Pitcher, Hugh, Wigley, Tom ML., 2001. Global and regional anthropogenic sulfur dioxide emissions. Global Planet. Change 29 (1), 99–119.
Stone, D.A., Pall, P., 2017. A Benchmark Estimate of the Effect of Anthropogenic Emissions on the Ocean Surface (In preparation).
Stott, P.A., Allen, M., Christidis, N., Dole, R., Hoerling, M., Huntingford, C., Pall, P., Perlwitz, J., Stone, D., 2013. Attribution of weather and climate-related events. Climate science for serving society. Springer 307–337.
Stott, Peter A., Christidis, Nikolaos, Otto, Friederike EL., Sun, Ying, Vanderlinden, Jean-Paul, van Oldenborgh, Geert Jan, Vautard, Robert, von Storch, Hans, Walton, Peter, Pascal Yiou, et al., 2016. Attribution of extreme weather and climate-related events. Wiley Interdisciplinary Reviews: Climate Change 7 (1), 23–41.
Taylor, Karl E., Doutriaux, Charles, 2010. CMIP5 Model Output Requirements: File Contents and Format, Data Structure and Metadata. http://cmip-pcmdi.llnl.gov/cmip5/docs/CMIP5_output_metadata_requirements.pdf.
Trenberth, Kevin E., Fasullo, John T., Shepherd, Theodore G., 2015. Attribution of climate extreme events. Nat. Clim. Change 5 (8), 725–730.
Taylor, Karl E., Williamson, David, Zwiers, Francis, 2000. The Sea Surface Temperature and Sea-ice Concentration Boundary Conditions for AMIP II Simulations. Program for Climate Model Diagnosis and Intercomparison. Lawrence Livermore National Laboratory, University of California.
Taylor, Karl E., Stouffer, Ronald J., Meehl, Gerald A., 2012. An overview of CMIP5 and the experiment design. Bull. Am. Meteorol. Soc. 93 (4), 485–498.
Tennant Warren, J., Shutts Glenn, J., Alberto, Arribas, Thompson Simon, A., 2011. Using a stochastic kinetic energy backscatter scheme to improve MOGREPS probabilistic forecast skill. Mon. Weather Rev. 139 (4), 1190–1206.
Vautard, R., Christidis, N., Ciavarella, A., Alvarez-Castro, C., Bellprat, O., Christiansen, B., Colfescu, I., Cowan, T., Doblas-Reyes, F., Eden, J., Hauser, M., Hegerl, G., Hempelmann, N., Klehmet, K., Lott, F.C., Nangini, C., Orth, R., Radanovics, S., Seneviratne, S.I., van Oldenborgh, G.J., Stott, P.A., Tett, S., 2018. Evaluation of the HadGEM3-A simulations in view of climate and weather event human influence attribution in Europe. Clim. Dynam. submitted for publication. https://doi.org/10.1007/s00382-018-4183-6.
Von Storch, Hans, Zwiers, Francis W., 2001. Statistical Analysis in Climate Research. Cambridge university press.
Walters, D., Brooks, M., Boutle, I., Melvin, T., Stratton, R., Vosper, S., Wells, H., Williams, K., Wood, N., Allen, T., Bushell, A., Copsey, D., Earnshaw, P., Edwards, J., Gross, M., Hardiman, S., Harris, C., Heming, J., Klingaman, N., Levine, R., Manners, J., Martin, G., Milton, S., Mittermaier, M., Morcrette, C., Riddick, T., Roberts, M., Sanchez, C., Selwood, P., Stirling, A., Smith, C., Suri, D., Tennant, W., Vidale, P.L., Wilkinson, J., Willett, M., Woolnough, S., Xavier, P., 2016. The met office unified model global atmosphere 6.0/6.1 and jules global land 6.0/6.1 configurations. Geosci. Model Dev. Discuss. (GMDD) 1–52, 2016.
Wehner, M., Stone, D., Shiogama, H., Wolski, P., Ciavarella, A., Christidis, N., Krishnan, H., 2017. Early 21st Century Anthropogenic Changes in Extremely Hot Days as Simulated by the C20C+ Detection and Attribution Multi-model Ensemble (submitted to WACE).
Weigel Andreas, P., Liniger, M.A., Appenzeller, C., 2008. Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? Q. J. R. Meteorol. Soc. 134 (630), 241–260.
Weisheimer, Antje, Palmer, T.N., 2014. On the reliability of seasonal climate forecasts. J. R. Soc. Interface 11 (96), 20131162.
Wilcox, L.J., Yiou, P., Hauser, M., Lott, F.C., Oldenborgh, G.J., Colfescu, I., Dong, B., Hegerl, G., Shaffrey, L., Sutton, R., 2017. Multiple Perspectives on the Attribution of the Extreme European Summer of 2012 to Climate Change (In preparation).
Wood, Nigel, Staniforth, Andrew, White, Andy, Allen, Thomas, Diamantakis, Michail, Gross, Markus, Melvin, Thomas, Smith, Chris, Vosper, Simon, Zerroukat, Mohamed, et al., 2014. An inherently mass-conserving semi-implicit semi-Lagrangian discretization of the deep-atmosphere global non-hydrostatic equations. Q. J. R. Meteorol. Soc. 140 (682), 1505–1520.