Decision-relevant characterisation of uncertainty in UK climate projections

Project CR20-3: Enabling the use and producing improved understanding of EuroCORDEX data over the UK

Richard Chandler*
Department of Statistical Science, UCL

Clair Barnes
Department of Statistical Science, UCL

Chris Brierley
Geography Department, UCL

July 2023

*Please direct all communications to r.chandler@ucl.ac.uk

Contents

1 Introduction **4**

2 Data sources and processing **6**

2.1 The UKCP18 regional ensemble . . . 6
2.2 The EuroCORDEX ensemble . . . 7
2.3 Estimates of observed climate . . . 7
2.4 Estimates of radiative forcing . . . 8
2.5 Preliminary data visualisation . . . 8

3 Combining climate projections: a brief review **9**

3.1 Overview of existing approaches . . . 10
3.2 Decision-relevant uncertainty assessments . . . 12
3.3 Shared discrepancies and ensemble structure . . . 14

4 Mimics for annual time series **15**

4.1 A mimic for the real climate system . . . 15
4.1.1 Parameter estimation . . . 18
4.1.2 Trend estimation and future projections . . . 20
4.2 Mimics for exchangeable ensembles . . . 22
4.2.1 The drift terms $(\beta_\omega(\cdot))$ and $\left(\tilde{\beta}_i(\cdot)\right)$ . . . 23
4.2.2 The trend terms $(\mu_\omega(\cdot))$ and $(\tilde{\mu}_i(\cdot))$ . . . 24
4.2.3 Simplifying the structure . . . 25
4.2.4 Identifiability . . . 26
4.2.5 State space formulation as a dynamic linear model (DLM) . . . 27
4.2.6 Connection with Sansom et al. (2021) . . . 28
4.3 Mimics for structured regional ensembles . . . 29
4.4 Full uncertainty quantification . . . 32
4.4.1 Sampling from $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$ . . . 33
4.4.2 Sampling from $\pi(\boldsymbol{S}|\boldsymbol{\theta},\boldsymbol{Y})$ . . . 34
4.4.3 Sampling from $\pi\left(\boldsymbol{Y}^{(\text{new})}|\boldsymbol{\theta},\boldsymbol{S}\right)$ . . . 34
4.4.4 Assessing and improving the posterior sampling accuracy . . . 35

5 Application to UK-averaged temperature and precipitation **38**

5.1 Annual temperature means and precipitation totals . . . 39
5.2 Seasonal results . . . 44

6 Summary and discussion **47**
6.1 Overall comments . . . 47
6.2 Implications for users of projections . . . 49
6.3 Unresolved issues and future work . . . 50

Acknowledgements **52**

Technical appendices

A1 Form of the EBM-inspired trend $\mu_0(t)$ **52**

A2 Choices of prior **53**
A2.1 Temperature in °C . . . 53
A2.2 Precipitation totals in cm . . . 55
A2.3 Logged precipitation totals . . . 56
A2.4 Initial values for the state vector . . . 57

A3 Initial values for numerical MAP estimation **58**
A3.1 The requirement for a negative definite Hessian matrix . . . 59

A4 Covariance matrix for constrained innovations in an exchangeable ensemble **59**

A5 Covariance matrix for constrained innovations in a structured regional ensemble **60**

Bibliography **63**

1 Introduction

The UK Climate Projections (UKCP) provide several products giving information about the future climate of the UK, including projections at global (60km), regional (12km) and local (2.2km) scales. While the regional projections are proving beneficial in understanding future climate risks (for example, Kennedy-Asser et al., 2021; Arnell et al., 2021), these were derived by perturbing the physical parameters of a single global climate model (GCM) and a single regional climate model (RCM), and so may not sample the full range of possible futures that is consistent with current scientific understanding. Project CR20-3 aims to complement the UKCP regional projections by combining them with information from the EuroCORDEX ensemble (Jacob et al., 2014) at the same resolution. This combined set will allow a better understanding of structural uncertainty in UK climate projections (Murphy et al., 2019; Arnell et al., 2021) and substantially enhances the information base available to support future climate change risk assessments.

A major contribution of the project has been to analyse and understand the EuroCORDEX outputs (both the GCM and the RCM components) over the UK region, in particular comparing them with the existing information from UKCP18. The results of this analysis are summarised in Barnes et al. (2022a): they include evaluations of global and regional projections with respect to historical observations, as well as descriptive summaries of future projections from the various sources. These analyses provide a comprehensive overview which is useful both as a summary of the information available, and to support the work of users who need qualitative climate information to provide context for their planning and decision-making.

A key message from the previous analyses is that at specified future time points the ranges of projections — and projected changes — can differ noticeably between ensembles, although the extent of these differences depends on the quantity being considered. Some of the differences can be resolved, to a greater or lesser degree, by considering changes over periods corresponding to fixed thresholds of global mean temperature change in the driving GCMs (Barnes et al., 2022b, Section 4): this 'warming level' approach has also been adopted in the IPCC's Sixth Assessment Report (Chen et al., 2021). In some respects however, such an approach addresses the symptom rather than the cause of differences between ensemble members; and the focus on warming levels is arguably unhelpful for users who typically need to plan for specified time frames e.g. relating to the design lifetimes of infrastructure projects or to the cost of adaptation measures over a specified planning horizon.

An alternative approach to resolving differences between ensembles is to postprocess their out-

puts, in such a way as to produce information that is deemed more relevant for users according to some criteria. Many approaches have been suggested for carrying out such postprocessing: Brunner et al. (2020) provide a recent review, and Section 3 provides more details. Nonetheless, the results of any such postprocessing exercise are inevitably subject to uncertainty; and few existing techniques attempt to provide information that is suitable for use with formal frameworks for decision-making under uncertainty (again, see Section 3). Indeed, decision-relevant characterisation of uncertainty and risk has been identified as a key knowledge gap for the UK Climate Resilience Programme under which the current project is funded.[1] This report describes a first step towards addressing this issue in the context of the UKCP18 and EuroCORDEX regional ensembles, noting that decision-relevant risk assessment requires consideration of multiple sources of uncertainty, and typically also requires an assessment of how, if at all, the ensemble simulations relate to reality (e.g. Stainforth et al. 2007).

The UKCP18 and EuroCORDEX ensembles each sample some but not all sources of uncertainty: in the work reported below, the omitted sources are handled by allowing for structured time-varying discrepancies between the ensemble members and the real climate system. The relationships between the ensemble simulations and reality are handled by extending the conceptual framework of Chandler (2013), which was developed explicitly to address many of the issues highlighted by Brunner et al. (2020). The extensions developed herein are twofold: the first is to move beyond a simplified conceptual framework to an operational procedure for postprocessing ensembles of time series, and the second is to handle structured ensembles in which the members are not necessarily exchangeable (see Section 4). The extension to time series uses ideas that are very similar to those proposed independently in the context of global temperature projections by Sansom et al. (2021), albeit differing in the precise details and in the computational implementation which, here, is designed to ensure that the methodology is suitable for routine application to large numbers of series.

The results reported below are potentially relevant for users who need more than a qualitative assessment of future UK climate, and who are comfortable with the use of probabilistic approaches to the assessment of risk as reviewed by, for example, Economou et al. (2016). Nonetheless, the results — particularly those relating to precipitation — should be regarded as a proof-of-concept pending more experience with the methodology. In particular, at the current stage of development the methodology only handles one quantity at a time (e.g. annual temperature, annual precipitation, summer temperature, winter precipitation and so forth) and does not attempt to

[1] See, for example, https://www.ukri.org/what-we-offer/browse-our-areas-of-investment-and-support/uk-climate-resilience-programme/ (accessed 28th November 2022).

characterise relationships between different quantities.

The next section summarises the datasets used in this report. Section 3 provides a brief review of approaches for postprocessing and combining information from ensembles of climate projections; while Section 4 describes the methodological development (this section can be skipped by readers who are more interested in the results). Section 5 reports the application of the methodology to UK-averaged temperature and precipitation, both annually and for each of the four seasons; while Section 6 concludes. Further temperature and precipitation analyses, for the main administrative regions of the UK, can be found online at https://github-pages.ucl.ac.uk/eurocordex-uk-plots/. An R package (Chandler, 2023) has also been developed, that can be used to apply the methodology to other series.

2 Data sources and processing

The analyses in this report focus on projections of annual and seasonal mean temperatures and precipitation totals to the year 2080, derived from the UKCP18 and EuroCORDEX regional ensembles under the RCP8.5 emissions scenario and averaged over the UK land surface. The ensembles are supplemented by historical estimates of the corresponding quantities from 1950 to 2021, and by estimates of global net effective radiative forcing for the corresponding time periods. Each of these datasets is now discussed in detail.

2.1 The UKCP18 regional ensemble

This is a twelve-member perturbed physics ensemble, produced using the HadREM3-GA7-05 regional model driven by the HadGEM3-GC3.05 global model under the RCP8.5 emissions scenario. Ensemble member 01 was run with the standard parameters for each model, and the remaining ensemble members were each run with slightly different perturbations to the model physics (Murphy et al., 2019). For each ensemble member, the same perturbations were used for the regional as for the global model.

The regional ensemble outputs used here are on a daily time scale and are available for the meteorological years 1980 through 2080 (i.e. from 1st December 1979 through 30th November 2080), at a spatial resolution of $12 \times 12 \text{km}^2$. UK-averaged annual and seasonal mean temperatures and precipitation totals have been derived from each ensemble member by averaging over all UK land surface grid cells and then aggregating the daily data over the relevant years or seasons.

2.2 The EuroCORDEX ensemble

The subset of the EuroCORDEX ensemble used here contains 64 members, each produced using one of ten different RCMs driven by one of ten different GCMs from Phase 5 of the Coupled Model Intercomparison Project (CMIP5) experiment (Taylor et al., 2012). All runs use the RCP8.5 emissions scenario. The regional models include the unperturbed version of HadREM3-GA7-05 that was used to produce UKCP18 regional ensemble member 01 as noted above (Tucker et al., 2021).

The EuroCORDEX simulations are run at a resolution of 0.11°, with the exact spatial extent varying according to the RCM used: all cover an area from approximately 27-72° N, and from 22° W-45° E (Jacob et al., 2014). The majority of GCM:RCM combinations contribute a single run to the ensemble; a few contribute three runs, and some are missing. Most of the 'missing' combinations have not been run, but four were excluded deliberately from the analysis: three because they are superseded in the ensemble by runs from later versions of the same RCMs driven by the same GCMs, and one because it contains inconsistent metadata regarding the driving GCM. More details are in Table 3 of Barnes et al. (2022a).

Like the UKCP18 ensemble, the EuroCORDEX outputs used here are on a daily time scale for meteorological years 1980 through 2080. Each RCM produces outputs on its own grid: annual and seasonal time series of UK-averaged temperature and precipitation have been obtained for each ensemble member by averaging over its own set of UK land surface grid cells.

2.3 Estimates of observed climate

Estimates of historical UK-averaged temperature and precipitation are obtained from the HadUK-Grid dataset, a collection of gridded climate variables over the UK land surface derived from the observation network (Perry et al., 2009). Precipitation and daily maximum and minimum temperatures are available at daily resolution from the end of the 19th century: daily mean temperature is calculated as the mean of the daily maximum and minimum.

In the work reported below, annual and seasonal UK-averaged mean temperature and precipitation have been calculated from this dataset for the meteorological years 1950 to 2021 inclusive. Data prior to 1950 have not been used, in part because they are less reliable due to a sparser station network. Other reasons for excluding these earlier data are (i) that they are relatively uninformative about responses to changes in radiative forcing (because forcings were relatively stable until the second half of the twentieth century — see Figure 2 below); and (ii) that their

inclusion increases the computational cost of the subsequent analyses.

2.4 Estimates of radiative forcing

To force the driving GCM runs in both the UKCP18 and EuroCORDEX ensembles, estimates of historical concentrations of greenhouse gases and other atmospheric forcing agents were used for the period 1980–2005 while, from 2006 onwards, concentrations were taken from the RCP8.5 emissions scenario (Taylor et al., 2012; Myhre et al., 2013). The combined effect of the various forcing agents can be summarised via an 'effective radiative forcing' (ERF) at a global scale (Smith et al., 2020): the annual ERF series used in the RCP8.5 emissions scenario is available from `https://github.com/IPCC-WG1/Chapter-7/tree/main/data_output/RCPs`, and is used in the analyses below to constrain the statistical representation of trends in the various projections.

2.5 Preliminary data visualisation

Figure 1 shows the annual time series of UK-averaged temperature and precipitation from both the UKCP18 and EuroCORDEX ensembles, together with the HadUK-Grid estimates (labelled as 'observations' in the figure). As far as temperature is concerned, both ensembles appear slightly too cool during the period for which HadUK-Grid estimates are available: the ensemble means generally lie below the observations. Subsequently however, the UKCP18 ensemble warms more rapidly than does EuroCORDEX and, by 2080, the mean temperatures across the members of the two ensembles differ by more than 1°C.

Next considering total annual precipitation, both ensembles tend to overestimate on average prior to 2020. Moreover, neither shows any obvious trend — in contrast with the HadUK-Grid estimates which suggest that average precipitation has slightly increased between 1950 and 2020. The ensembles' absence of trend continues beyond 2020: neither projects any obvious change over time, at least compared with the magnitude of interannual variation.

These points reflect features of the ensembles that are discussed in more depth in Barnes et al. (2022a,b). For present purposes, the key points are that the ensembles differ on average in the magnitude of projected future change, and that each ensemble's members tend to share discrepancies compared with the observations (e.g. the cool bias in temperatures, and the absence of trend in precipitation). A further feature, that will be exploited in the subsequent methodological development, is that the magnitude of interannual variability appears stable over time in both ensembles and for both variables. The same is true for each individual ensemble member, although

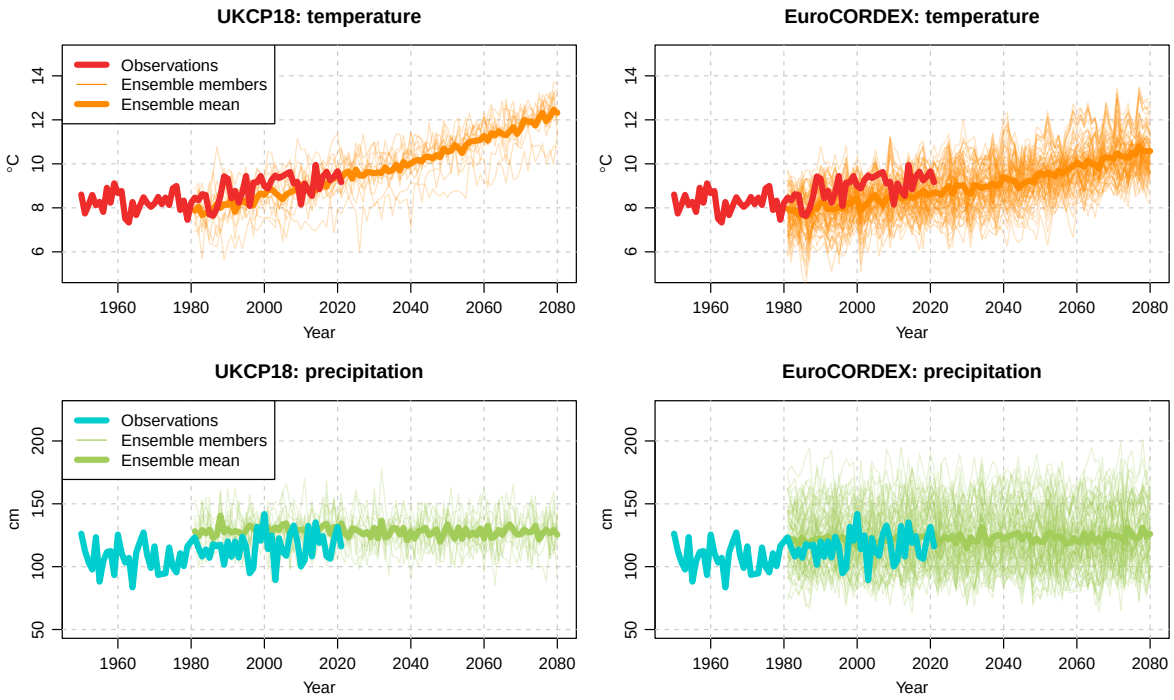

Figure 1: Annual (meteorological year) time series of UK-averaged mean temperature and total precipitation, 1950–2080: HadUK observations, and simulations from UKCP18 and EuroCORDEX ensembles.

this is less easy to see for the EuroCORDEX ensemble due to the large number of members.

For completeness, Figure 2 shows the annual time series of effective radiative forcings (henceforth ERFs) corresponding to the ensemble projections. Downward spikes correspond to volcanic eruptions. The forcings are plotted starting in 1900, to justify the earlier claim that they were relatively stable in the first half of the twentieth century.

3 Combining climate projections: a brief review

As noted in Section 1, many approaches are available for postprocessing ensembles of climate projections. The categorisation here is closely based on the review by Brunner et al. (2020). Climate models are referred to as 'simulators' throughout, to avoid confusion with other types of models.

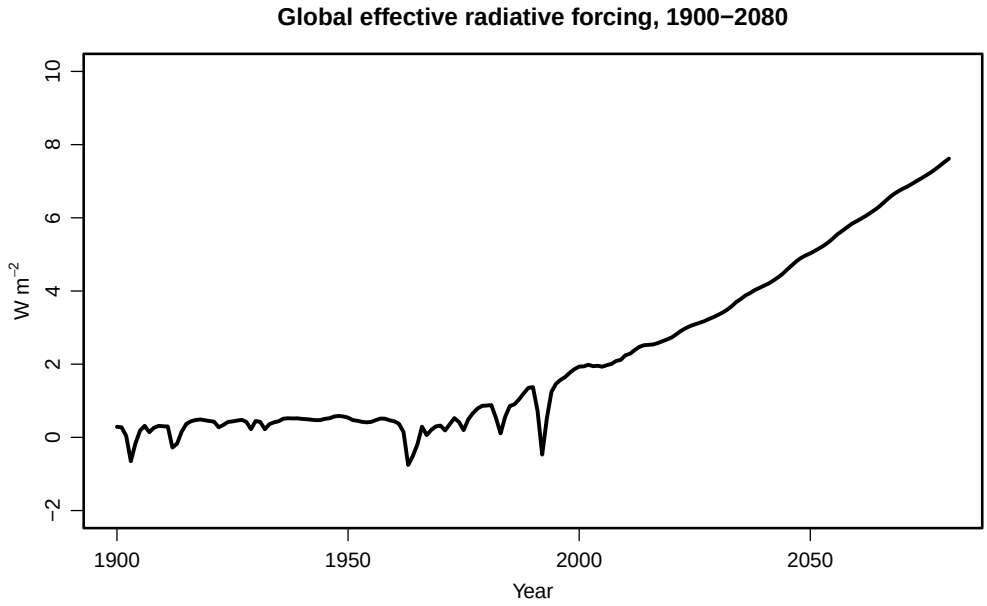

Figure 2: Annual time series of global effective radiative forcings underpinning the simulations from both the UKCP18 and EuroCORDEX ensembles. Values to 2000 are historical estimates; those from 2001 onwards correspond to the RCP8.5 scenario.

3.1 Overview of existing approaches

Perhaps the simplest and most widespread approach is to pool or average the ensemble members, optionally attaching weights to individual simulators or ensemble members based on criteria such as historical performance. This approach has long been criticised on both theoretical and empirical grounds; see, for example, Knutti (2010), Kjellström and Giorgi (2010) and Chandler (2013). The main problems are the potentially arbitrary choice of criteria for assigning weights; the absence of any rigorous treatment of uncertainty; and the fact that no single simulator is uniformly 'better' than other so that any kind of simple averaging is guaranteed to be sub-optimal.

A second type of approach is based on methodology developed by the "detection and attribution" community, which attempts to use an ensemble to identify space-time patterns in some climate variable of interest that are associated with changes in external forcings; and to use historical observations to determine how these patterns should be scaled using what is essentially a regression model (e.g. Allen et al. 2000; Stott and Kettleborough 2002; Kettleborough et al. 2007). The approach is commonly known as the "ASK" ("Allen, Stott and Kettleborough") methodology. The underlying assumption is that the ensemble provides information on patterns of change, but historical observations are needed to constrain the magnitudes of these patterns. The character-

isation of these methods by Brunner et al. (2020) suggests that they are fundamentally heuristic in nature, and notes that they rely on the ability to identify a forcing-related component of (observed) historical change.

Generalisations of the ASK approach are often described as being based on "observational constraints" or "emergent constraints". Hegerl et al. (2021) describe the idea as being to "rely on statistical relationships between present day, observable, climate properties and the magnitude of future change", while Shiogama et al. (2016) assert the underlying premise that "if a GCM overestimates the observed magnitude of historical climate change (not only the global mean change, but also spatiotemporal patterns), such a GCM will overestimate future climate changes by a proportional amount, and vice versa". The preliminary analysis in Section 2.5 suggests that this premise does not hold in the current context, for example because the historical biases in simulated temperatures are offset by biases in warming rates. Williamson and Sansom (2019) also note that the approach does not fully capture the uncertainty in climate projections, apart from anything else because the underpinning assumptions (which are rarely investigated in detail) are unrealistic: the weakest such assumption is that "there are no processes systematically missing from the models, but present in reality, that might cause us to view the behavior of the real world to be distinguishable from that of the models". Again, the analysis in Section 2.5 shows that this assumption does not hold: there are systematic differences between the properties of the observed and modelled series. Williamson and Sansom (2019) propose to modify the approach using an extended statistical model that allows for a shared discrepancy between each of the simulators and the real climate system: indeed, this is a key requirement of *any* method that seeks to provide credible uncertainty assessments, as discussed in more detail below.

A further difficulty with the emergent constraints methodology is that the observable properties used to develop the required statistical relationships will, in general, differ depending on the quantity of interest. For example, Williamson and Sansom (2019) examine the application of the methodology to estimate the earth's equilibrium climate sensitivity (the quantity of interest) based on statistical relationships with a metric of global temperature variability (the observable property) proposed by Cox et al. (2018): this metric will not necessarily be relevant when considering other quantities of interest. This dependence on the quantity of interest suggests that the approach may not be practical when bulk processing projections for many different quantities, as is required to produce region-specific information to supplement the UKCP18 projections.

The third category of approach considered by Brunner et al. (2020) is based on Bayesian methods — although their coverage of such methods is rather selective and omits many references that

are relevant in the current context, such as as Tebaldi et al. (2005); Greene et al. (2006); Furrer et al. (2007); Smith et al. (2009); Buser et al. (2009); Tebaldi and Sansó (2009); Buser et al. (2010); Kang et al. (2012) and Rougier et al. (2013). In its most general formulation, a Bayesian approach requires an explicit statistical characterisation of the structure of the ensemble outputs and their relation to the real climate system, from which posterior probability distributions of quantities of interest can be derived. The precise nature of the characterisation can vary widely, however: for example, the UKCP18 probabilistic projections are based on a statistical emulator that is designed to approximate the behaviour of a single simulator and is calibrated using a perturbed physics ensemble (Sexton et al., 2012), while Qasmi and Ribes (2022) use a direct statistical representation of the ensemble outputs themselves.[2]

It is also worth noting that there is nothing intrinsically Bayesian about developing an explicit statistical characterisation of ensemble structure: however, the resulting characterisations are typically too complex to be handled using conventional statistical methods and it is therefore convenient to take advantage of modern Bayesian computational machinery such as Markov Chain Monte Carlo (MCMC) methods (e.g. Robert and Casella 2011). A further advantage of Bayesian approaches is that they allow external information — for example, regarding physical constraints on quantities of interest — to be incorporated formally into an analysis and propagated into the resulting uncertainty assessments. Williamson and Sansom (2019) provide a clear example of this process.

The final category of methods considered by Brunner et al. (2020) is designed for single-model ensembles, often designed to explore the effects of uncertainty in initial conditions. This is not particularly relevant in the context of the EuroCORDEX and UKCP18 regional projections, and hence is not considered further here.

3.2 Decision-relevant uncertainty assessments

The approach adopted in the present work is motivated primarily by the need to produce defensible uncertainty assessments that can be incorporated into formal frameworks for decision-making

[2]It is perhaps worth clarifying that recent usage of the term 'emulator' in the climate science literature differs from its precise statistical meaning given, for example, by Goldstein and Rougier (2005). In statistical terms, an emulator is a fast and (relatively) simple statistical model that aims to approximate the workings of a complex computer simulator (as in Rougier and Sexton 2007 for example), whereas in the recent climate literature (e.g. Jackson et al. 2022; Nicholls et al. 2022) the term is often used to denote *any* fast simplified statistical or physical model. For the remainder of this report, the term will be used according to its precise statistical meaning: following Chandler (2013), we use the term 'mimic' to refer to a statistical model that merely describes the structure of an ensemble's outputs instead of trying to emulate the workings of the simulators themselves.

under uncertainty. Such frameworks require the specification of probability distributions for quantities of interest, which are used in essence to make bets on the eventual outcome — the decision-maker's goal then being to minimise their expected loss (or maximise their expected gain), defined as appropriate to the application. Economou et al. (2016) illustrate the main concepts, in the context of a simplified example relating to uncertainties in hazard warning systems.

Seen from this perspective, the aim of an uncertainty analysis must be to provide probability distributions for uncertain quantities of interest that are suitable for making bets. For example, in the context of weather forecasting one could assemble a large number of instances when the forecast probability of rain was 0.7: a necessary (but not sufficient) condition for the distributions to be considered decision-relevant is that rain subsequently occurred in roughly 70% of these instances. If this is not the case, the forecast probabilities clearly will not provide reliable estimates of expected losses, and hence will not be suitable for use in formal decision-making frameworks.

When evaluating the performance of probabilistic weather forecasting systems, as in the example above the forecasts can be evaluated by comparing with the outcomes that subsequently materialised. When considering ensembles of climate projections however, this is not possible: projections are typically made over decadal time periods, so it will be decades before such evaluations projections can be made. In this setting therefore, there are limited opportunities to demonstrate that uncertainty assessments are fit for purpose. A necessary condition for this is that they are based on defensible and transparent judgements; Section 6 discusses some other possibilities. From this perspective, it is critical to account for ensemble structure — in particular, for relationships among the ensemble members, and between the ensemble and the real climate system. This is not the case for many of the non-Bayesian approaches reviewed above, which perhaps focus more on the underlying physical processes. By contrast, the approach developed below has the ensemble structure at its core, with the physical processes informing the detailed implementation.

The importance of ensemble structure for uncertainty assessments may not be immediately obvious. Perhaps the simplest demonstration arises in introductory statistics courses: when testing for differences between the means of two groups of observations, if the data are paired then the analysis must acknowledge that each pair of observations shares a common component of variation — for example, by using a paired t-test instead of the usual two-group t-test — in order to obtain correct standard errors that are fundamental to the subsequent inference. In the same way, there are shared components of variation in any ensemble of climate simulator outputs, and these must be considered in any attempt to provide decision-relevant uncertainty assessments.

3.3 Shared discrepancies and ensemble structure

As hinted above, Bayesian approaches are natural candidates for use in situations requiring probabilistic uncertainty assessments accounting for the structure of the available data. To set the scene for the detailed development in Section 4, it is helpful to introduce the key challenges when analysing ensembles of projections from multiple climate simulators, and to outline a generic conceptual framework for addressing these challenges. The description here is based on Chandler (2013): several of the other Bayesian references above are similar in spirit.

The starting point is the observation that climate projections do not attempt to predict the detailed day-to-day sequence of weather: rather, they are intended to provide information on the statistical properties of the climate system at time scales of decades and upwards (e.g. von Storch and Zwiers 1999, Section 7.1). Leith and Chandler (2010) noted that 'statistical properties' of any system can be regarded as parameters in a statistical model, subsequently referred to as a *mimic* in the vocabulary of Chandler (2013); and, moreover, that ensembles of projections are produced by climate simulators that, in broad terms, represent the same dynamical processes so that the ensemble members should have a similar statistical structure. This in turn implies that when focusing on a specific quantity such as annual mean temperature for a particular region, the *form* of the mimics used to summarise the structure of each ensemble member should be similar; but that differences between members will be reflected predominantly in the values of their statistical model parameters — referred to as *descriptors*.

When analysing the outputs from any ensemble of M (say) projections, as in Section 2 the ensemble outputs typically span a time period that allows comparison with historical observations of the real climate system. The ensemble outputs can be represented as vectors $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_M$, each corresponding to one member; and the corresponding observations by $\boldsymbol{Y}_0$. The ensemble members will include information relating to both historical and future periods but, by definition, $\boldsymbol{Y}_0$ can only include historical information. The corresponding descriptors for the various information sources are denoted by vectors $\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$. Framed in this way, the goal of analysing the available information is to use the available data $\boldsymbol{Y}_0, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_M$ to learn about the real-climate descriptor $\boldsymbol{\theta}_0$ — in particular, about any components of $\boldsymbol{\theta}_0$ relating to future climate for which no observations are (yet) available.

As noted above while discussing approaches based on emergent constraints, a key feature of almost all ensembles is the potential presence of shared discrepancies between the properties of the simulators and those of the real climate. These arise for several reasons, including the omission of

processes from the simulators and the sharing of information between climate modelling groups; and their existence has long been acknowledged (Knutti et al., 2010) as well as being visible in some of the plots from Section 2.5. Chandler (2013) considers the situation where the descriptors $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M$ share a single common discrepancy $\boldsymbol{\omega}$ compared with the real climate system, represented in the statistical framework by specifying the conditional expectation $\mathbb{E}(\boldsymbol{\theta}_i|\boldsymbol{\theta}_0) = \boldsymbol{\theta}_0 + \boldsymbol{\omega}$ for $i = 1, \ldots, M$. In this case an explicit expression can be derived for the posterior distribution of $\boldsymbol{\theta}_0$ in a Bayesian framework, using other judgements about the deviations of each individual simulator descriptor from its consensus, and about the behaviour of the shared discrepancy itself.

In the present setting, there are two fundamental differences compared with the framework outlined in Chandler (2013). The first is that it is hard to capture the structure of the nonlinear trends in projections such as those shown in Section 2.5 over long time periods, using a low-dimensional descriptor vector $\boldsymbol{\theta}$ (both Leith and Chandler 2010 and Chandler 2013 considered fixed 30-year time slices instead of continuous series), while the second is the application to structured regional ensembles in which some members share the same GCM while others share the same RCM, necessitating an extension to the notion of a 'single common discrepancy' $\boldsymbol{\omega}$. Our approach to the first issue is to use a flexible state-space time series model formulation for the mimic in which — as in Sansom et al. (2021) — the underlying trend is considered as a realisation of a stochastic process and estimated using the Kalman Smoother (Chandler and Scott, 2011, Section 5.5). For the second, we rewrite the basic framework in a form that is better suited to the analysis of structured ensembles. The next section provides the technical details.

4 Mimics for annual time series

This section develops mimics that are appropriate for the analysis of ensembles of climate projections where the outputs of interest are time series at an annual resolution. These could be annual precipitation totals or temperature means as in Section 2.5; or, alternatively, time series of seasonal totals or means (e.g. the time series of winter precipitation totals). In the first instance, a mimic structure is developed for the real climate system; frameworks for the analysis of both simple and structured ensembles are described subsequently.

4.1 A mimic for the real climate system

Let $Y_0(t)$ denote the value of some quantity of interest — such as UK-averaged temperature — in the real climate system for year t. A simple representation of the structure of the series $(Y_0(t))$

is

$$Y_0(t) \;=\; \mu_0(t) + \varepsilon_0(t) \;, \tag{1}$$

where $\mu_0(\cdot)$ represents a smooth trend and $(\varepsilon_0(t))$ is a series of 'irregular' variations about that trend. In the first instance, it is convenient to specify that the $(\varepsilon_0(t))$ are mutually independent random variables, each with expectation and common variance σ_0^2. Inspection of Figure 1 suggests that the magnitude of year-on-year variation in observed temperature and precipitation does indeed seem fairly constant over the period for which observations are available, so that the constant variance assumption initially seems reasonable. The assumption of independence is harder to judge from the plots: this will be checked later.

To complete the mimic specification it is necessary to specify a form for the trend $\mu_0(\cdot)$ which is, at least in part, a response to greenhouse gas emissions and other forcings. It is therefore of interest to relate the trend to the ERF time series shown in Figure 2. It is, of course, unlikely that there will be a direct relationship between regional UK trends and this ERF series which, in addition to representing emissions at a global rather than regional scale, aggregates over multiple chemical species arising from both natural and anthropogenic sources. The aim of incorporating the ERF information is therefore not to provide a detailed representation of regional response to complex combinations of forcings: rather, it is to provide plausible constraints on the structure of trends in the regional climate response. To achieve this, we use simple energy balance models (EBMs) as a guide — in a similar spirit to that of Qasmi and Ribes (2022) who used EBMs to represent both local and global temperature responses to natural forcings in a related context (but using a fundamentally different methodology from that considered here).

The simplest EBM is a zero-dimensional representation of an aquaplanet experiencing small perturbations about an equilibrium level: this model is described (e.g. Bates 2007) via the linear first-order differential equation

$$c_0 \frac{dT}{dt} = -bT(t) + f(t) \;, \tag{2}$$

where $f(t)$ represents the forcing perturbation at time t and $T(t)$ is the corresponding perturbation about the equilibrium global temperature; c_0 and b are parameters representing the physical characteristics of the system. The solution of (2), for $t \geq 0$, is

$$T(t) = T(0)e^{-\kappa t} + c_0^{-1} \int_0^t f(t-u)e^{-\kappa u} du \;, \tag{3}$$

where the "characteristic decay rate" κ is equal to b/c_0. Bates (2007) also describes a more

sophisticated version of the model in which energy is transported between 'tropical' and 'polar' latitudinal zones: in this more complex model, the solution involves two exponential terms with different decay rates. Similarly, Geoffroy et al. (2013) study a model with two layers ('atmosphere' and 'ocean') and find that the time series of global surface temperature has a similar form involving two exponential terms (see their equations B8 and B9): the underlying differential equations for their layers are similar to (2). Analytical solutions to other EBM variants (e.g. Hansen et al. 1981, who consider a single-layer one-dimensional model in which the earth is divided into latitudinal bands) have not been published, but the models are all essentially systems of equations similar to (2) and have solutions that can be expressed as combinations of exponentially-weighted integrals of past and present forcings: these integrals represent the effects of thermal inertia within each compartment of the system.

In view of the similarities between EBM solutions for both global and 'regional' (e.g. zonal bands) temperature responses, expressions such as (2) and (3) may provide a useful guide to the relationship between ERF and regional temperature trends — and, possibly, also to trends in other variables such as precipitation, although the justification in this case is less clear. It would, of course, be naïve to suggest that these expressions could be used directly to model the temperatures or precipitation directly: apart from anything else, if the forcing series is smooth then the solution to a differential equation such as (2) will also be a smooth curve with no irregular year-on-year variation. The trend $\mu_0(\cdot)$ *will* be smooth in this situation, however: hence this particular issue is resolved by considering the EBM solution merely as providing guidance on the form of the trend rather than the detailed structure of the climate variable(s) of interest.

A further complication is that EBMs treat time as a continuous quantity whereas the present analysis focuses on discrete annual series. To address this, we replace derivatives and integrals with differences and sums respectively. The resulting analogue of (2) is

$$c_0 \left(\mu_0(t) - \mu_0(t-1)\right) = b\mu_0(t-1) + f(t)$$

where now $f(t)$ represents the forcing at time t. Rearranging and reparameterising, this can be written as

$$\mu_0(t) = \phi_0\mu_0(t-1) + \gamma_0 f(t) \ ,$$

where $\phi_0 = 1 - b/c_0$ is related to the inertia of the system and $\gamma_0 = 1/c_0$ can be interpreted as an instantaneous response to the forcing.

Finally, noting that the connection with EBMs is perhaps best regarded as a metaphor to find an

approximate representation of the trend, one might allow for systematic deviation, or 'drift' away from this relationship over time, by including an additional time-varying term:

$$\mu_0(t) = \phi_0\mu_0(t-1) + \gamma_0 f(t) + \beta_0(t-1) \ . \tag{4}$$

Several options are available to represent the drift process $(\beta_0(\cdot))$: the approach taken here is to represent it as a random walk

$$\beta_0(t) \ = \ \beta_0(t-1) + Z_0(t) \ , \tag{5}$$

where $(Z_0(t))$ is a sequence of independent zero-mean random variables with mean zero and variance τ_0^2. The rationale for this is that a random walk has the flexibility to move arbitrarily far from its starting point over long enough time intervals (Chandler and Scott, 2011, Section 5.2), so that its inclusion in (5) allows the mimic to adapt to whatever trend is indicated by the data if this is not consistent with the trajectory suggested by the EBM. On the other hand, if the random walk variance τ_0^2 is very small then the drift process $\beta_0(\cdot)$ will itself change very slowly over time: this will be appropriate if the EBM-inspired trend structure provides a good fit to the observations.

It can be shown (see Appendix A1) that under (4) and (5), the trend $\mu_0(t)$ can be written as

$$\mu_0(t) = \phi_0^t\mu_0(0) + \gamma_0 \sum_{u=0}^{t-1} \phi_0^u f(t-u) + \mathcal{J}_0(t) \ , \tag{6}$$

where $(\mathcal{J}_0(t))$ has a complex structure induced by the drift process $\beta_0(\cdot)$. The analogue with the continuous-time result (3) is clear if ϕ_0 is restricted to be strictly positive, in which case we can write $\phi_0 = \exp(-\kappa)$ so that (6) becomes $e^{-\kappa t}\mu_0(0) + \gamma_0 \sum_{u=0}^{t-1} e^{-\kappa u} f(t-u) + \mathcal{J}_0(t)$.

4.1.1 Parameter estimation

The mimic defined by equations (1), (4) and (5) contains three unknown parameters: σ_0^2, τ_0^2 and ϕ_0, which collectively form the descriptor vector $\boldsymbol{\theta}_0$ for the real climate series (note that γ does not appear in the descriptor vector, for reasons explained below). Given data $\{Y(1), Y(2), \ldots, Y(T)\}$ (which may include missing values — if, for example, T represents a time in the future for which projections are required), this descriptor vector can be estimated using a likelihood function constructed under the assumption that the (ε_t) and (Z_t) in (1) and (5) are normally distributed. The likelihood is constructed by writing the mimic as a dynamic linear model (DLM) in state space form

and using the Kalman Filter: the general procedure is described in Durbin and Koopman (2012, Chapter 7). The state space formulation involves the state vector $\boldsymbol{S}_0(t) = (\mu_0(t)\ \beta_0(t)\ \gamma_0)'$ so that $Y_0(t) = (1\ 0\ 0)\,\boldsymbol{S}_0(t) + \varepsilon_0(t)$ and

$$\boldsymbol{S}_0(t) = \begin{pmatrix} \phi_0 & 1 & f(t) \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \boldsymbol{S}_0(t-1) + \begin{pmatrix} 0 \\ Z_0(t) \\ 0 \end{pmatrix} . \tag{7}$$

Note that γ appears here as part of the state vector: this may seem unnatural, but is needed to exploit the flexibility offered by the state space formulation. The transition matrix contains the forcing series $f(t)$ and hence is time-varying. This is handled straightforwardly by most software implementations of the Kalman Filter and Smoother, including the `dlm` library in `R` (Petris, 2010) which is used to produce the results reported below.[3]

To estimate the parameters using the likelihood function, the most natural approach is to find the maximum likelihood estimators (MLEs). The maximisation must be done numerically. We maximise over the transformed descriptor vector $(\log \sigma_0^2\ \ \log \tau_0^2\ \ \log [\phi_0/(1-\phi_0)])'$ to ensure that the variance estimates are non-negative and that the estimate of ϕ_0 lies within its meaningful range of $(0, 1)$; this transformation also improves the accuracy of quadratic approximations to the log-likelihood that are used in the subsequent assessment of uncertainty (Section 4.4). The transformation does, however, cause potential convergence problems in situations where the data suggest that one or more variances are very small: such situations are not unusual. This issue is also reported by Petris (2010), and is caused by the log-likelihood being very flat for large negative values of the log variances where the variances themselves are all very close to zero.

To address this 'flat likelihood' problem, instead of maximising the log-likelihood itself we maximise a penalised version, where the penalty is a sum of quadratic functions of $\log \sigma_0^2$, $\log \tau_0^2$ and $\log [\phi_0/(1-\phi_0)]$. The penalty is chosen to be small enough to have minimal effect except in regions of the parameter space where the log-likelihood is flat: in these regions, it induces a small amount of curvature. This procedure can be formalised by noting that any such quadratic penalty can be regarded (up to additive constants that do not depend on the parameters of interest) as a sum of logarithms of Gaussian probability density functions (pdfs). The penalised log-likelihood function can therefore be regarded as the logarithm of a Bayesian posterior density, where the individual Gaussian pdfs are taken as independent prior distributions for the transformed descrip-

[3] At the time of writing, there is an unresolved memory leak in the underlying `C` code for one of the key routines in the latest version (1.1-6) of the `dlm` library. This has been brought to the attention of the package author; until it is resolved, a corrected version can be obtained from Richard Chandler at r.chandler@ucl.ac.uk.

tor elements (this follows from the usual 'posterior $\propto$ prior $\times$ likelihood' formula, which becomes 'log posterior = log prior + log-likelihood + constant' on taking logs). Maximisation of the log posterior is therefore equivalent to finding the posterior mode, and yields the *maximum a posteriori* (MAP) estimator of the descriptor vector — which also forms the basis of the Laplace approximation to a full Bayesian analysis (e.g. Kass and Raftery 1995). For the maximisation, we use the `nlm()` routine in R (R Core Team, 2022); the choice of prior parameters is summarised in Appendix A2, while the setting of initial values for the optimisation algorithm is discussed in Appendix A3.

After estimating the descriptor vector, a time series of standardised residuals can be calculated (see Chandler and Scott 2011, Section 5.5.4) and used to check the mimic structure. Figure 3 shows the time series and autocorrelation functions for standardised residuals from the mimic fitted to the observations shown in Figure 1. These provide checks: first, that the mimic structure captures the trend in the original series (no long-term trend is discernible in the residual time series); second, that the assumption of constant variance is reasonable; and finally that the irregular series $(\varepsilon_0(t))$ does indeed appear to be uncorrelated, as evidenced by the fact that at lags of a year and above the residual autocorrelations lie mostly within the 95% limits indicated by the dashed lines.

4.1.2 Trend estimation and future projections

Having estimated the descriptor vector for the mimic, the Kalman Smoother can be used to estimate the underlying trend $\{\mu_0(1), \mu_0(2), \ldots, \mu_0(T)\}$ and associated uncertainties. Prediction intervals can also be formed for the quantities of interest themselves (i.e. the $(Y_0(t))$), using equation (1) to infer that if $\hat{\mu}_0(t)$ denotes the Kalman smoothed estimate of $\mu_0(t)$ then

$$Y_0(t) = \mu_0(t) + \varepsilon_0(t) = \hat{\mu}_0(t) - (\hat{\mu}_0(t) - \mu_0(t)) + \varepsilon_0(t) \ .$$

It follows that the variance of $Y_0(t) - \hat{\mu}_0(t)$ is equal to $\mathrm{Var}\left[\hat{\mu}_0(t) - \mu_0(t)\right] + \mathrm{Var}\left[\varepsilon_0(t)\right]$. The first of these two variances is provided directly from the output of the Kalman Smoother, while the second is σ_0^2.

If the $(\varepsilon_0(t))$ can be considered as normally distributed, then a 95% prediction interval for $Y_0(t)$ is $\hat{\mu}_0(t) \pm 1.96\sqrt{\mathrm{Var}\left[Y_0(t) - \hat{\mu}_0(t)\right]}$ — this is often the case for annual mean temperatures, and may also hold approximately for annual and seasonal total precipitation, although the lower limit of the interval is not guaranteed to be positive. For this reason, to construct a prediction interval for a precipitation total $Y_0(t)$ we use the 2.5th and 97.5th percentiles of a gamma distribution

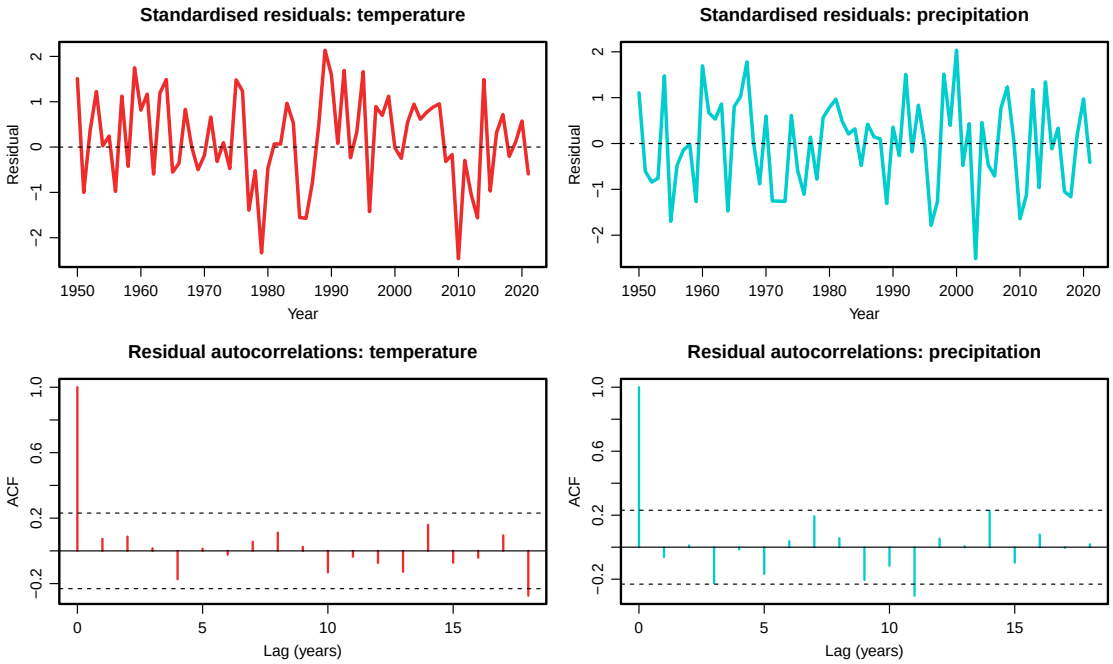

Figure 3: Analysis of standardised residuals from EBM-inspired mimic fitted to observations of UK-averaged annual mean temperature and total precipitation, from 1950 to 2020. Top: residual time series, bottom: residual autocorrelation functions. Dashed lines in the top plots are at zero; those in the bottom plots define approximate 95% limits under the assumption that the underlying autocorrelations are zero.

---

with expectation $\hat{\mu}_0(t)$ and variance $\mathrm{Var}\left[Y_0(t) - \hat{\mu}_0\left(t\right)\right]$.[4]

Figure 4 shows the projections and prediction intervals for UK-averaged annual temperature and precipitation, obtained from this mimic fitted to the observations. It is premature to comment on these projections in detail, because they involve a substantial amount of extrapolation based on data to 2020. Nonetheless, it is notable that the future projections are reasonably tightly constrained: this presumably reflects the fact that past changes are closely related to the forcing series shown in Figure 2. It is noteworthy that the estimated temperature trend dips briefly in the early 1960s and again in the early 1990s: these dips coincide with relatively cool observed temperatures, and are associated with volcanically-induced reductions in ERF.

Of course, this close relationship between forcings and trends may change in the future: no purely statistical analysis based on the observations alone could detect this. However, the ensembles

[4] An alternative strategy for non-negative variables such as precipitation would be to work on a log scale; however, the physical justification for this is unclear.

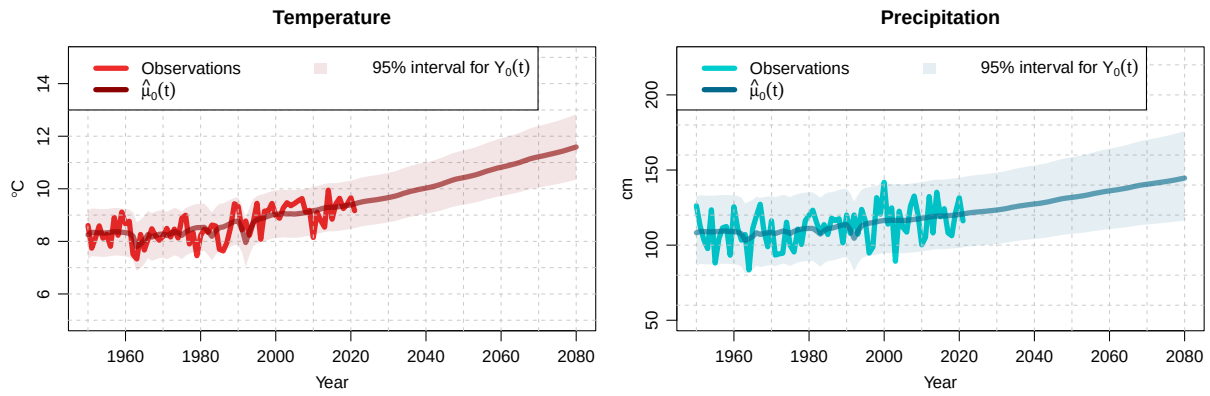

Figure 4: Projections of UK-averaged annual mean temperature and total precipitation, based on EBM-inspired mimic fitted to observations from 1950 to 2020.

do provide information about potential future trends: the next step is therefore to specify corresponding mimics for the ensemble members and to link their descriptors to that of the real climate system.

4.2 Mimics for exchangeable ensembles

The simplest ensemble structure arises when all members are exchangeable: formally this means that the statistical representation of the ensemble structure would not change if the labels of the individual members were permuted (e.g. Williamson and Sansom 2019). Exchangeability is a reasonable assumption for the UKCP18 regional ensemble since, as described in Section 2.1, all members are produced using variants of the same GCM-RCM pair.

For an exchangeable ensemble with M members, let $(Y_i(t))$ denote the output time series from the ith member. The EBM-inspired mimic defined by (1), (4) and (5) can be applied to each of the ensemble members, which have their own descriptor vectors, as well as their own drift processes and values of γ. Thus we have

$$Y_i(t) = \mu_i(t) + \varepsilon_i(t) \; ; \qquad \mathbb{E}\left(\varepsilon_i(t)\right) = 0, \mathrm{Var}\left[\varepsilon_i(t)\right] = \sigma_i^2 \; ; \tag{8}$$

$$\mu_i(t) = \phi_i \mu_i(t-1) + \gamma_i f(t) + \beta_i(t-1) \; ; \tag{9}$$

$$\text{and} \quad \beta_i(t) = \beta_i(t-1) + Z_i(t) \; ; \qquad \mathbb{E}\left(Z_i(t)\right) = 0, \mathrm{Var}\left[Z_i(t)\right] = \tau_i^2 \; , \tag{10}$$

where $(\varepsilon_i(\cdot))$ and $(Z_i(\cdot))$ are uncorrelated sequences.

The processes (μ_i) and $(\beta_i(\cdot))$ are not independent of each other, due to the potential for shared

discrepancies between the ensemble members and reality. The structure of these processes is entirely determined by the forcings, the process $(Z_i(t))$, the descriptor vector and the value of γ_i: shared discrepancies thus imply dependence among the $\{Z_i(t) : i = 1, \ldots, M\}$.

To account for the discrepancies, start by writing $\mu_i(t) = \mu_\omega(t) + \tilde{\mu}_i(t)$ and $\beta_i(t) = \beta_\omega(t) + \tilde{\beta}_i(t)$, say. The shared components $\mu_\omega(\cdot)$ and $\beta_\omega(\cdot)$ are referred to below as the 'consensus trend' and 'consensus drift' respectively. Here and throughout, the tilde notation $\tilde{\cdot}$ denotes quantities that relate to discrepancies, either between individual ensemble members and the consensus or between the consensus and reality: these discrepancies are mutually independent of each other, and of $(\mu_0(\cdot))$ and $(\beta_0(\cdot))$.

The next step is to use the mimic structure (8)–(10) to determine the dynamics of the components $(\mu_\omega(\cdot))$, $(\tilde{\mu}_i(\cdot))$, $(\beta_\omega(\cdot))$ and $\left(\tilde{\beta}_i(\cdot)\right)$; and to relate the consensus components to their real-world counterparts $(\mu_0(\cdot))$ and $(\beta_0(\cdot))$. It is easiest to start with the drift terms $(\beta_\omega(\cdot))$ and $\left(\tilde{\beta}_i(\cdot)\right)$, which contribute to the trend components but are otherwise not affected by them.

4.2.1 The drift terms $(\beta_\omega(\cdot))$ and $\left(\tilde{\beta}_i(\cdot)\right)$

Consider the drift increment $\nabla\beta_i(t) := \beta_i(t) - \beta_i(t-1)$ which, from (10), is equal to $Z_i(t)$. With $\beta_i(t) = \beta_\omega(t) + \tilde{\beta}_i(t)$, we also have $\nabla\beta_i(t) = \nabla\beta_\omega(t) + \nabla\tilde{\beta}_i(t)$; hence $\nabla\beta_\omega(t) + \nabla\tilde{\beta}_i(t) = Z_i(t)$. To ensure that $(Z_i(\cdot))$ is an uncorrelated sequence as in (10), $(\nabla\beta_\omega(\cdot))$ and $\left(\nabla\tilde{\beta}_i(\cdot)\right)$ are also specified as uncorrelated sequences so that $(\beta_\omega(\cdot))$ and $\left(\tilde{\beta}_i(\cdot)\right)$ are themselves random walks: $\beta_\omega(t) = \beta_\omega(t-1) + Z_\omega(t)$ and $\tilde{\beta}_i(t) = \tilde{\beta}_i(t-1) + \tilde{Z}_i(t)$ say, with $\mathbb{E}\left(Z_\omega(t)\right) = \mathbb{E}\left(\tilde{Z}_i(t)\right) = 0$, $\text{Var}\left[Z_\omega(t)\right] = \tau_\omega^2$ and $\text{Var}\left[\tilde{Z}_i(t)\right] = \tilde{\tau}_i^2$.

Next, consider the relationship between the consensus drift process $(\beta_\omega(\cdot))$ and its real-world counterpart $(\beta_0(\cdot))$. It is tempting to write $\beta_\omega(t) = \beta_0(t) + \tilde{\beta}_\omega(t)$, so that $\beta_i(t) = \beta_0(t) + \tilde{\beta}_\omega(t) + \tilde{\beta}_i(t)$ and $Z_i(t) = \nabla\beta_i(t) = \nabla\beta_0(t) + \nabla\tilde{\beta}_\omega(t) + \nabla\tilde{\beta}_i(t) = Z_0(t) + \tilde{Z}_\omega(t) + \tilde{Z}_i(t)$; $\left(\tilde{Z}_\omega(\cdot)\right)$ here is an uncorrelated process with $\mathbb{E}\left(\tilde{Z}_\omega(t)\right) = 0$ and $\text{Var}\left[\tilde{Z}_\omega(t)\right] = \tilde{\tau}_\omega^2$. Considering the variances of both sides of this expression, it follows that $\text{Var}\left(Z_i(t)\right) := \tau_i^2 = \tau_0^2 + \tilde{\tau}_\omega^2 + \tilde{\tau}_i^2$, which cannot be less than $\tau_0^2 = \text{Var}\left[Z_0(t)\right]$. Under this specification therefore, no ensemble member can drift away from its EBM approximation more slowly than does the real climate. It is hard to justify such a constraint: to remove it, the drift consensus is redefined as

$$\beta_\omega(t) = \alpha\beta_0(t) + \tilde{\beta}_\omega(t) \tag{11}$$

for some $\alpha \in \mathbb{R}$. In this case, the increment variance for the ith ensemble member is $\tau_i^2 = \alpha^2\tau_0^2 +$

$\tilde{\tau}_\omega^2 + \tilde{\tau}_i^2$, which can be less than τ_0^2 if $|\alpha| < 1$. Moreover, $Z_i(t)$ is partitioned as $\alpha Z_0(t) + \tilde{Z}_\omega(t) + \tilde{Z}_i(t)$: the common component $\tilde{Z}_\omega(t)$ induces the required dependence among the ensemble members.

The implications for the mimic specification (8)–(10) are: first, that $\beta_i(t-1)$ must be replaced with $\beta_0(t-1) + \alpha\tilde{\beta}_\omega(t-1) + \tilde{\beta}_i(t-1)$ in (9); and second, that (10) must be replaced with

$$\tilde{\beta}_\omega(t) = \tilde{\beta}_\omega(t-1) + \tilde{Z}_\omega(t) \; ; \qquad \mathbb{E}\left(\tilde{Z}_\omega(t)\right) = 0, \mathsf{Var}\left[\tilde{Z}_\omega(t)\right] = \tilde{\tau}_\omega^2 \; , \tag{12}$$

$$\text{and} \quad \tilde{\beta}_i(t) = \tilde{\beta}_i(t-1) + \tilde{Z}_i(t) \; ; \qquad \mathbb{E}\left(\tilde{Z}_i(t)\right) = 0, \mathsf{Var}\left[\tilde{Z}_i(t)\right] = \tilde{\tau}_i^2 \; . \tag{13}$$

With this structure, information from the ensemble feeds back into the post-processed projections primarily via the drift terms $\left\{\tilde{\beta}_i(\cdot) : i = 1, \ldots, M\right\}$, although the interdependence between all elements of the state vector ensures that this information also influences other elements to some degree. Informally, if the ensemble projections provide evidence that the trends $\{\mu_i(\cdot)\}$ may start to move away from the EBM approximation in the future, then this will be attributed to the drift processes $\{\beta_i(\cdot) = \alpha\beta_0(\cdot) + \tilde{\beta}_\omega(\cdot) + \tilde{\beta}_i(\cdot)\}$. If for the sake of argument, $\tilde{\tau}_\omega^2$ and $\tilde{\tau}_i^2$ were both zero then $\tilde{\beta}_\omega(\cdot)$ and $\tilde{\beta}_i(\cdot)$ would both be time-invariant: any future change in the drift would be attributed to the process $\beta_0(\cdot)$ therefore, thus influencing the future projections of the real-world trend $\mu_0(\cdot)$. Conversely, if τ_0^2 is very small — suggesting that the real-world trend follows the EBM approximation very closely — then the ensemble will contribute little additional information.

4.2.2 The trend terms $(\mu_\omega(\cdot))$ and $(\tilde{\mu}_i(\cdot))$

We now examine the implications of the above results for the trend dynamics in (9). As noted already, the evolution of $(\mu_i(\cdot))$ is completely determined by other quantities: without loss of generality therefore, and for consistency with the development above, we define $\tilde{\mu}_\omega(t) := \mu_\omega(t) - \alpha\mu_0(t)$ so that $\mu_\omega(t) = \alpha\mu_0(t) + \tilde{\mu}_\omega(t)$ and $\mu_i(t) = \alpha\mu_0(t) + \tilde{\mu}_\omega(t) + \tilde{\mu}_i(t)$. We also write $\gamma_i = \alpha\gamma_0 + \tilde{\gamma}_\omega + \tilde{\gamma}_i$ and $\phi_i = \phi_\omega + \tilde{\phi}_i$ (note that there is no need to relate ϕ_i directly to ϕ_0 because, as an element of the descriptor vector, it is not stochastic in nature).

With this notation, (9) becomes

$$\begin{aligned}\alpha\mu_0(t) + \tilde{\mu}_\omega(t) + \tilde{\mu}_i(t) = {} & \left(\phi_\omega + \tilde{\phi}_i\right)\left[\alpha\mu_0(t-1) + \tilde{\mu}_\omega(t-1) + \tilde{\mu}_i(t-1)\right] \\ & + \left(\alpha\gamma_0 + \tilde{\gamma}_\omega + \tilde{\gamma}_i\right) f(t) + \alpha\beta_0(t-1) + \tilde{\beta}_\omega(t-1) + \tilde{\beta}_i(t-1) \; .\end{aligned}$$

Clearly, any term on the right-hand side of this expression that varies between ensemble members

(i.e. with an 'i' subscript) is associated with the dynamics of $\tilde{\mu}_i(t)$; the remaining terms are most naturally associated with the consensus trend $\mu_\omega(t) = \alpha\mu_0(t) + \tilde{\mu}_\omega(t)$. We can therefore write

$$\tilde{\mu}_i(t) = \tilde{\phi}_i \left[\alpha\mu_0(t-1) + \tilde{\mu}_\omega(t-1)\right] + \left(\phi_\omega + \tilde{\phi}_i\right) \tilde{\mu}_i(t-1) + \tilde{\gamma}_i f(t) + \tilde{\beta}_i(t-1) \tag{14}$$

for $1 \leq i \leq M$; while the consensus can be written as

$$\begin{aligned}
\alpha\mu_0(t) + \tilde{\mu}_\omega(t) &= \phi_\omega \left[\alpha\mu_0(t-1) + \tilde{\mu}_\omega(t-1)\right] \\
&\qquad + \left(\alpha\gamma_0 + \tilde{\gamma}_\omega\right) f(t) + \alpha\beta_0(t-1) + \tilde{\beta}_\omega(t-1) \\
&= \alpha \left[\phi_\omega\mu_0(t-1) + \gamma_0 f(t) + \beta_0(t-1)\right] \\
&\qquad + \phi_\omega\tilde{\mu}_\omega(t-1) + \tilde{\gamma}_\omega f(t) + \tilde{\beta}_\omega(t-1) \\
&= \alpha \left[\phi_0\mu_0(t-1) + \gamma_0 f(t) + \beta_0(t-1)\right] + \alpha\left(\phi_\omega - \phi_0\right)\mu_0(t-1) \\
&\qquad + \phi_\omega\tilde{\mu}_\omega(t-1) + \tilde{\gamma}_\omega f(t) + \tilde{\beta}_\omega(t-1) \\
&= \alpha\mu_0(t) + \alpha\left(\phi_\omega - \phi_0\right)\mu_0(t-1) + \phi_\omega\tilde{\mu}_\omega(t-1) + \tilde{\gamma}_\omega f(t) + \tilde{\beta}_\omega(t-1) \ ,
\end{aligned}$$

the last step using the definition of $\mu_0(t)$ at (4). It follows immediately that

$$\tilde{\mu}_\omega(t) = \alpha\left(\phi_\omega - \phi_0\right)\mu_0(t-1) + \phi_\omega\tilde{\mu}_\omega(t-1) + \tilde{\gamma}_\omega f(t) + \tilde{\beta}_\omega(t-1) \ . \tag{15}$$

4.2.3 Simplifying the structure

The representation above involves a large number of descriptors: σ_0^2, τ_0^2, ϕ_0, α, $\tilde{\tau}_\omega^2$, ϕ_ω and $\left\{\left(\sigma_i^2, \tilde{\tau}_i^2, \tilde{\phi}_i\right) : i = 1, \ldots, M\right\}$. Numerical maximisation of a log-likelihood or log-posterior with respect to such a large descriptor set is likely to be difficult. The analysis can be simplified considerably if the ensemble members are considered to have common expected interannual variability σ_1^2, common drift increment discrepancy variance $\tilde{\tau}_1^2$ and common inertia coefficient ϕ_1. This reduces the total number of parameters to eight, and corresponds to a judgement that the ensemble members are strictly exchangeable.

A side-effect of setting common inertia coefficients $\{\phi_i = \phi_1\}$ is that there are no member-specific discrepancies: $\phi_i = \phi_1 = \phi_\omega$, and $\tilde{\phi}_i = 0$ for $i = 1, \ldots, M$. In this case, on the right-hand side of (14) the first term vanishes to yield

$$\tilde{\mu}_i(t) = \phi_1\tilde{\mu}_i(t-1) + \tilde{\gamma}_i f(t) + \tilde{\beta}_i(t-1) \ . \tag{16}$$

Note that there is little gain in terms of computational stability from restricting the $\{\gamma_i\}$ in this way, because they can be estimated straightforwardly using the Kalman Smoother. However, if $\phi_1 \neq \phi_0$ then the ensemble forcing coefficients $\{\gamma_i : i = 1, \ldots, M\}$ no longer have the same interpretation as the observation-derived coefficient γ_0 because they represent the effects of different weighted averages of current and previous forcings (see equation (6)). These forcing coefficients should therefore be interpreted with care, if at all.

4.2.4 Identifiability

A difficulty with the representation above is that the combined data structure is represented via $(M + 2)$ underlying trends $\mu_0(\cdot)$, $\tilde{\mu}_\omega(\cdot)$ and $\tilde{\mu}_1(\cdot)$ to $\tilde{\mu}_M(\cdot)$, whereas the available data contain just $M + 1$ series. The trends are not uniquely defined, therefore. For example, the trend for the ith ensemble member is $\mu_0(\cdot) + \tilde{\mu}_\omega(\cdot) + \tilde{\mu}_i(\cdot)$: this trend is unaffected by replacing $\tilde{\mu}_\omega(\cdot)$ and $\tilde{\mu}_i(\cdot)$ with $\tilde{\mu}_\omega(\cdot) + K$ and $\tilde{\mu}_i(\cdot) - K$ respectively, for any constant K.

This non-identifiability can be resolved by imposing an appropriate constraint on the trends: we choose the constraint $\sum_{i=1}^{M} \tilde{\mu}_i(t) = 0$ so that $\tilde{\mu}_M(t) = -\sum_{i=1}^{M-1} \tilde{\mu}_i(t)$ for all t. This 'sum-to-zero' constraint ensures that the consensus trend $\mu_\omega(\cdot) = \mu_0(\cdot) + \tilde{\mu}_\omega(\cdot)$ has an interpretation as an 'ensemble average': similar approaches are adopted routinely in analyses of variance e.g. Davison (2003, Section 9.2.2), and in the construction of generalised additive models (Wood, 2006, Chapter 5).

From equation (16), a sum-to-zero constraint on the $\{\tilde{\mu}_i(\cdot) : i = 1, \ldots, M\}$ implies corresponding constraints on the $\{\tilde{\gamma}_i\}$ and on the drift discrepancy processes $\left\{\tilde{\beta}_i(\cdot) : i = 1, \ldots, M\right\}$: if $\sum_{i=1}^{M} \tilde{\mu}_i(t) = 0$ and $\sum_{i=1}^{M} \tilde{\mu}_i(t-1) = 0$ then we must also have $\sum_{i=1}^{M} \tilde{\gamma}_i = 0$ and $\sum_{i=1}^{M} \tilde{\beta}_i(t-1) = 0$. The first of these conditions is easily handled by setting $\tilde{\gamma}_M = -\sum_{i=1}^{M-1} \tilde{\gamma}_i$. The second needs more care.

From (13), the constraint $\sum_{i=1}^{M} \tilde{\beta}_i(t-1) = 0$ implies in turn that $\sum_{i=1}^{M} \tilde{Z}_i(t) = 0$. A naïve way to achieve this is to set $\tilde{Z}_M(t) = -\sum_{i=1}^{M-1} \tilde{Z}_i(t)$: under this specification however, we have $\text{Var}\left(\tilde{Z}_M(t)\right) = \sum_{i=1}^{M-1} \tilde{\tau}_i^2$, which is equal to $(M-1)\tau_1^2$ under the assumption of exchangeability between the first $M-1$ ensemble members. But this implies that the Mth ensemble member is *not* exchangeable with the others — which, since the labelling is arbitrary, is unsatisfactory.

A preferable alternative is to condition probabilistically on the constraint. Specifically, let $\tilde{Z}_\cdot(t) = \sum_{i=1}^{M} \tilde{Z}_i(t)$ and consider the vector

$$\tilde{\boldsymbol{Z}}^*(t) = \left(\tilde{Z}_1(t)\ \tilde{Z}_2(t)\ \ldots\ \tilde{Z}_M(t)\ \tilde{Z}_\cdot(t)\right)' , \tag{17}$$

which, under the assumption of a common drift discrepancy variance $\tilde{\tau}_1^2$ for all ensemble members, has mean vector $\mathbf{0}$ and covariance matrix

$$\text{Var}\left(\tilde{\boldsymbol{Z}}^*(t)\right) = \begin{pmatrix} \tilde{\tau}_1^2 & 0 & \cdots & 0 & \tilde{\tau}_1^2 \\ 0 & \tilde{\tau}_1^2 & \cdots & 0 & \tilde{\tau}_1^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \tilde{\tau}_1^2 & \tilde{\tau}_1^2 \\ \tilde{\tau}_1^2 & \tilde{\tau}_1^2 & \cdots & \tilde{\tau}_1^2 & M\tilde{\tau}_1^2 \end{pmatrix} . \tag{18}$$

If we now define $\tilde{\boldsymbol{Z}}(t) = \left(\tilde{Z}_1(t)\ \tilde{Z}_2(t)\ \ldots\ \tilde{Z}_M(t)\right)'$ and condition on $\tilde{Z}_\cdot(t) = 0$, then a little algebra (see Appendix A4) shows that $\mathbb{E}\left(\tilde{\boldsymbol{Z}}(t)|\tilde{Z}_\cdot(t) = 0\right) = \mathbf{0}$ and that

$$\text{Var}\left(\tilde{\boldsymbol{Z}}(t)|\tilde{Z}_\cdot(t) = 0\right) = \tilde{\tau}_1^2 M^{-1} \begin{pmatrix} M-1 & -1 & \cdots & -1 \\ -1 & M-1 & \cdots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \cdots & M-1 \end{pmatrix} , \tag{19}$$

so that exchangeability is maintained between all ensemble members.

4.2.5 State space formulation as a dynamic linear model (DLM)

As with the observation-only mimic of Section 4.1, the combined mimics for all of the ensemble members and historical observations can be written as a DLM in state space form, so that the Kalman Filter can be used to construct a log-likelihood for the descriptors. This in turn enables the descriptors to be estimated, whence the Kalman Smoother can be used to estimate the trends and produce prediction intervals for future observations.

There are several ways to represent the combined mimics in state space form. In all cases, the observation vector at time t is $\boldsymbol{Y}(t) = (Y_0(t)\ \ Y_1(t)\ \ \ldots\ \ Y_M(t))'$; and, in the implementation considered here, the state vector is defined as

$$\boldsymbol{S}(t) = \left(\mu_0(t)\ \ \beta_0(t)\ \ \gamma_0\ \ \tilde{\mu}_\omega(t)\ \ \tilde{\beta}_\omega(t)\ \ \tilde{\gamma}_\omega\ \ \tilde{\mu}_1(t)\ \ \tilde{\beta}_1(t)\ \ \tilde{\gamma}_1\ \ \ldots\ \ \tilde{\mu}_{M-1}(t)\ \ \tilde{\beta}_{M-1}(t)\ \ \tilde{\gamma}_{M-1}\right)' .$$

From (1) and (8) with $\mu_i(t) = \alpha\mu_0(t) + \tilde{\mu}_\omega t + \tilde{\mu}_i(t)$ and $\tilde{\mu}_M(t) = -\sum_{i=1}^{M-1} \tilde{\mu}_i(t)$, we have

$$\boldsymbol{Y}(t) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \alpha & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \alpha & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \alpha & 0 & 0 & 1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 & \cdots & -1 & 0 & 0 \end{pmatrix} \boldsymbol{S}(t) + \begin{pmatrix} \varepsilon_0(t) \\ \varepsilon_1(t) \\ \varepsilon_2(t) \\ \vdots \\ \varepsilon_M(t) \end{pmatrix}.$$

Under the simplified representation of Section 4.2.3, the state vectors evolve as

$$\boldsymbol{S}(t) = \begin{pmatrix} \phi_0 & 1 & f(t) & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \alpha(\phi_1 - \phi_0) & 0 & 0 & \phi_1 & 1 & f(t) & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \phi_1 & 1 & f(t) & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & \phi_1 & 1 & f(t) \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{pmatrix} \boldsymbol{S}(t-1) + \begin{pmatrix} 0 \\ Z_0(t) \\ 0 \\ 0 \\ \tilde{Z}_\omega(t) \\ 0 \\ 0 \\ \tilde{Z}_1(t) \\ 0 \\ \vdots \\ 0 \\ \tilde{Z}_{M-1}(t) \\ 0 \end{pmatrix}, \tag{20}$$

where the covariance matrix of the final term is constructed from the elements of (19).

4.2.6 Connection with Sansom et al. (2021)

As noted in the Introduction, the framework introduced above is similar in many respects to that proposed by Sansom et al. (2021), who analysed projections of global mean surface temperature from the CMIP5 ensemble (Taylor et al., 2012) under the RCP4.5 emissions scenario. The closest points of similarity are the use of a state space formulation to represent the time evolution of both historical observations and ensemble outputs, the use of energy balance models with a random-walk drift term to represent the time evolution of the system, and the specification of shared discrepancies between the ensemble members and the real climate system.

There are some key differences between the approaches, however. The main one is that the EBM used by Sansom et al. (2021) is substantially more complex than equation (2): it has three

compartments instead of one, and includes a representation of net top-of-atmosphere radiation flux that is coupled with the global temperature series, instead of working directly with the underlying trend as in the present work. Moreover, the random-walk 'drift' is also applied to the radiation flux term rather than to the overall trend as in the development above: this can perhaps be interpreted as adjusting for the approximate nature of the EBM equations within the dynamics itself, whereas the drift terms in (4) and (9) are applied instead to the EBM outputs. The more complex EBM formulation is likely to provide a better representation of system dynamics at a global scale, but the increased complexity is perhaps less easily justified in the current context when considering local and regional quantities for which *any* EBM can provide at best a rough approximation of anticipated behaviour.

An EBM of increased complexity will also increase the computational cost of the analysis. Another feature adding — probably more substantially — to the computational cost of of Sansom et al. (2021) is the use of a full MCMC approach to estimation and inference, in place of the MAP estimation used here. This has the advantage of providing a full assessment of uncertainty in a single step, but can be very time-consuming and hence unsuitable for bulk processing of many sets of projections. In the current approach, the full treatment of uncertainty is achieved using importance sampling after the mimics have been fitted; this is discussed in Section 4.4.

There are other differences between the present approach and that of Sansom et al. (2021) — for example, they do not impose an explicit sum-to-zero constraint on their drift terms, instead relying on the zero-mean increments to ensure identifiability, analogously to the use of random-effects structures in analysis of variance — but these are relatively minor compared with the points above. Nonetheless, the approaches are sufficiently different that it would be interesting in future work to apply them both to the same sets of projections and compare their results and performance.

4.3 Mimics for structured regional ensembles

The development so far has focused on exchangeable ensembles, but is not applicable to regional ensembles where some members are run using the same GCM and others using the same RCM. For such ensembles, the structure of the observations is unchanged and continues to be described using equations (1), (4) and (5). The mimics and descriptors for the ensemble members must be modified however, to account for the ensemble structure.

Consider an ensemble involving R RCMs and G GCMs, potentially with multiple replicates of each RCM:GCM combination (e.g. with different initial conditions); and denote by $Y_{irg}(t)$ the

simulated quantity of interest at time t from the ith replicate of the (r, g) RCM:GCM combination. By direct analogy with (8)–(10), the ensemble structure is represented as

$$\begin{aligned} Y_{irg}(t) &= \mu_{rg}(t) + \varepsilon_{irg}(t) \; ; \qquad \mathbb{E}\left[\varepsilon_{irg}(t)\right] = 0, \mathsf{Var}\left[\varepsilon_{irg}\left(t\right)\right] = \sigma_{rg}^2 \; ; & (21) \\ \mu_{rg}(t) &= \phi_{rg}\mu_{rg}(t-1) + \gamma_{rg} f(t) + \beta_{rg}(t-1) & (22) \\ \text{and } \beta_{rg}(t) &= \beta_{rg}(t-1) + Z_{rg}(t) \quad ; \quad \mathbb{E}\left[Z_{rg}\left(t\right)\right], \mathsf{Var}\left[Z_{rg}\left(t\right)\right] = \tau_{rg}^2 \; . & (23) \end{aligned}$$

To account for the ensemble structure, shared discrepancies and components of variation are now separated into RCM and GCM effects, as

$$\begin{aligned} \mu_{rg}(t) &= \mu_{\omega}(t) + \tilde{\mu}_{r\cdot}(t) + \tilde{\mu}_{\cdot g}(t) + \tilde{\mu}_{rg}(t) \; ; & (24) \\ \beta_{rg}(t) &= \beta_{\omega}(t) + \tilde{\beta}_{r\cdot}(t) + \tilde{\beta}_{\cdot g}(t) + \tilde{\beta}_{rg}(t) \; ; & (25) \\ \gamma_{rg} &= \gamma_{\omega} + \tilde{\gamma}_{r\cdot} + \tilde{\gamma}_{\cdot g} + \tilde{\gamma}_{rg} \; ; & (26) \\ \beta_{\omega}(t) &= \alpha\beta_0(t) + \tilde{\beta}_{\omega}(t) \; ; \qquad \text{and} \qquad \gamma_{\omega} = \alpha\gamma_0 + \tilde{\gamma}_{\omega} \; . & (27) \end{aligned}$$

This structure represents the variation among the ensemble members in terms of an overall consensus trend $\mu_{\omega}(\cdot)$, contributions from the individual RCMs and GCMs, and finally 'interactions' representing any additional effect of specific RCM:GCM combinations. The consensus trend can further be decomposed as $\mu_{\omega}(t) = \alpha\mu_0(t) + \tilde{\mu}_{\omega}(t)$ as before: this is needed for the state-space formulation of the structure, but not for the following mathematical development.

Following the same steps as previously, (24)–(27) can be substituted into (22). The left-hand side becomes $\mu_{\omega}(t) + \tilde{\mu}_{r\cdot}(t) + \tilde{\mu}_{\cdot g}(t) + \tilde{\mu}_{rg}(t)$, and the right-hand side is

$$\begin{aligned} &\left(\phi_{\omega} + \tilde{\phi}_{r\cdot} + \tilde{\phi}_{\cdot g} + \tilde{\phi}_{rg}\right)\left[\mu_{\omega}(t) + \tilde{\mu}_{r\cdot}(t) + \tilde{\mu}_{\cdot g}(t) + \tilde{\mu}_{rg}(t)\right] \\ &\quad + \left(\gamma_0 + \tilde{\gamma}_{\omega} + \tilde{\gamma}_{r\cdot} + \tilde{\gamma}_{\cdot g} + \tilde{\gamma}_{rg}\right) f(t) + \alpha\beta_0(t) + \tilde{\beta}_{\omega}(t) + \tilde{\beta}_{r\cdot}(t) + \tilde{\beta}_{\cdot g}(t) + \tilde{\beta}_{rg}(t) \; , \end{aligned}$$

where ϕ_{rg} has also been partitioned as $\phi_{rg} = \phi_{\omega} + \tilde{\phi}_{r\cdot} + \tilde{\phi}_{\cdot g} + \tilde{\phi}_{rg}$. Equating terms with regard

to their dependence on r and g, we now find that

$$\begin{aligned}
\mu_\omega(t) &= \phi_\omega \mu_\omega(t-1) + (\alpha\gamma_0 + \tilde{\gamma}_\omega) f(t) + \alpha\beta_0(t-1) + \tilde{\beta}_\omega(t-1) \; ; && (28)\\
\tilde{\mu}_{r\cdot}(t) &= \left(\phi_\omega + \tilde{\phi}_{r\cdot}\right) \tilde{\mu}_{r\cdot}(t-1) + \tilde{\phi}_{r\cdot}\mu_\omega(t-1) + \tilde{\gamma}_{r\cdot} f(t) + \tilde{\beta}_{r\cdot}(t-1) \; ; && (29)\\
\tilde{\mu}_{\cdot g}(t) &= \left(\phi_\omega + \tilde{\phi}_{\cdot g}\right) \tilde{\mu}_{\cdot g}(t-1) + \tilde{\phi}_{\cdot g}\mu_\omega(t-1) + \tilde{\gamma}_{\cdot g} f(t) + \tilde{\beta}_{\cdot g}(t-1) \text{ ; and} && (30)\\
\tilde{\mu}_{rg}(t) &= \left(\phi_\omega + \tilde{\phi}_{r\cdot} + \tilde{\phi}_{\cdot g} + \tilde{\phi}_{rg}\right) \tilde{\mu}_{rg}(t-1) + \tilde{\phi}_{rg}\left[\mu_\omega(t-1) + \tilde{\mu}_{r\cdot}(t-1) + \mu_{\cdot g}(t-1)\right] \\
&\quad + \tilde{\phi}_{r\cdot}\tilde{\mu}_{\cdot g}(t-1) + \tilde{\phi}_{\cdot g}\tilde{\mu}_{r\cdot}(t-1) + \tilde{\gamma}_{rg} f(t) + \tilde{\beta}_{rg}(t-1) \; . && (31)
\end{aligned}$$

As before, the ERF coefficients $\gamma_0, \tilde{\gamma}_\omega, \ldots, \tilde{\gamma}_{rg}$ are time-constant while the approximation drift processes $\beta_0(\cdot)$, $\tilde{\beta}_\omega(\cdot), \ldots, \tilde{\beta}_{rg}(\cdot)$ are random walks.

Once again, it is convenient to reduce the number of descriptors so as to simplify the problem. Noting that the most of the key discrepancy structure is captured by the trends, forcing coefficients and drift terms; that the precise trend structure for the individual ensemble members is not of direct interest; and that numerical optimisation over large numbers of descriptors is likely to be slow and potentially unstable, we proceed as though the $\left\{\sigma^2_{rg}\right\}$, $\left\{\tau^2_{rg}\right\}$ and $\{\phi_{rg}\}$ are common to all ensemble members. In this case, the descriptors are σ^2_0, ϕ_0, τ^2_0, σ^2_1, $\phi_1 = \phi_\omega$, α, $\tilde{\tau}^2_\omega$ and $\tilde{\tau}^2_1$ as in the exchangeable case; and (28)–(31) simplify to

$$\begin{aligned}
\mu_\omega(t) &= \phi_1 \mu_\omega(t-1) + (\gamma_0 + \tilde{\gamma}_\omega) f(t) + \alpha\beta_0(t-1) + \tilde{\beta}_\omega(t-1) \; ; \\
\tilde{\mu}_{r\cdot}(t) &= \phi_1 \tilde{\mu}_{r\cdot}(t-1) + \tilde{\gamma}_{r\cdot} f(t) + \tilde{\beta}_{r\cdot}(t-1) \; ; \\
\tilde{\mu}_{\cdot g}(t) &= \phi_1 \tilde{\mu}_{\cdot g}(t-1) + \tilde{\gamma}_{\cdot g} f(t) + \tilde{\beta}_{\cdot g}(t-1) \text{ ; and} \\
\tilde{\mu}_{rg}(t) &= \phi_1 \tilde{\mu}_{rg}(t-1) + \tilde{\gamma}_{rg} f(t) + \tilde{\beta}_{rg}(t-1) \; .
\end{aligned}$$

The last three equations here are very similar to the corresponding equation (16) for the exchangeable case; and, writing $\mu_\omega(t) = \alpha\mu_0(t) + \tilde{\mu}_\omega(t)$, the corresponding representation (15) of the shared trend discrepancy $\tilde{\mu}_\omega(\cdot)$ carries over exactly. This similarity arises from the partitioning of each underlying trend into components of discrepancy that can be considered independently under the simplified structure — although they are combined in a structured way via equation (24) to induce dependencies among the ensemble time series.

With this simplified structure, one can also contemplate omitting the final 'interaction' terms from (24) to (26): this will considerably reduce the size of the state vector and hence potentially speed up computations.

The specification above uses $1 + R + G + RG$ trends in total, to describe the structure of

at most RG combinations of RCMs and GCMs. $R + G + 1$ constraints must be imposed, therefore, to ensure that the mimic is identifiable: again, these are chosen to ensure that the various quantities in the mimic have a natural interpretation. Specifically, we use sum-to zero constraints $\sum_{r=1}^{R} \tilde{\mu}_{r\cdot}(t) = 0$, $\sum_{g=1}^{G} \tilde{\mu}_{\cdot g}(t) = 0$, $\left\{\sum_{g=1}^{G} \tilde{\mu}_{rg}(t) = 0 : r = 1, \ldots, R\right\}$ and $\left\{\sum_{r=1}^{R} \tilde{\mu}_{rg}(t) = 0 : g = 1, \ldots, G-1\right\}$ (these constraints also imply that $\sum_{r=1}^{R} \tilde{\mu}_{rG}(t) = 0$, since $\sum_{r=1}^{R} \tilde{\mu}_{rG}(t) = \sum_{r=1}^{R} \left(\sum_{g=1}^{G} \tilde{\mu}_{rg}(t) - \sum_{g=1}^{G-1} \tilde{\mu}_{rg}(t)\right) = \sum_{r=1}^{R} \sum_{g=1}^{G} \tilde{\mu}_{rg}(t) - \sum_{g=1}^{G-1} \sum_{r=1}^{R} \tilde{\mu}_{rg}(t) = 0 - 0$). The ensemble structure can then be written as a DLM in state space form via the state vector

$$\boldsymbol{S}(t) = \left(\boldsymbol{S}_0'(t)\ \tilde{\boldsymbol{S}}_{\omega}'(t)\ \tilde{\boldsymbol{S}}_{1\cdot}'(t)\ \ldots\ \tilde{\boldsymbol{S}}_{(R-1)\cdot}'(t)\ \tilde{\boldsymbol{S}}_{\cdot 1}'(t)\ \ldots\ \tilde{\boldsymbol{S}}_{\cdot(G-1)}'(t)\ \tilde{\boldsymbol{S}}_{11}'(t)\ \ldots\ \tilde{\boldsymbol{S}}_{(R-1)(G-1)}'(t)\right)',$$

where $\boldsymbol{S}_0(t) = (\mu_0(t)\ \ \beta_0(t)\ \ \gamma_0)'$, $\boldsymbol{S}_{\omega}(t) = \left(\tilde{\mu}_{\omega}(t)\ \ \tilde{\beta}_{\omega}(t)\ \ \tilde{\gamma}_{\omega}\right)'$ and $\boldsymbol{S}_{rg}(t) = \left(\tilde{\mu}_{rg}(t)\ \ \tilde{\beta}_{rg}(t)\ \ \tilde{\gamma}_{rg}\right)'$ for $1 \leq r \leq R-1$, $1 \leq g \leq G-1$. The detailed representation is mostly straightforward and is omitted for reasons of space, with the exception of the covariance matrix of the constrained innovations $\left\{\tilde{Z}_{rg}(t) : r = 1, \ldots, R-1; g = 1, \ldots, G-1\right\}$ which is derived in Appendix A5.

4.4 Full uncertainty quantification

For all of the mimics developed above, the state-space formulation enables estimation of the descriptors using MAP estimation; and the resulting fits can be used in conjunction with the Kalman Smoother to produce trend estimates and prediction intervals for the real-world quantities of interest as in Figure 4. This procedure does not account for uncertainty in the descriptor estimates, however: the prediction intervals in Figure 4 are therefore likely to be too narrow. The easiest way to characterise the full uncertainty distribution, including the effect of descriptor uncertainty, is to draw samples from it. We now discuss how to achieve this.

For a given mimic, denote the descriptor vector by $\boldsymbol{\theta}$; the complete collection of state vectors over the time period of interest by $\boldsymbol{S}$; the collection of observations and ensemble outputs used for descriptor estimation by $\boldsymbol{Y}$; and any 'observable but not yet observed' quantities by $\boldsymbol{Y}^{(\text{new})}$. In the current context, $\boldsymbol{Y}^{(\text{new})}$ will often consist of the vector of real-world quantities of interest $(Y_0(T+1), Y_0(T+2), \ldots, Y_0(T+\ell))$ where T is the final year for which observations are available (2020 for the HadUK data considered in Section 2.3) and $T + \ell$ is the final year of ensemble outputs. Other possibilities are considered in Section 6. The following algorithm produces a realisation from the joint distribution of $\boldsymbol{Y}^{(\text{new})}$ under an assumed mimic structure, conditional on the data $\boldsymbol{Y}$:

1. Sample a value of $\boldsymbol{\theta}$ from the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$. Here and throughout, the notation $\pi(\cdot|\cdot)$ denotes a generic joint conditional pdf.

2. Using the value of $\boldsymbol{\theta}$ sampled in step 1, sample a value of $\boldsymbol{S}$ from the distribution $\pi(\boldsymbol{S}|\boldsymbol{Y}, \boldsymbol{\theta})$.

3. Using the values of $\boldsymbol{S}$ from step 2 and $\boldsymbol{\theta}$ from step 1, sample a value of $\boldsymbol{Y}^{(\text{new})}$ from the distribution $\pi\left(\boldsymbol{Y}^{(\text{new})}|\boldsymbol{S}, \boldsymbol{\theta}\right)$. Note that under the mimic structures considered here, $\boldsymbol{Y}^{(\text{new})}$ is conditionally independent of the observations $\boldsymbol{Y}$ given $\boldsymbol{S}$ and $\boldsymbol{\theta}$.

Repeating these steps a large number of times will produce the equivalent number of sampled values of $\boldsymbol{Y}^{(\text{new})}$, conditioned on $\boldsymbol{Y}$. Any feature of the joint pdf $\pi\left(\boldsymbol{Y}^{(\text{new})}|\boldsymbol{Y}\right)$ can then be estimated, to a degree of precision limited only by the number of samples.

Each of these steps is now discussed in more detail.

4.4.1 Sampling from $\pi(\theta|Y)$

Much of the literature on DLMs focuses on the development of MCMC schemes for sampling from the posterior distribution of parameters and state vectors conditional on the data: see, for example, Carter and Kohn (1994); Frühwirth-Schnatter (1994); Durbin and Koopman (2002); Reis et al. (2006). As discussed in Section 4.2.6 however, the computational expense of these schemes is undesirable in applications where it is required to carry out routine analyses of many time series of ensemble outputs. As an alternative therefore, we proceed on the basis (established by experimentation with many different sets of projections) that the log-posterior distributions are invariably unimodal in the region of 'reasonable' descriptor values. Moreover, standard theory suggests that in large samples the log-posterior should be approximately quadratic in the region of parameter space supported by the data (Cox, 2006, Section 6.2). In the current work, the use of transformed parameters such as the logarithms of variances, together with the use of Gaussian priors (see Section 4.1.1), is designed in part with this quadratic approximation in mind: for example, the parameter transformations avoid the pronounced asymmetric log-likelihood profiles that can otherwise occur if the maximum is close to a boundary of the parameter space.

If the log posterior is approximately quadratic in the neighbourhood of the MAP estimator $\hat{\boldsymbol{\theta}}$, it follows that the posterior itself is approximately multivariate normal, with mean equal to the maximum a posterior estimate and covariance matrix ($\boldsymbol{\Omega}$, say) equal to the negative Hessian matrix of the log posterior at its maximum. This is the Laplace approximation to the posterior, mentioned in Section 4.1.1. The Hessian is computed during the estimation procedure, as a by-product of the numerical maximisation: approximate sampling from $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$ can therefore be

achieved by sampling from the corresponding multivariate normal distribution. The accuracy of this multivariate normal approximation is considered further in Section 4.4.4.

4.4.2 Sampling from $\pi(\boldsymbol{S}|\boldsymbol{\theta}, \boldsymbol{Y})$

Given a parameter vector $\boldsymbol{\theta}$ and observations $\boldsymbol{Y}$, the problem of simulating from the conditional distribution of the state vectors has received considerable attention in the literature. Algorithms include that of Durbin and Koopman (2002), as well as the "forward-filtering-backward-sampling" (FFBS) approach of Frühwirth-Schnatter (1994); Carter and Kohn (1994) that forms the basis for the `dlmBSample()` routine in the dlm library (Petris, 2010), used here.

One potential difficulty with the FFBS approach used here is that it is based on Gaussian specifications for all of the stochastic components of the DLM. As noted in Section 4.1.1, this Gaussian specification could be problematic when applied to non-negative quantities such as precipitation: in the context of sampling from $\pi(\boldsymbol{S}|\boldsymbol{\theta}, \boldsymbol{Y})$ based on the mimics discussed above, the sampled values of the trend $\mu_0(t)$ are not guaranteed to be positive. Experience suggests that negative values are encountered rather rarely when postprocessing precipitation projections from the UKCP18 and EuroCORDEX ensembles. As a pragmatic solution when working with precipitation projections therefore, samples with any negative value of μ_0 are discarded and the sampling is repeated until all values of $\mu_0(t)$ are non-negative. Again as noted in Section 4.1.1, an alternative would be to carry out the analysis on a log scale for precipitation and similar quantities.

Another difficulty, encountered occasionally, is that the FFBS algorithm occasionally fails, either during the Kalman filtering or during the subsequent backward sampling. This is associated with failure of the underlying `LAPACK` routines used in the matrix calculations, and cannot easily be fixed without rewriting either the `LAPACK` routines or the `dlm` library that calls them. We have examined the situations leading to such failures, which do not appear to be systematic (e.g. they are not associated with unusual sampled parameter values). In view of this, and given that it happens rather infrequently (typically up to a handful of times in a thousand samples), they have been dealt with simply by drawing replacement samples.

4.4.3 Sampling from $\pi\left(\boldsymbol{Y}^{(\mathbf{new})}|\boldsymbol{\theta}, \boldsymbol{S}\right)$

The final sampling step is straightforward: for all mimics under consideration, we have $\mathbb{E}\left[Y_*(t)|\boldsymbol{S}\right] = \mu_*(t)$ and $\mathrm{Var}\left[Y_*(t)|\boldsymbol{S}\right] = \sigma_*^2$ independently for all series and time points, where the asterisk "$*$" denotes an arbitrary subscript as appropriate to the mimic and quantity being considered. Since $\boldsymbol{Y}^{(\mathrm{new})}$ includes no elements that have already been observed, its elements can be sampled merely

by drawing the corresponding values independently from the appropriate distributions — which will usually be normal distributions with the specified mean and variance, but could also be gamma distributions for non-negative quantities as discussed in Section 4.1.2.

4.4.4 Assessing and improving the posterior sampling accuracy

In Section 4.4.1 it was suggested that a multivariate normal distribution can be used to draw approximate samples from the posterior $\pi\left(\boldsymbol{\theta}|\boldsymbol{Y}\right)$. The accuracy of this procedure depends on that of the quadratic approximation to the log posterior in the relevant neighbourhood of its mode. Of course, for any value of $\boldsymbol{\theta}$ the log posterior itself can be evaluated up to an additive constant since $\log \pi\left(\boldsymbol{\theta}|\boldsymbol{Y}\right) = \log h(\boldsymbol{\theta};\boldsymbol{Y}) + \log K$ say, where $\log h(\boldsymbol{\theta};\boldsymbol{Y})$ is the function that is maximised to obtain the MAP estimate (thus $\pi\left(\boldsymbol{\theta}|\boldsymbol{Y}\right) = Kh(\boldsymbol{\theta};\boldsymbol{Y})$, so that K is the normalising constant required to ensure that $\int \pi\left(\boldsymbol{\theta}|\boldsymbol{Y}\right) d\boldsymbol{\theta} = 1$). For any value $\boldsymbol{\theta}$, the accuracy of the approximation can therefore be assessed by comparing $\log h\left(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}\right) - \log h\left(\boldsymbol{\theta}|\boldsymbol{Y}\right)$ with $\log g\left(\hat{\boldsymbol{\theta}}\right) - \log g(\boldsymbol{\theta})$ in an appropriate neighbourhood of $\hat{\boldsymbol{\theta}}$, where $g(\cdot)$ is the joint density of the approximating multivariate normal distribution.

To determine an appropriate neighbourhood of $\hat{\boldsymbol{\theta}}$ for this purpose, note that if p denotes the dimension of $\boldsymbol{\theta}$ and if $\boldsymbol{\Omega}$ denotes the covariance matrix of the approximating multivariate normal density $g(\cdot)$ as above, then

$$g(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{p/2}\left|\boldsymbol{\Omega}\right|^{1/2}} \exp\left[-\frac{1}{2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)\right] \qquad (\boldsymbol{\theta} \in \mathbb{R}^p)$$

so that $\log g\left(\hat{\boldsymbol{\theta}}\right) - \log g(\boldsymbol{\theta}) = \frac{1}{2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)'\boldsymbol{\Omega}^{-1}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)$. Defining $\boldsymbol{Z} := \boldsymbol{\Omega}^{-1/2}\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}\right)$ where $\boldsymbol{\Omega}^{-1/2}$ is a matrix square root of $\boldsymbol{\Omega}^{-1}$ (typically computed via its eigendecomposition), we therefore have $2\left[\log g\left(\hat{\boldsymbol{\theta}}\right) - \log g(\boldsymbol{\theta})\right] = \boldsymbol{Z}'\boldsymbol{Z} = \sum_{i=1}^{p} Z_i^2$ where Z_i is the ith element of $\boldsymbol{Z}$. But with this definition of $\boldsymbol{Z}$, the $\{Z_i\}$ are independent standard normal random variables so that $\sum_{i=1}^{p} Z_i^2 \sim \chi_p^2$. Under the multivariate normal density $g(\cdot)$ therefore, values of $\boldsymbol{\theta}$ such that $2\left[\log g\left(\hat{\boldsymbol{\theta}}\right) - \log g(\boldsymbol{\theta})\right]$ falls in the extreme upper tail of this χ_p^2 distribution can be deemed inconsistent with the data: there is no strong requirement for the Gaussian approximation to hold for such values of $\boldsymbol{\theta}$. A 'relevant' region of $\boldsymbol{\theta}$ values can thus be identified using a high but non-extreme quantile of this χ_p^2 distribution. This leads to the following procedure, to check the accuracy of the Gaussian approximation for operational purposes:

1. Choose a high but non-extreme quantile of the χ_p^2 distribution, for example its 99th percentile. Call this quantile Q.

2. Define a set of 2^p vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{2^p}$, in which each element is set to either $\sqrt{Q/p}$ or $-\sqrt{Q/p}$ in such a way that every combination of values appears in the set.

3. For each $i \in \{1, \ldots, 2^p\}$, calculate $\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}} + \boldsymbol{\Omega}^{1/2}\boldsymbol{z}_i$, and the corresponding log-posterior ratio $\log h\left(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}\right) - \log h\left(\boldsymbol{\theta}_i|\boldsymbol{Y}\right)$.

As calculated in step 3 above, the $\{\boldsymbol{\theta}_i : i = 1, \ldots, 2^p\}$ lie on the principal axes of the ellipsoid defined by the equation $2\left[\log g\left(\hat{\boldsymbol{\theta}}\right) - \log g(\boldsymbol{\theta})\right] = Q$ — which can be considered as an approximate boundary for the region of $\boldsymbol{\theta}$ values that are 'reasonably consistent' with the data. If the Gaussian approximation to the posterior $\pi\left(\boldsymbol{\theta}|\boldsymbol{Y}\right)$ holds within this region therefore, the log ratios $\left\{\log h\left(\hat{\boldsymbol{\theta}}|\boldsymbol{Y}\right) - \log h\left(\boldsymbol{\theta}_i|\boldsymbol{Y}\right)\right\}$ will all be close to $Q/2$. If this condition does not hold, a relatively straightforward remedy is to sample randomly from the approximating Gaussian distribution and then to reweight the samples. This is the idea behind *importance sampling*, as suggested for use in this kind of application by Durbin and Koopman (2012, Chapter 13).

For present purposes, importance sampling is most easily described by considering the posterior probability $P\left(\boldsymbol{\theta} \in \mathcal{A}|\boldsymbol{Y}\right)$ where $\mathcal{A}$ is some region of the parameter space. Noting that this probability can be regarded as the expected value of $I\left(\boldsymbol{\theta} \in \mathcal{A}\right)$ where $I(\cdot)$ is the indicator function taking the value 1 if its argument is true and 0 otherwise, we have

$$\begin{aligned} P\left(\boldsymbol{\theta} \in \mathcal{A}|\boldsymbol{Y}\right) = \mathbb{E}\left[I\left(\boldsymbol{\theta} \in \mathcal{A}\right)|\boldsymbol{Y}\right] &= \int_{\boldsymbol{\theta}} I\left(\boldsymbol{\theta} \in \mathcal{A}\right)\pi\left(\boldsymbol{\theta}|\boldsymbol{Y}\right)d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\theta}} I\left(\boldsymbol{\theta} \in \mathcal{A}\right)\frac{\pi\left(\boldsymbol{\theta}|\boldsymbol{Y}\right)}{g\left(\boldsymbol{\theta}\right)}g\left(\boldsymbol{\theta}\right)d\boldsymbol{\theta} \\ &= \mathbb{E}_{\boldsymbol{\theta}\sim g}\left[I\left(\boldsymbol{\theta} \in \mathcal{A}\right)\frac{\pi\left(\boldsymbol{\theta}|\boldsymbol{Y}\right)}{g\left(\boldsymbol{\theta}\right)}\right], \end{aligned}$$

where $\boldsymbol{E}_{\boldsymbol{\theta}\sim g}\left[\cdot\right]$ denotes expectation with respect to the random vector $\boldsymbol{\theta}$, evaluated over the distribution with joint pdf $g(\cdot)$. The expectation can be estimated by drawing a large number of samples $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_N$ from this distribution (note that these are not the same as the values used in the procedure to check the accuracy of the Gaussian approximation above), and computing the sample average $N^{-1}\sum_{i=1}^{N} I\left(\boldsymbol{\theta}_i \in \mathcal{A}\right)\pi\left(\boldsymbol{\theta}_i|\boldsymbol{Y}\right)/g\left(\boldsymbol{\theta}_i\right)$. Recalling that $\pi\left(\boldsymbol{\theta}|\boldsymbol{Y}\right) = Kh\left(\boldsymbol{\theta};\boldsymbol{Y}\right)$ as defined at the start of this section, this sample average can be written as

$$KN^{-1}\sum_{i=1}^{N} I\left(\boldsymbol{\theta}_i \in \mathcal{A}\right)h\left(\boldsymbol{\theta}_i;\boldsymbol{Y}\right)/g\left(\boldsymbol{\theta}_i\right) = KN^{-1}\sum_{i=1}^{N} I\left(\boldsymbol{\theta}_i \in \mathcal{A}\right)w_i \text{ , say,} \tag{32}$$

where $w_i = h\left(\boldsymbol{\theta}_i;\boldsymbol{Y}\right)/g\left(\boldsymbol{\theta}_i\right)$.

It remains to find the value of K, determined by the requirement $\int_{\boldsymbol{\theta}} \pi\left(\boldsymbol{\theta}|\boldsymbol{Y}\right) d\boldsymbol{\theta} = 1$. This implies that

$$K \int_{\boldsymbol{\theta}} h\left(\boldsymbol{\theta}|\boldsymbol{Y}\right) d\boldsymbol{\theta} = 1 \quad \Rightarrow \quad K \int_{\boldsymbol{\theta}} \frac{h\left(\boldsymbol{\theta};\boldsymbol{Y}\right)}{g\left(\boldsymbol{\theta}\right)} g\left(\boldsymbol{\theta}\right) d\boldsymbol{\theta} = 1 \quad \Rightarrow \quad K\mathbb{E}_{\boldsymbol{\theta}\sim g}\left[\frac{h\left(\boldsymbol{\theta};\boldsymbol{Y}\right)}{g\left(\boldsymbol{\theta}\right)}\right] = 1$$

and $K = 1/\mathbb{E}_{\boldsymbol{\theta}\sim g}\left[h\left(\boldsymbol{\theta};\boldsymbol{Y}|\right)/g\left(\boldsymbol{\theta}\right)\right]$. Thus the samples $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$ can be used to estimate K as $N/\sum_{i=1}^{N} w_i$ and, in turn, $P\left(\boldsymbol{\theta} \in \mathcal{A}|\boldsymbol{Y}\right)$ can be estimated as $\left[\sum_{i=1}^{N} w_i I\left(\boldsymbol{\theta}_i \in \mathcal{A}\right)\right] / \left[\sum_{i=1}^{N} w_i\right]$. Each sampled value of $\boldsymbol{\theta}$ is thus reweighted to account for the fact that it was sampled from the distribution with density $g(\cdot)$ instead of the exact posterior; and the sampled $\boldsymbol{\theta}$ values can be used to estimate any desired summaries or functions of the posterior distribution by propagating this reweighting through the subsequent calculations.

Although importance sampling provides a quick and easy way to improve the Gaussian approximation to the posterior distribution, its use requires that the distribution of the importance sampling weights $\{w_i\}$ has a finite variance. Monahan (2001, Chapter 12) gives a thorough discussion and some suggested checks — the simplest of which is to examine the distribution of the weights for extreme outliers. Such situations are most likely to arise when the posterior distribution is heavy-tailed by comparison with its Gaussian approximation so that the ratio $h\left(\boldsymbol{\theta}_i;\boldsymbol{Y}\right)/g\left(\boldsymbol{\theta}_i\right)$ can become arbitrarily large within the region of interest. In this case, a simple solution is to sample from a heavier-tailed distribution such as a multivariate t distribution and, when calculating the importance weights $\{w_i\}$, to use the pdf of this heavier-tailed distribution in place of the multivariate normal pdf $g(\cdot)$.

A further consideration when using importance sampling is that the estimates are subject to sampling uncertainty: if the original Gaussian approximation to the log posterior is already accurate therefore, a very large number N of samples may be needed in order for the improvement to be distinguishable from sampling variation. This comes with its own computational overhead of course, since the log posterior must be computed N times. To reduce the cost, it is therefore worth adopting sampling schemes that are designed to reduce the number of samples required to characterise specified features of the posterior distribution with a given level of accuracy. Durbin and Koopman (2012, Chapter 11) recommend the use of antithetic variables for this purpose, to ensure that the simulated samples have both the correct mean vector and dispersion. The present work has also explored the possibility of using antithetic sampling to ensure the correct correlation structure.

Finally, a point of computational efficiency is that having obtained a sample from $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$, the

first step of the FFBS algorithm for sampling from $\pi(\boldsymbol{S}|\boldsymbol{\theta}, \boldsymbol{Y})$ is to apply the Kalman Filter to obtain the distribution of the state vector at the final time point, conditional on all of the observations: indeed, the input to the `dlmBSample()` routine (see Section 4.4.2) is the filtered observations rather than the original values. The Kalman Filter is also used to calculate the log posteriors $\{h(\boldsymbol{\theta}_i; \boldsymbol{Y})\}$ (Section 4.1.1), which are needed to define the importance weights $\{w_i\}$. For the mimics considered here, Kalman Filtering is one of the most time-consuming parts of the process: ideally therefore, the filter results from the FFBS algorithm would be stored for re-use when calculating the importance sampling weights. At present, unfortunately this is not possible using the routines provided in the `dlm` library.

5 Application to UK-averaged temperature and precipitation

This section presents the results of postprocessing both the UKCP18 and EuroCORDEX ensemble projections of UK-averaged temperature and precipitation, using the methodology described in Section 4. Implementation decisions, such as the choices of prior distributions for the descriptors and initial values for numerical maximisation of the posteriors, are documented in Appendices A2 and A3.

The analysis of UKCP18 projections follows the exchangeable ensembles methodology of Section 4.2, while the EuroCORDEX analysis uses the structured methodology of Section 4.3 — but omitting the interaction terms in the mimic specifications (24)–(26) because the ensemble contains too few runs to estimate them. Moreover, EuroCORDEX ensemble members driven by different runs from the same GCM are considered to be replicates so that the GCM trend contributions to these members are all the same — although the interannual variation about the trends is independent between runs.

Results for annual temperature means and precipitation totals are presented first, with separate results for each season considered briefly afterwards. Further visualisations of these results, together with corresponding analyses for each of the main administrative regions of the UK, are available from `https://github-pages.ucl.ac.uk/eurocordex-uk-plots/`.

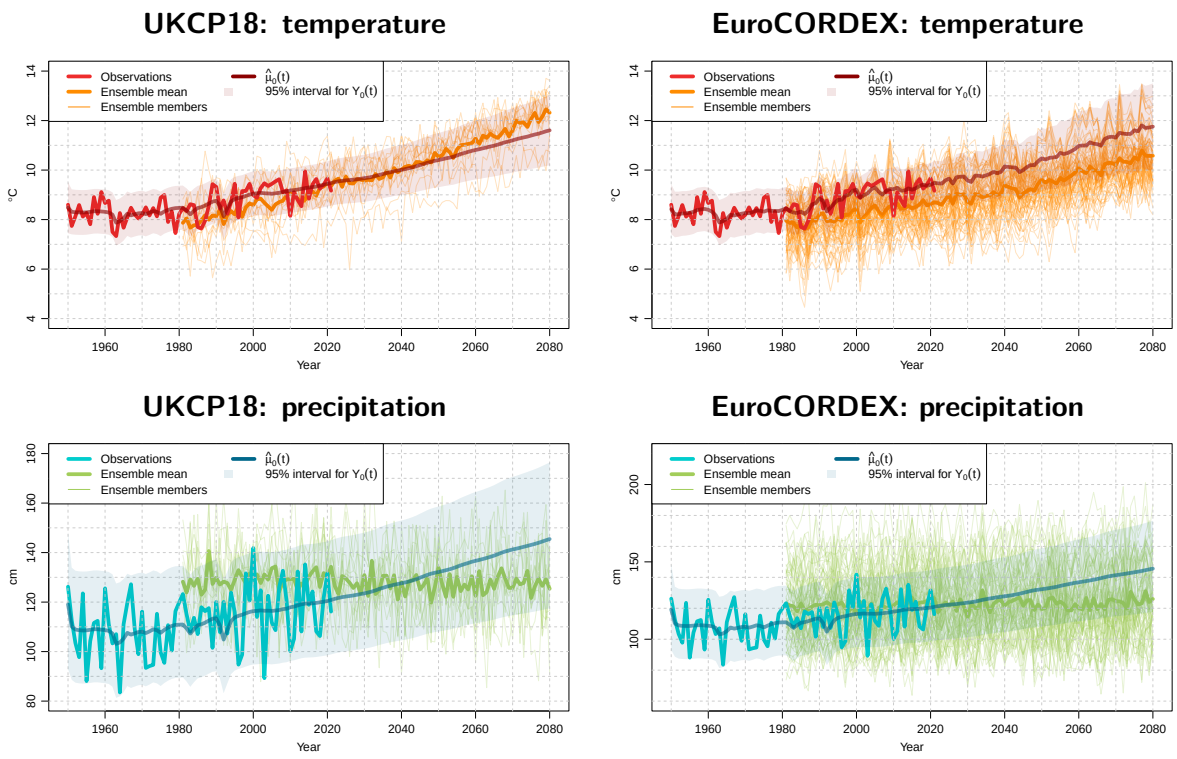

Figure 5: Postprocessed ensembles of annual UK-averaged temperature means and precipitation totals, based on ensemble outputs from 1980–2080 and observations from 1950–2020.

5.1 Annual temperature means and precipitation totals

Figure 5 shows the postprocessed UKCP18 and EuroCORDEX projections for both temperature and precipitation, obtained using the MAP estimates of the relevant descriptors in the same way as for Figure 4. The data shown in each panel are identical to those in Figure 1, but the postprocessed results also show the estimated real-world trend $\hat{\mu}_0(\cdot)$ together with 95% prediction intervals for the real-world annual values themselves.

Focusing first on the temperature projections in Figure 5, both the trend estimates $\hat{\mu}_0(\cdot)$ and the 95% postprocessed prediction intervals are very similar for the two ensembles: by 2080, $\hat{\mu}_0(\cdot)$ has reached just under 12°C in both cases, with a 95% prediction interval running from 10°C to around 13°C (slightly higher for the EuroCORDEX ensemble). This is despite appreciable differences between the raw ensembles, as discussed in Section 2.5: in particular, the postprocessed trend estimates suggest that the UKCP18 ensemble starts too cool and warms too rapidly, whereas the EuroCORDEX ensemble remains slightly too cool throughout. There are small differences between

the postprocessed projections and those obtained without using the ensemble data in Figure 4: the projected temperatures in Figure 5 warm slightly more rapidly and the uncertainty intervals are slightly wider than those based solely on the observations. The differences (particularly the increased uncertainty) are presumably attributable to the fact that the ensemble members indicate that future trends may depart slightly from the EBM approximation upon which Figure 4 is based.

The apparent differences between the postprocessed precipitation projections on the bottom row of Figure 5 are due solely to different vertical scales on the plots (the result of the bulk processing software used to produce them). In fact, the postprocessed projections from both ensembles are once again in close agreement, with the estimated trends $\hat{\mu}_0(\cdot)$ both reaching around 145cm per annum by 2080 and with 95% prediction intervals ranging from 120cm to just under 180cm. Once again, the differences compared with the 'observation-only' results in Figure 4 are small — in this case however, it is notable that neither ensemble captures the increasing trend in observed precipitation, so the form of the trends in the postprocessed projections is strongly determined by the assumed connection with the forcing series shown in Figure 2. This assumption is arguably harder to justify for precipitation than for temperature, for reasons discussed in Section 4.2; hence the precipitation results should be treated with more caution than those for temperature. Nonetheless, it is clear that the ensembles themselves completely fail to capture the observed increase in precipitation over the last 70 years: this suggests either that there are deficiencies in the representation of precipitation in the GCMs and RCMs, or that there is a problem with the observational dataset used to represent the real climate in the present analysis.

Figure 6 shows the analogue of Figure 5, but now accounting for uncertainty in the descriptor estimates using sampling from the approximate Gaussian posterior as described in Section 4.4.1. The trend estimates $\{\hat{\mu}_0(t)\}$ here are computed as means across 1000 antithetic samples from the approximate posterior, balanced for mean and dispersion; the 95% prediction intervals are computed from the quantiles of the same 1000 samples. A comparison with Figure 5 reveals very little difference, suggesting that descriptor parameter uncertainty is relatively unimportant in the present context.

The posterior distributions shown in Figure 6 are approximate, due to the use of Gaussian distributions when sampling from the posterior distribution of the descriptors. Importance sampling can be used to reweight these distributions and hence characterise the exact posterior, as discussed in Section 4.4.4. The results are shown in Figure 7. The first three plots here are very similar to the corresponding results in Figures 5 and 6 — once again, suggesting that descriptor uncertainty is relatively unimportant here.

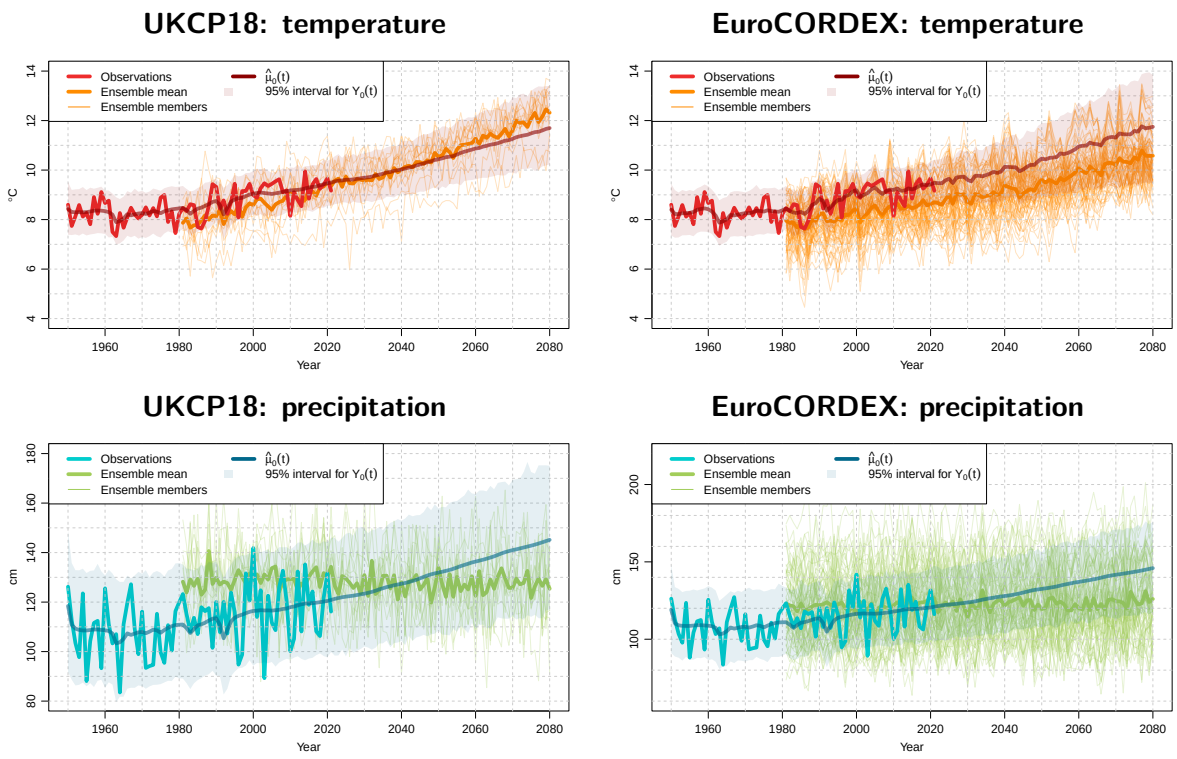

Figure 6: As Figure 5, but accounting for parameter uncertainty using Gaussian approximation to the posterior distribution of descriptors for each mimic.

No prediction interval is visible for the EuroCORDEX precipitation projections in Figure 7, however. The reason for this can be deduced from Figure 8 in which, for each set of projections, the importance weights have been sorted in decreasing order and cumulated. If the weights were all equal then these cumulative traces would appear as diagonal lines: this scenario corresponds to the case when the posterior distribution is exactly Gaussian so that importance sampling is not needed (see the definition of importance weights following equation (32)). Conversely, if the cumulative trace increases immediately to 100% as in the final panel of Figure 8, then the total weight must be concentrated in just one or two samples: this indicates that the importance sampling weights may have an infinite variance, which invalidates the procedure (see Section 4.4.4). For the remaining three panels, the traces lie substantially above the diagonal but the total weight is distributed over a substantial proportion of the samples: this suggests that the importance sampling has made meaningful adjustments to the initial Gaussian approximation in the corresponding analyses.

In situations where the importance sampling weights may have infinite variance, as noted in

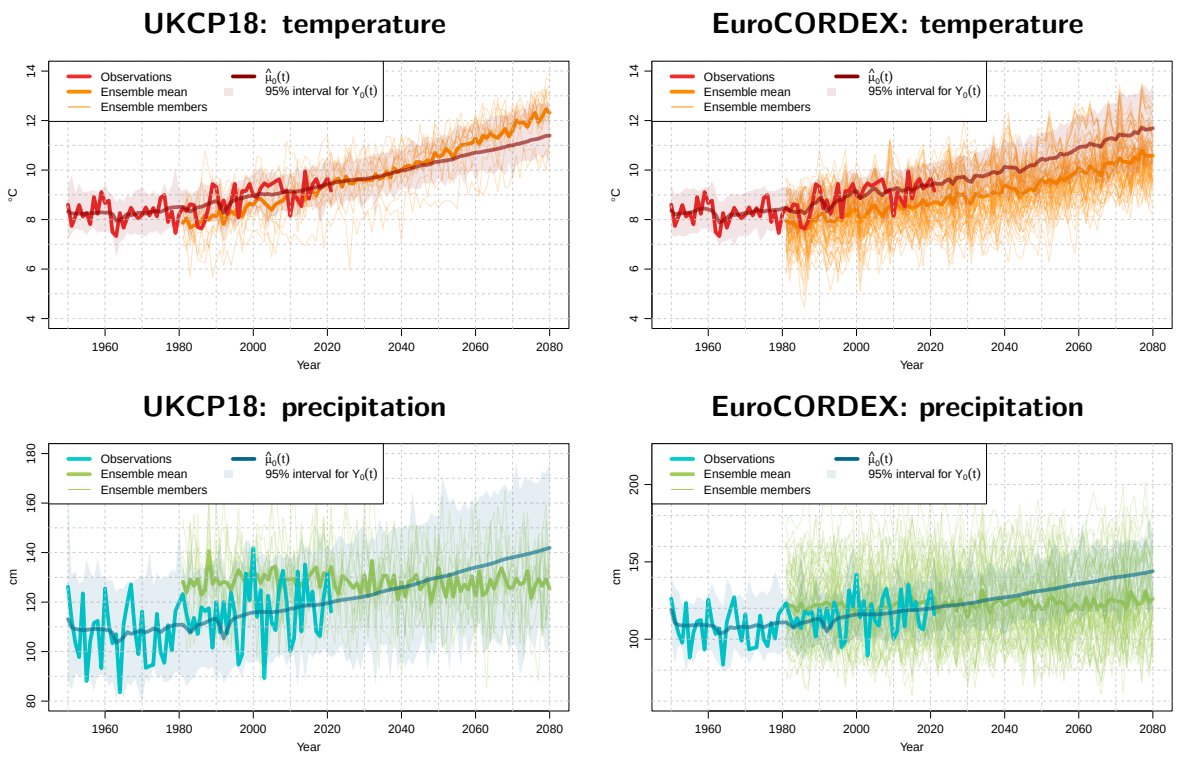

Figure 7: As Figure 6, but using importance sampling weights.

Section 4.4.4 a possible resolution is to replace the Gaussian approximation to $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$ with a heavier-tailed distribution when sampling the descriptors. Accordingly, the analysis of EuroCORDEX precipitation projections has been repeated using a multivariate Student t distribution on five degrees of freedom. In Figure 9, the resulting postprocessed projections are now much more similar to those from Figures 5 and 6; and the plot of cumulative importance weights no longer gives any cause for concern.

From the analysis of these four sets of annual projections, the implication so far is that any attempt to incorporate descriptor uncertainty, either with or without importance sampling, makes no material difference to the postprocessed projections — except when the importance sampling fails. Moreover, the calculations needed for a full uncertainty assessment are relatively time-consuming: for the exchangeable UKCP18 ensemble the time taken to estimate the descriptors and run the Kalman Smoother for each variable is around 90 seconds on a fast laptop, then an additional 4 minutes to sample from the Gaussian approximation to the full posterior, or 7 minutes for the importance sampling. The corresponding timings for the structured EuroCORDEX ensemble are around 3 minutes for the initial estimation and Kalman smoothing; 6 minutes for

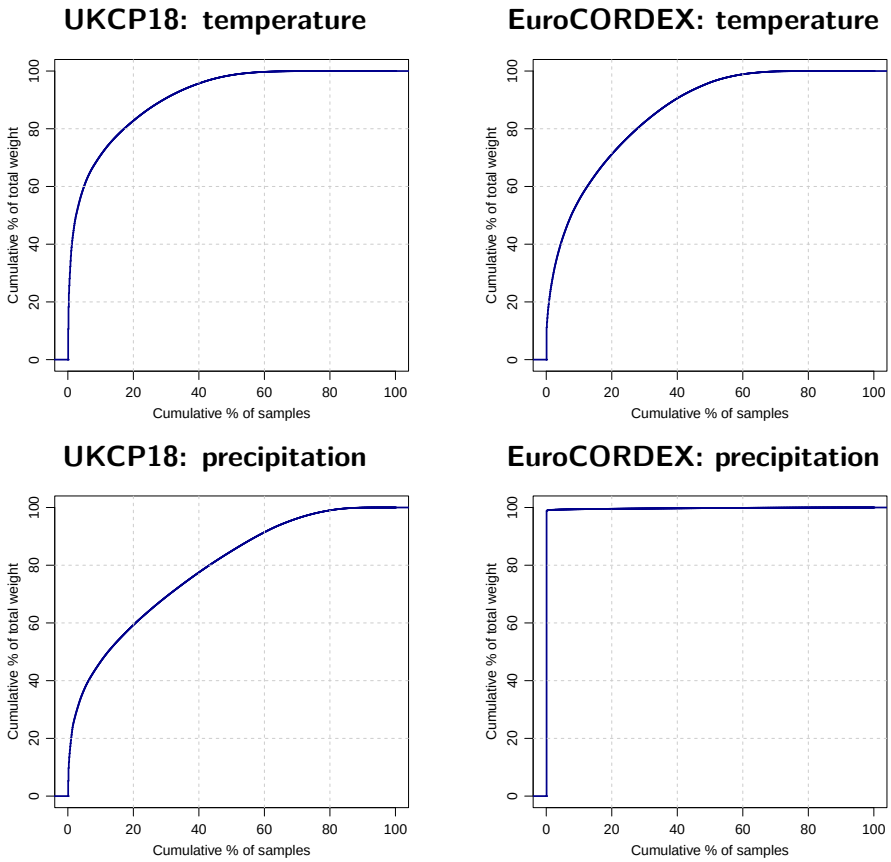

Figure 8: Cumulative contributions of sorted importance sampling weights for the analyses shown in Figure 7.

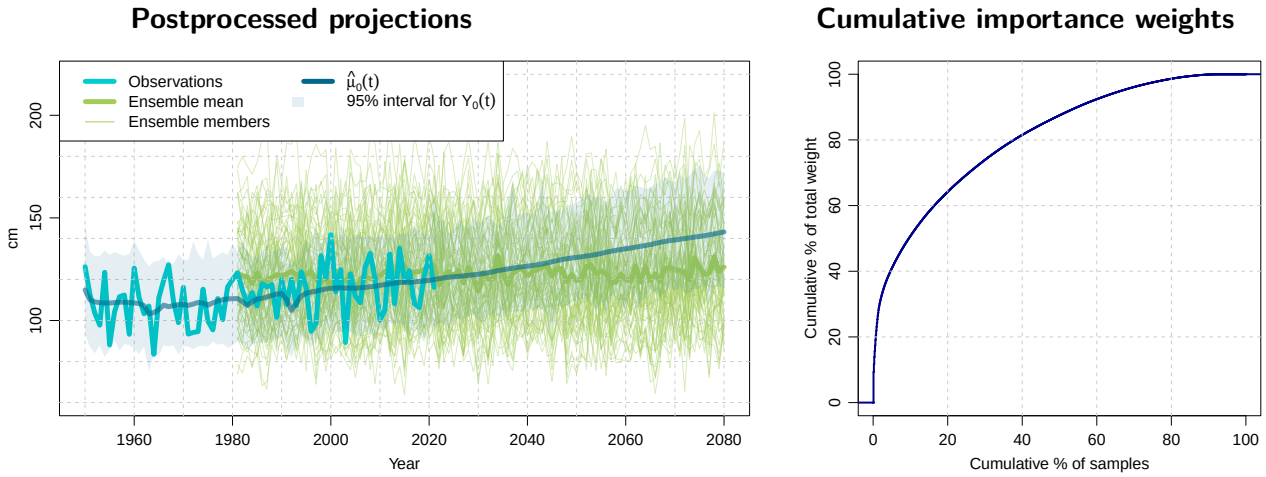

Figure 9: EuroCORDEX precipitation projections: full posterior distribution obtained via importance sampling with a t_5 initial approximation to the posterior, in place of the Gaussian approximation used in Figures 7 and 8.

the Gaussian approximation to the posterior; and 11–12 minutes for the importance sampling. A substantial portion of the difference between the timings for the Gaussian approximation and importance sampling is associated with the need to run the Kalman Filter twice to carry out importance sampling, as noted in Section 4.4.4. These conclusions appear to hold in general, as evidenced by the online archive of postprocessing results for national, regional, annual and seasonal analyses. For the reporting of seasonal results below therefore, for brevity we focus solely on the initial plots neglecting descriptor uncertainty.

5.2 Seasonal results

Figures 10–13 show the postprocessing results for each of the four seasons. In broad terms, the results are similar to those from the annual analyses above. In particular, the postprocessed projections from the two ensembles are very similar each other: in almost all cases where they appear different, this is due to the use of different vertical axis scales to accommodate the raw ensemble members. This suggests, once again, that the proposed postprocessing methodology is able to resolve the inter-ensemble differences to a large extent: this is mainly achieved by correcting for the systematic discrepancies between the members of each ensemble and the real climate system.

It is perhaps unnecessary to discuss all of the systematic ensemble discrepancies in Figures 10–13 in detail. It is nonetheless worth highlighting some features of the results, in the context of the generally accepted wisdom that the UK is heading for warmer, wetter winters and hotter, drier summers (Murphy et al., 2019). Thus:

- The postprocessed trend $\hat{\mu}_0(\cdot)$ in winter temperatures is in close agreement with the UKCP18 ensemble mean and, by 2080, almost 2°C warmer than the EuroCORDEX ensemble mean. The upper end of the 95% uncertainty interval suggests that it would not be surprising for some individual winters to be warmer than the warmest of the UKCP18 ensemble members by the second half of the century.
- Both ensembles suggest a small increase in winter total precipitation before postprocessing. The postprocessed projections indicate a much more substantial increase however, with the mean winter precipitation in 2080 (around 55cm) being around the maximum observed to date. As noted previously, the mimics for precipitation are potentially less reliable than those for temperature because the physical basis for the EBM-approximated trend is less convincing. Nonetheless, the observed winter precipitation series clearly shows a much faster increase than either of the ensembles prior to 2020: it would be surprising if this

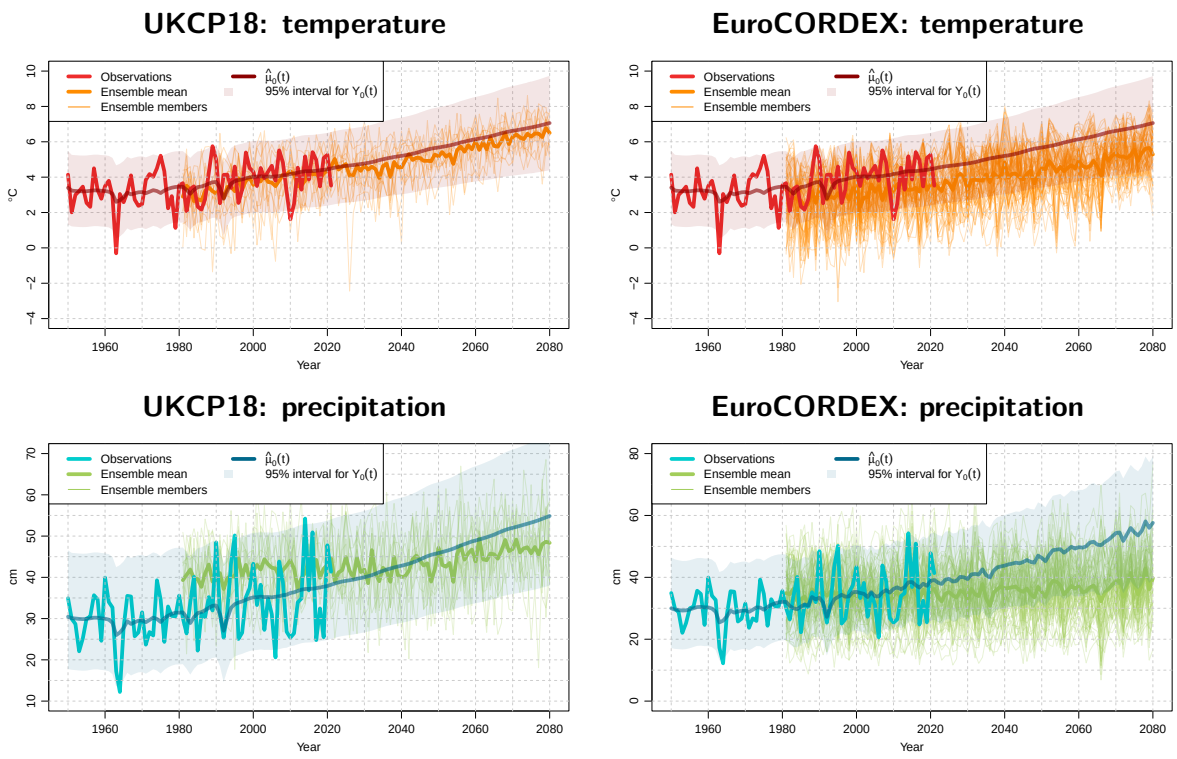

Figure 10: Postprocessed ensembles of winter (DJF) UK-averaged temperature means and precipitation totals, based on ensemble outputs from 1980–2080 and observations from 1950–2020.

trend does not continue at least into the near future as indicated by the postprocessed projections.

- The effect of postprocessing in spring is mainly to remove biases in the ensembles. The estimated temperature trends are similar to those in the ensemble means — albeit warming slightly more slowly than UKCP18 and slightly faster than EuroCORDEX — while there is little trend in spring precipitation, with the postprocessing merely removing wet biases in the ensembles during this season.

- Postprocessed trends in summer temperatures are much closer to the EuroCORDEX ensemble mean than to the UKCP18 ensemble: the postprocessing suggests that the latter warms much too quickly in this season, with the ensemble mean close to the upper end of the postprocessed uncertainty intervals.

- Both unprocessed ensembles project a decline in total summer precipitation, which is greater for UKCP18 than for EuroCORDEX. With the usual caveats regarding the EBM approx-

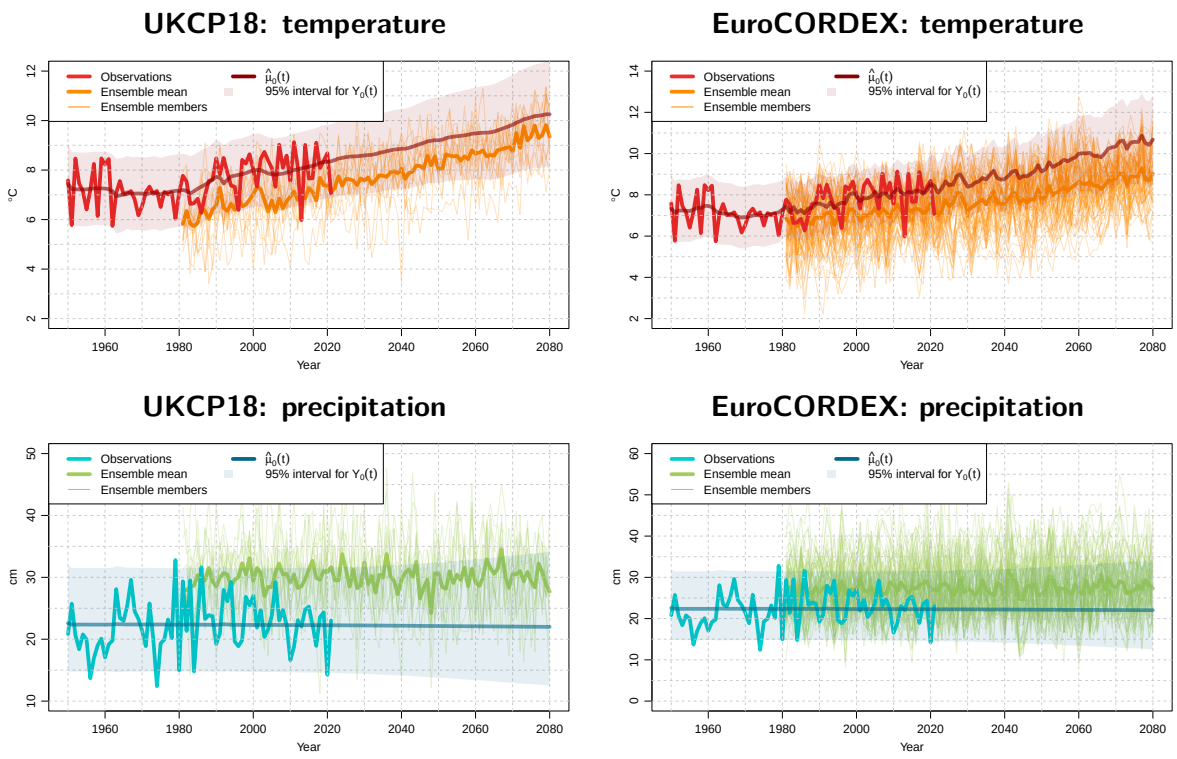

Figure 11: Postprocessed ensembles of spring (MAM) UK-averaged temperature means and precipitation totals, based on ensemble outputs from 1980–2080 and observations from 1950–2020.

imation to precipitation trends, the postprocessed projections here indicate *increases* in summer precipitation — on the basis, presumably, of a small but steady increase between 1980 and 2020 that is not reflected in the ensembles. There are also discernible differences between the postprocessing results for UKCP18 and EuroCORDEX: the former trend is smoother and, prior to 1980, does not track the apparent decline in the observations so closely as the latter. The reason for these differences is unclear: they remain after accounting for descriptor uncertainty by sampling from the full posterior either with or without importance sampling (plots available online). A more coherent picture emerges for regional summer precipitation, however (again, plots are online): the projected increase in summer precipitation is strongest in the north of the UK and weakest in the south-east, where the postprocessed trends are almost flat. It is also worth noting that trends in total precipitation can arise from changes in both frequency and intensity of precipitation events: with some consensus that summer precipitation in the UK may be getting less frequent but more intense (Murphy et al., 2019), it is possible that the analyses in Figure 12 may be conflating

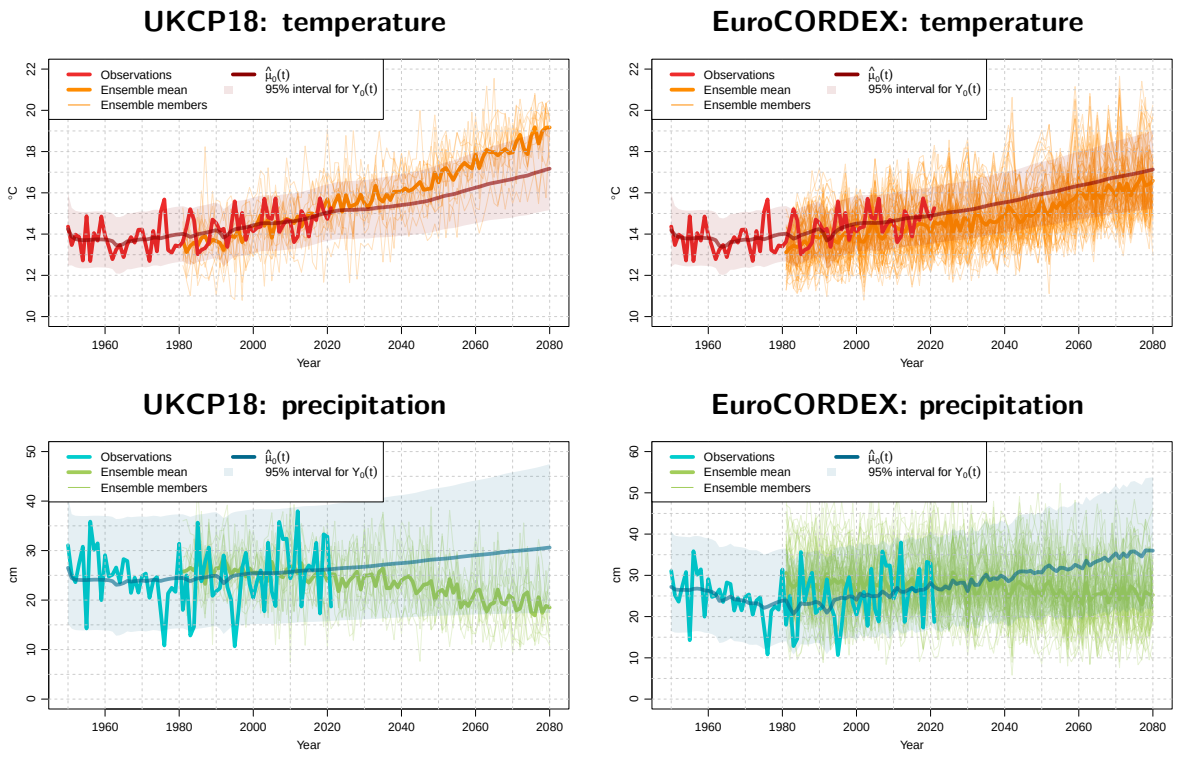

Figure 12: Postprocessed ensembles of summer (JJA) UK-averaged temperature means and precipitation totals, based on ensemble outputs from 1980–2080 and observations from 1950–2020.

two different sources of variation and hence that it would be preferable to examine them separately.

- In autumn, the postprocessed trend corrects for a slight cool bias in the EuroCORDEX ensemble, and warms more slowly than the UKCP18 ensemble mean; there is also an indication of a slight increase in total autumn precipitation by 2080, which is not present in either of the unprocessed ensembles.

6 Summary and discussion

6.1 Overall comments

The methodology developed in this report is designed to account for many of the known features of ensembles of climate projections, within a flexible framework that imposes relatively few as-

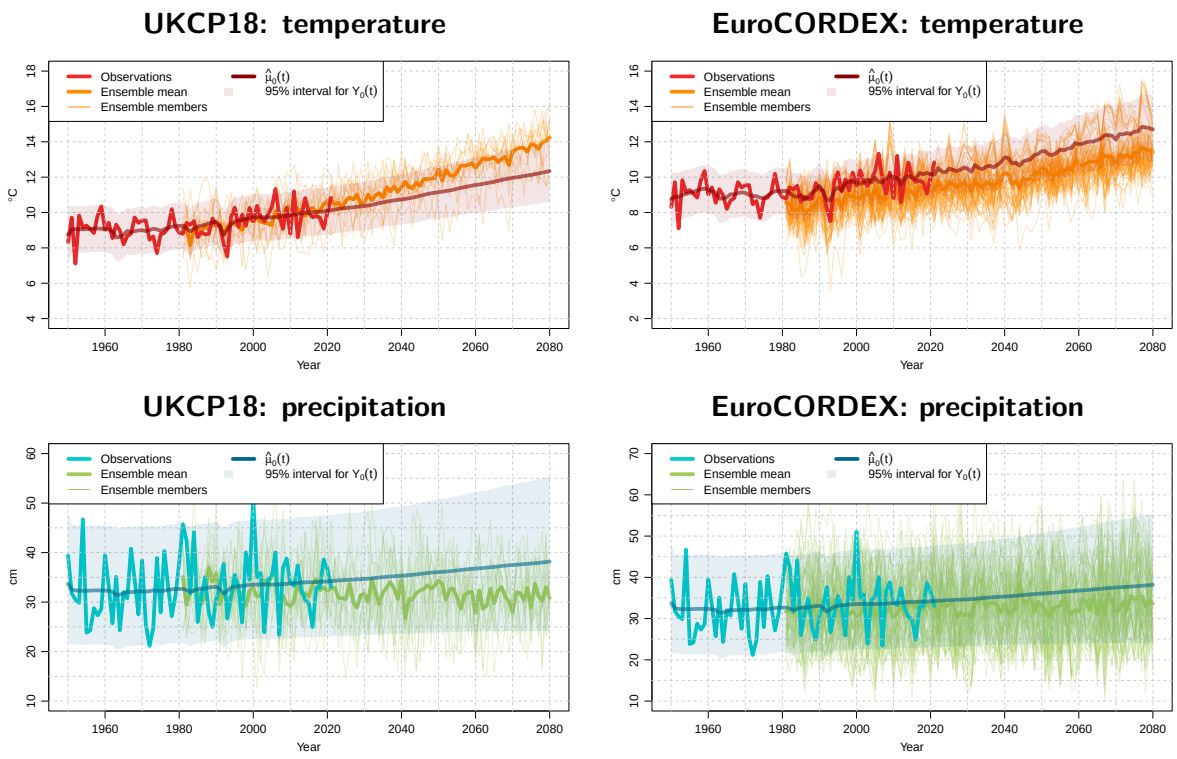

Figure 13: Postprocessed ensembles of autumn (SON) UK-averaged temperature means and precipitation totals, based on ensemble outputs from 1980–2080 and observations from 1950–2020.

sumptions and that attempts to incorporate physical insights into the statistical representation by appealing to the dynamics of (albeit very simple) energy balance models. In many respects it is similar to the approach proposed by Sansom et al. (2021), although differing in important details — notably the choices made here with computational tractability in mind, that allow the routine processing of large numbers of projections. The treatment of structured regional ensembles in Section 4.3 is also completely new.

To indicate the scale of analyses that have been undertaken: the full set of results available from https://github-pages.ucl.ac.uk/eurocordex-uk-plots/ includes analyses of annual and seasonal temperature means and precipitation totals, averaged over the entire UK and for each of 14 administrative regions and for four different ensembles (postprocessed results for the UKCP18 global and CMIP5 ensembles are also available). The time taken to process the entire set of 600 projections on UCL's general-purpose HPC cluster was around three days in total — including the generation of samples from the full posterior distribution, both with and without importance

sampling. In all of the examples considered here, this full posterior sampling was found to make little difference to the final results: this potentially offers the opportunity to cut computational costs still further in the future, since the full posterior sampling is the most time-consuming part of the process.

The flexibility of the proposed framework is illustrated by the range of 'adjustments' to the raw ensembles that are found in Section 5: the postprocessed results are seen to adjust for biases in the raw ensembles, and also to 'correct' for rates of change in the ensembles that appear too fast or too slow by comparison with the observations. Importantly, these adjustments are not specified directly: rather, they are implied by the statistical representation of the ensemble structure and its relation to reality. This approach is intended to ensure that the resulting uncertainty assessments are as transparent and defensible as possible, and hence suitable for use with formal decision-making frameworks.

The strongest assumption in the methodology is the link between the underlying trend and the output of a simple EBM, which is calibrated as part of the postprocessing procedure (via the estimation of the inertia coefficient ϕ as part of the descriptor vector, and the forcing coefficient γ as part of the state vector). The random-walk drift processes $\{\beta_0\}$ are intended to alleviate this assumption to some extent, because they have the flexibility to allow the estimated trend to move arbitrarily far from the EBM solution over sufficiently long time periods if this is indicated by the data and ensemble outputs.

Although these statistical assumptions are fairly weak therefore, there is one situation in which they may have a substantial influence on the results: this is when the historical observations contain a trend but the ensembles do not (as encountered throughout the precipitation analyses in this report). Essentially the difficulty here is that the ensemble value of γ is zero, so that the ensemble can provide no information on the plausibility of the EBM approximation to the trend by comparison with the observations. The problem with precipitation totals could potentially be resolved by studying frequency and intensity separately, since trends in each of these components may be better defined. At present however, this is just speculation.

6.2 Implications for users of projections

Faced with large ensembles of climate projections, it is natural for users to ask the questions (1) which members to select for their analyses (2) how to correct for biases. According to the methodology developed in this report, the answer to the first question is: *select none of them, but use all of them*; and the second question is then redundant because the postprocessing

methodology directly targets the real climate system.

A purist would perhaps argue that, in an ideal world and conditional on the appropriateness of the assumed mimic structure, the formally correct procedure is to work with samples from the posterior distribution of the quantity of interest. At present however, it is likely that rather few users would be prepared to accept this argument — although some may be interested in using the postprocessed projections as a means of placing the ensembles in the context of what may be expected in reality. For example, the postprocessed temperature projections in Section 5 showed good agreement with the UKCP18 ensemble in winter, and with the EuroCORDEX ensemble in summer: users could exploit this information to choose the most appropriate ensemble for their application.

A more sophisticated interpretation of the postprocessed projections does require an examination of samples from the posterior distribution, however. In the plots of postprocessed projections seen so far, uncertainty has been represented by a coloured band representing a 95% prediction interval. This is a traditional means of indicating uncertainty in forecasts, projections or curves: however, it does not provide a clear indication of what the future time series of future temperatures might actually look like. It is not clear, for example, whether all of the potential futures share the same trend with an increasing level of interannual variation; or whether different possible futures have markedly different trends. To address this issue, Bowman (2019) suggests using animations in which each individual frame shows a sample from the posterior distribution (strictly speaking, from a joint distribution with the same mean and covariance as the posterior) but in which the transitions between frames are smooth: the smoothness of the transitions helps the viewer to process the information as it is presented. The animations are produced by sampling a small number of 'primary' frames from the posterior $\pi\left(\boldsymbol{Y}^{(\mathrm{new})}|\boldsymbol{Y}\right)$, and then interpolating between these primary frames in such a way that each interpolation also has the correct mean and covariance structure. Examples are provided as animated `GIF` images in the online archive of results: these provide a clear indication of the range of potential outcomes that can be expected — which include periods running counter to the overall trend, as well as individual years or seasons falling outside the ranges of the uncertainty bands shown in the static displays of this report.

6.3 Unresolved issues and future work

There are a few unresolved issues arising out of the work reported above. The most obvious one relates to trends in precipitation, which require further investigation as outlined above. In some way related to this: the mimics developed in Section 4 are designed for use with time

series containing no seasonal structure (hence the separate treatment of each season in Section 5) and with a constant level of interannual variation over time. Moreover, much of the underlying software is designed for use with Gaussian time series — although the relevant calculations only involve means and variances, which can be exploited in some situations (e.g. the use of gamma distributions in this report to ensure that sampled precipitation totals are non-negative). Such workarounds are not a universal panacea however; and, particularly when working with series that are highly non-Gaussian, it is likely that a more principled approach will be needed.

A further aspect that is not considered here is the simultaneous postprocessing of multiple variables: for example, temperature and precipitation have been considered separately throughout. In principle, the mimics of Section 4 can be extended straightforwardly to this setting: all that is required is to specify correlations between the irregular interannual variations $(\varepsilon_* (\cdot))$ and the trend drift increments $(Z_* (\cdot))$ for each quantity, although care will once again be needed with the specification of shared discrepancies and sum-to-zero constraints to obtain a self-consistent system.

In Section 3 it was noted that the performance of probabilistic weather forecasting systems can be evaluated by comparing with the subsequent outcomes; but that this is not possible in the context of climate projections. Nonetheless, the existence of multiple ensembles and greenhouse gas emissions scenarios does provide a synthetic testbed for any postprocessing methodology: for example, one could envisage an experiment in which the postprocessing technique is applied to the UKCP18 ensemble but with one of the EuroCORDEX ensemble members substituted for the real climate — the goal in that case being to use the postprocessed UKCP18 ensemble to infer the trajectory of the EuroCORDEX member beyond the year 2020. An alternative would be to calibrate the mimics using projections from one emissions scenario, and then to use them to make inferences about the corresponding simulator outputs for an alternative scenario. Either approach would provide an opportunity to evaluate the credibility of the postprocessing methodology under a fairly realistic setting in which there are systematic differences between the ensemble and the system for which projections are ultimately required.

As a final comment: the posterior sampling methodology in Section 4.4 generates realisations of 'potentially observable' quantities (denoted $\boldsymbol{Y}^{\text{new}}$ there). Throughout the report, the focus here has been on quantities arising in the real climate system. However, this restriction is not necessary: one could also envisage, for example, generating series corresponding to missing ensemble members e.g. to impute the missing members from an unbalanced regional ensemble. In this context, the mimics can arguably be regarded as very simple emulators in the sense that they

provide an approximation to the function mapping ERF (as input) to the climate time series of interest (as output).

Acknowledgements

The authors acknowledge the World Climate Research Programme's Working Group on Regional Climate, and the Working Group on Coupled Modelling, former coordinating body of CORDEX and responsible panel for CMIP5. We also thank the climate modelling groups for producing and making available their model output; the Earth SystemGrid Federation infrastructure, an international effort led by the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison; the European Network for Earth System Modelling and other partners in the Global Organisation for Earth System Science Portals (GO-ESSP).

Finally, we express our deep appreciation to Ed Lowther and Idil Ozdemir of UCL's Centre for Advanced Research Computing, for their assistance with the HPC runs and for their development of the online interface to the results.

Technical appendices

A1 Form of the EBM-inspired trend $\mu_0(t)$

In the proposed framework, the form of the underlying trend $\mu_0(t)$ is specified via the difference equation (4). The solution to this equation can be obtained by writing the equation as

$$(1 - \phi_0 B)\mu_0(t) = \gamma_0 f(t) + B\beta_0(t) \ ,$$

where B denotes the backshift operator (i.e. $B\mu_0(t) = \mu_0(t-1)$, $B\beta_0(t) = \beta_0(t-1)$ and so on — see Chandler and Scott 2011, Section 5.1.8 for an introduction to this operator and its properties). Thus

$$\mu_0(t) = \gamma_0 \left(1 - \phi_0 B\right)^{-1} f(t) + \mathcal{I}_0(t) \ ,$$

where the process $(\mathcal{I}_0(t))$ represents the cumulative effect of the drift $(\beta_0(\cdot))$. As a transformation of the random walk $(\beta_0(\cdot))$, this processes has a complex stochastic structure. If $|\phi_0| < 1$ then the standard properties of the backshift operator allow the first term to be written as a convergent

infinite sum, so that

$$\mu_0(t) = \gamma_0 \sum_{u=0}^{\infty} \phi_0^u f(t-u) + \mathcal{I}_0(t) \ . \tag{A.1}$$

Putting $t = 0$ in this expression, we have

$$\mu_0(0) = \gamma_0 \sum_{v=0}^{\infty} \phi_0^v f(-v) + \mathcal{I}_0(0) = \gamma_0 \sum_{u=t}^{\infty} \phi_0^{u-t} f(t-u) + \mathcal{I}_0(0) \ ,$$

the final step obtained by putting $t - u = v$ in the sum so that $v = u - t$. It follows that $\phi_0^t \mu_0(t) = \gamma_0 \sum_{u=t}^{\infty} \phi_0^u f(t-u) + \phi_0^t \mathcal{I}_0(0)$ which, from (A.1), is equal to $\mu_0(t) - \gamma_0 \sum_{u=t-1}^{\infty} \phi_0^u f(t-u) - \mathcal{I}_0(t) + \phi_0^t \mathcal{I}_0(0)$. Hence $\mu_0(t) = \phi_0^t \mu_0(t) + \gamma_0 \sum_{u=t-1}^{\infty} \phi_0^u f(t-u) + \mathcal{I}_0(t) - \phi_0^t \mathcal{I}_0(0) = \phi_0^t \mu_0(t) + \gamma_0 \sum_{u=t-1}^{\infty} \phi_0^u f(t-u) + \mathcal{J}_0(t)$ say, in agreement with result (6).

A2 Choices of prior

The mimics in Sections 4.2 and 4.3 both use the same collection of descriptors: σ_0^2, ϕ_0, τ_0^2, σ_1^2, ϕ_1, α, $\tilde{\tau}_\omega^2$ and $\tilde{\tau}_1^2$. Their interpretation is summarised in Table A.1. As noted in Section 4.1, transformations of the descriptors are targeted in the MAP estimation procedure (log transformations for variances, and logit transformations for inertia coefficients). This section documents the choices of prior distribution that have been used for each of these transformed descriptors.

Prior choices will in general depend on the anticipated behaviour of the quantities of interest; therefore separate settings must be considered for temperature and precipitation, and for annual and seasonal series. Potential prior settings are also suggested here for an analysis of log-transformed precipitation, although the results of this analysis do not appear in the main report. Throughout, choices are deliberately conservative in the sense that they are designed to have a minimal impact on the results except where the log-likelihood is flat. This has been checked for several cases: in all instances where unpenalised maximum likelihood estimation was numerically stable, the estimates and projections were visually indistinguishable from those based on MAP estimation.

A2.1 Temperature in °C

The following prior choices are adopted for the analysis of annual mean temperatures:

- $\log \sigma_0^2$: we can confidently assert that σ_0 is between 0.005 and 5. This is because: if $\sigma_0 < 0.005$, then 95% of annual temperatures will be within 0.01°C of the overall trend;

Descriptor	Interpretation	Equation
σ_0^2	Variance of interannual variations about trend for real climate	(1)
ϕ_0	Trend inertia coefficient for real climate	(4)
τ_0^2	Variance of EBM approximation drift increments for real climate	(5)
σ_1^2	Common variance of interannual variations about trend for ensemble members	(8)†
ϕ_1	Common trend inertia coefficient for ensemble members	(9)†
α	Scaling of real-climate trends and approximation drifts in specification of ensemble consensus processes	(11)
$\tilde{\tau}_\omega^2$	Variance of shared discrepancy drift increments in ensemble	(12)†
$\tilde{\tau}_1^2$	Common variance of increments, for individual member / simulator drift discrepancies relative to ensemble consensus	(13)†

Table A.1: Descriptors used in the mimics of Sections 4.2 and 4.3. Equation numbers indicate where descriptor is introduced; '†' superscripts indicate where the original specification has been simplified by setting common values for all ensemble members.

and if $\sigma_0 > 5$ then fewer than 95% of annual temperatures will be within 10°C of the trend. It follows that $\log \sigma_0$ is between $\log 0.005 = -5.3$ and $\log 5 = 1.6$; and $\log \sigma_0^2$ is between -10.6 and 3.2. Interpreting this interval as covering the central 95% portion of the required normal distribution (i.e. 'mean $\pm$ 2 standard deviations'), the prior mean and standard deviation are -3.7 and 3.45 respectively.

- $\log\left[\phi_0/\left(1-\phi_0\right)\right]$: a $N(0, 5^2)$ prior is adopted. This ensures that ϕ_0 lies between 0 and 1, and is likely to be in the range $(0.0001, 0.9999)$ — on the basis that values less than 0.0001 (resp. greater than 0.9999) are unlikely to be distinguishable from 0.0001 (resp. 0.9999) in practice so that these values can be considered as an effective range.

- $\log \tau_0^2$: over a period of T years, the EBM approximation drift will change by an amount that is normally distributed with mean 0 and variance $T\tau_0^2$. Over a century, this cumulative change is unlikely to be less than ± 0.01°C or more than $\pm 10°$. This implies that the standard deviation of the 100-year change is between 0.005 and 5, whence the variance is between 0.005^2 and 5^2 and τ_0^2 is between $0.005^2/100$ and $5^2/100$. $\log \tau_0^2$ is therefore between -15.2 and -1.4, so that the prior mean and standard deviation are -8.3 and 3.45 respectively.

- α: this is expected to be around 1 (so that the ensemble consensus trend is roughly equal to the real-world trend on average), and it is unlikely to be negative. We set a prior of

$N(1,1)$, therefore.

- $\log \tilde{\tau}_{\omega}^2$: although this is likely to be substantially smaller in magnitude than τ_0^2, we use the same prior specification — on the basis that the choice is guaranteed to be conservative and is unlikely to influence any aspects of the estimation except its numerical stability.
- $\log \sigma_1^2$: the same prior is used for $\log \sigma_1^2$ as for $\log \sigma_0^2$ because the two parameters have similar interpretations relating to the magnitude of interannual variation.
- $\log \tilde{\tau}_1^2$: as with $\tilde{\tau}_1^2$, we use the same prior as for τ_0^2.
- $\log\left[\phi_1/\left(1-\phi_0\right)\right]$: The same prior is used as for $\log\left[\phi_0/\left(1-\phi_0\right)\right]$.

The same arguments — and hence the same prior choices — apply to seasonal mean temperatures.

A2.2 Precipitation totals in cm

Priors for the parameters in precipitation models are the same as for temperature except in cases that are dependent on measurement units. A further consideration is that although one might still reasonably expect some correspondence between precipitation trends and forcings, the specific EBM-inspired mimic formulation is less defensible for regional precipitation than for regional temperature: as a result, the approximation drift processes $\{\beta_{\cdot}(\cdot)\}$ are likely to be relatively more important for precipitation than for temperature.

For annual precipitation totals, the judgements made herein are as follows:

- $\log \sigma_0^2$ and $\log \sigma_1^2$: it is unlikely that more than 5% of annual precipitation totals will be within 10cm of, or more than 1m from, their expected values. This suggests that σ_0 is between 5cm and 50cm, so that $\log \sigma_0^2$ is between 3.2 and 7.8. Interpreting this interval as covering the central 95% portion of the required normal distribution, the prior mean and standard deviation for $\log \sigma_0^2$ are 5.5 and 1.15 respectively. The same prior is adopted for $\log \sigma_1^2$.
- $\log \tau_0^2$, $\log \tilde{\tau}_{\omega}^2$ and $\log \tilde{\tau}_1^2$: over a century, the cumulative change $\beta_0(t+100)-\beta_0(t)$ in the EBM approximation drift is unlikely to be less than 0.1cm or more than 10cm. This implies that the standard deviation of the 100-year cumulative drift is between 0.05cm and 5cm, whence the variance is between 0.05^2 and 5^2 so that τ_0^2 is between $0.05^2/100$ and $5^2/100$. $\log \tau_0^2$ is therefore between -10.6 and -1.4, and its prior mean and standard deviation are -6 and 2.3 respectively. As for temperature, the same priors are used for $\log \tilde{\tau}_{\omega}^2$ and $\log \tilde{\tau}_1^2$ as for $\log \tau_0^2$.

The same arguments can be applied to seasonal totals, but dividing the respective benchmarks by 4 to account for the fact that seasonal accumulations are over three months rather than 12. The respective variances are decreased by a factor of 16 therefore, whence the prior means for $\log \sigma_0^2$ and $\log \sigma_1^2$ are reduced from 5.5 to $5.5 - \log 16 = 2.7$. Similarly, the prior means for $\log \tau_0^2$, $\log \tilde{\tau}_\omega^2$ and $\log \tilde{\tau}_1^2$ are reduced from -6 to $-6 - \log 16 = -8.8$. The prior standard deviations are unaffected.

A2.3 Logged precipitation totals

If the precipitation analysis is done on the log scale, the prior considerations for variance parameters become slightly different because judgements apply to relative rather than absolute changes. The judgements are as follows for annual precipitation totals:

- $\log \sigma_0^2$ and $\log \sigma_1^2$: it is unlikely that more than 5% of annual precipitation totals will be within 1% of, or more than five times, their median values. The corresponding bounds for log precipitation totals are $\log 1.01 = 0.01$ and $\log 5 = 1.6$ respectively. σ_0 is between 0.005 and 0.8 therefore, so that $\log \sigma_0^2$ is between -10.6 and -0.4 Interpreting this as a 95% interval in the usual way, the prior mean and standard deviation for $\log \sigma_0^2$ are -5.5 and 2.6 respectively. The same prior is adopted for $\log \sigma_1^2$.
- $\log \tau_0^2$, $\log \tilde{\tau}_\omega^2$ and $\log \tilde{\tau}_1^2$: the EBM approximation drift process $\beta_0(\cdot)$ now refers to errors in the EBM approximation to the dynamics of log total precipitation. Over a century, the cumulative change $\beta_0(t+100) - \beta_0(t) = \sum_{i=1}^{100} Z_0(t+i)$ corresponds to a multiplicative change of $\exp\left[\sum_{i=1}^{100} Z_0(t+i)\right]$ in approximation error on the original precipitation scale. One might be surprised if the error were to change by more than a factor of 100 or by less than a factor of 1.01 in each direction on this time scale; this corresponds to $\sum_{i=1}^{100} Z_0(t+i)$ lying in the range $(\log 1.01, \log 100) = (0.01, 4.6)$. This implies that the standard deviation of the cumulative 100-year change in drift is between 0.005 and 2.3, whence the variance is between 0.005^2 and 2.3^2 and τ_0^2 is between $0.005^2/100$ and $2.3^2/100$. $\log \tau_0^2$ is therefore between -15.2 and -2.9, so that the prior mean and standard deviation are -9.1 and 3 respectively. As usual, the same priors are used for $\log \tilde{\tau}_\omega^2$ and $\log \tilde{\tau}_1^2$ as for $\log \tau_0^2$.

As these arguments apply to relative rather than actual changes, the same values can be used without modification when working with logged seasonal, rather than annual, precipitation totals.

A2.4 Initial values for the state vector

In the state-space formulation of a DLM, the system dynamics are expressed via linear relationships between $\boldsymbol{S}(t)$ and $\boldsymbol{S}(t-1)$ where $\boldsymbol{S}(t)$ denotes the state vector at time t: see equations (7) and (20), for example. To formulate the likelihood — and hence the posterior distribution — for the descriptor vector based on data for times $t = 1, \ldots, T$, it is therefore necessary to specify estimates of $\boldsymbol{S}(0)$, together with the covariance matrix of the associated estimation error. In the time series literature, these are typically specified via the 'equilibrium' stochastic properties of the state process (Chandler and Scott, 2011, Section 5.5). Since some aspects of the state dynamics are adopted primarily for convenience, however (for example, the random-walk specification of the drift process $(\beta_0(t))$ in (5), and its ensemble discrepancy counterparts), it is arguably undesirable to rely on these notional equilibria. An alternative, adopted here, is to take a Bayesian perspective and treat $\boldsymbol{S}(0)$ as an uncertain quantity requiring an informed prior distribution.

For both the exchangeable and structured ensembles, the state vector $\boldsymbol{S}(0)$ involves $\mu_0(0)$ (the expected value of the real-world quantity of interest); the drift approximation term $\beta_0(0)$; the forcing dependence coefficient γ_0; and the shared and member-specific discrepancies associated with each of these components. With the exception of μ_0, the prior means for all of these quantities are set to zero (for example, there is no *a priori* reason to believe that discrepancies or drift approximation terms should be either positive or negative); while the prior means for $\mu_0(0)$ are based on roughly typical values for the UK climate. For the prior covariance matrix, initially a diagonal matrix is created, with diagonal elements set to common values deduced from conservative assessments of likely ranges for the value of $\mu_0(0)$ (for other terms such as discrepancies, the likely ranges are invariably smaller than those for $\mu_0(0)$ so that the use of common prior variances throughout maintains conservatism): for example, for all of the temperature analyses it is considered very unlikely that the actual value of $\mu_0(0)$ will be more than 10°C away from its prior expectation, so the prior standard deviation is set to 5°C and the prior variance is 25. Having created this diagonal matrix, the corresponding conditional covariance matrix is computed to respect the sum-to-zero constraints on the various components of $\boldsymbol{S}(0)$ — in the same way as is done for the innovation vectors themselves (see Appendices A4 and A5).

Table A.2 summarises the prior means specified for $\mu_0(0)$ for each different quantity of interest, together with the prior variances used for the initial diagonal covariance matrix in the procedure described above.

Quantity of interest	Prior mean for $\mu_0(0)$	Initial prior variance
Annual mean temperature (°C)	10	25
DJF mean temperature (°C)	5	25
MAM mean temperature (°C)	10	25
JJA mean temperature (°C)	15	25
SON mean temperature (°C)	10	25
Annual total precipitation (cm)	120	10^4
DJF total precipitation (cm)	120	10^4
MAM total precipitation (cm)	30	625
JJA total precipitation (cm)	30	625
SON total precipitation (cm)	30	625

Table A.2: Prior means and initial variances used for the initial state vector $\boldsymbol{S}(0)$ for each quantity of interest. For precipitation analysis on a log scale, a log transformation is applied to the prior means but the variances are unchanged.

A3 Initial values for numerical MAP estimation

The MAP estimates of descriptors in the mimics of Sections 4.1, 4.2 and 4.3 are obtained via numerical optimisation of the log-likelihood or log posterior. For any such numerical scheme it is necessary to specify initial values for the optimisation algorithm.

For the observation-only mimic of Section 4.1, initial values for the variances σ_0^2 and τ_0^2 are obtained by fitting a smooth local linear trend model (Chandler and Scott, 2011, Section 5.5.5) to the series: this is a special case of (4) with $\gamma_0 = 0$ and $\phi_0 = 1$, and is stochastically equivalent to an ARIMA(0,2,2) process so that it can be fitted by transforming parameter estimates obtained using the `arima()` command in R. For the EBM-inspired mimic however, the meaningful range of values for ϕ_0 is $0 < \phi_0 < 1$ (the restriction $|\phi_0| < 1$ is required for the derivation of equation (6) in Appendix A1, and the restriction $\phi > 0$ is required to demonstrate equivalence with the continuous-time EBM solution as discussed following equation (6)). The initial value of ϕ_0 is thus set to 0.99, so that overall the initial values for the observation-only mimic are close to the optimum for a special case of the model.

For the ensemble mimics of Sections 4.2 and 4.3, initial values for σ_0^2, τ_0^2 and ϕ_0 are taken from a fit of the observation-only mimic to the historical climate observations. τ_ω^2 and ϕ_1 are initialised by fitting this observation-only mimic to the ensemble mean time series; $\tilde{\tau}_\omega^2$ is then computed as $\max(10^{-12}, \tau_\omega^2 - \tau_0^2)$. For the member-specific discrepancies about the ensemble consensus,

initial values for σ_1^2 and $\tilde{\tau}_1^2$ are obtained by fitting a smooth local linear trend to the centred ensemble members (i.e. after subtracting the ensemble mean from each member) and averaging the resulting estimates; and α is initialised to 1.

A3.1 The requirement for a negative definite Hessian matrix

Although the use of MAP estimation — together with the choice of initial values outlined above — is intended to provide a reasonable assurance of successful optimisation, it does not completely guarantee that the computed Hessian of the log-posterior will be negative definite. Negative definitiness is a necessary requirement for a maximum; and also a requirement for the uncertainty calculations outlined in Section 4.4.1, since the inverse of the Hessian is the large-sample posterior covariance matrix of the descriptor estimates. The reason for this could be either failure of the optimisation algorithm, or (more frequently) numerical inaccuracies in situations where the log posterior remains fairly flat despite the quadratic penalty introduced in the MAP estimation (see Section 4.1.1).

Several solutions to this issue have been explored, including the replacement of a calculated Hessian with the nearest positive definite matrix using the `nearPD()` function in the `Matrix` library in `R` (Bates et al., 2023), but this yields very unreliable standard errors. The eventual solution adopted is to shrink the eigenvalues of the computed Hessian towards their mean, via the formula given in equation 1 of Ollila et al. (2021).

A4 Covariance matrix for constrained innovations in an exchangeable ensemble

Here we derive expression (19) for the covariance matrix of the innovations $\tilde{Z}_1(t), \ldots \tilde{Z}_M(t)$ in an exchangeable M-member ensemble subject to the constraint $\tilde{Z}_\cdot(t) = \sum_{i=1}^{M} \tilde{Z}_i(t) = 0$. The derivation uses the standard result (Krzanowski, 1988, Section 7.2) that if $\boldsymbol{X}$ has a multivariate distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, and if $\boldsymbol{X}$ is partitioned as $\boldsymbol{X} = (\boldsymbol{X}_1' \;\; \boldsymbol{X}_2')'$ with a corresponding partitioning of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} ,$$

then the conditional distribution of $\boldsymbol{X}_1$ given $\boldsymbol{X}_2 = \boldsymbol{x}_2$ is itself multivariate normal with mean vector $\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\left(\boldsymbol{x}_2 - \boldsymbol{\mu}_2\right)$ and covariance matrix $\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}$.

Applying this result to the vector $\tilde{\boldsymbol{Z}}^*(t)$ defined in (17) and its covariance matrix given at (18), it is easily seen that conditioning on $\tilde{Z}_{\cdot}(t) = 0$ yields $\mathbb{E}\left(\tilde{\boldsymbol{Z}}(t)|\tilde{Z}_{\cdot}(t) = 0\right) = \boldsymbol{0}$ and

$$\text{Var}\left(\tilde{\boldsymbol{Z}}(t)|\tilde{Z}_{\cdot}(t) = 0\right) = \begin{pmatrix} \tilde{\tau}_1^2 & 0 & \cdots & 0 \\ 0 & \tilde{\tau}_1^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \tilde{\tau}_1^2 \end{pmatrix} - \begin{pmatrix} \tilde{\tau}_1^2 \\ \tilde{\tau}_1^2 \\ \vdots \\ \tilde{\tau}_1^2 \end{pmatrix} \begin{pmatrix} \tilde{\tau}_1^2 & \tilde{\tau}_1^2 & \cdots & \tilde{\tau}_1^2 \end{pmatrix} / M\tilde{\tau}_1^2 \,,$$

which simplifies to (19) as required. ■

A5 Covariance matrix for constrained innovations in a structured regional ensemble

In the mimic for structured regional ensembles considered in Section 4.3, constraints were imposed on the trends $\{\tilde{\mu}_{r\cdot}(t)\}$, $\{\tilde{\mu}_{\cdot g}(t)\}$ and $\{\tilde{\mu}_{rg}(t)\}$ for purposes of identifiability and interpretability. As with the exchangeable case of Section 4.2, these constraints imply corresponding sum-to-zero constraints on the innovations $\left\{\tilde{Z}_{r\cdot}(t)\right\}$, $\left\{\tilde{Z}_{\cdot g}(t)\right\}$ and $\left\{\tilde{Z}_{rg}(t)\right\}$ in the respective approximation error discrepancy processes $\left\{\tilde{\beta}_{r\cdot}(t)\right\}$, $\left\{\tilde{\beta}_{\cdot g}(t)\right\}$ and $\left\{\tilde{\beta}_{rg}(t)\right\}$.

The constraints $\sum_{r=1}^{R}\tilde{Z}_{r\cdot}(t) = 0$ and $\sum_{g=1}^{G}\tilde{Z}_{\cdot g}(t) = 0$ have the same form as those considered in Appendix A4, so the corresponding blocks of the innovation covariance matrix have the same form as (19) but with M replaced by R and G respectively. The constraints on the 'interaction' terms $\left\{\tilde{Z}_{rg}(t)\right\}$ lead to a more complex structure however, which we now derive.

For compactness of notation, we write $\tilde{Z}_{rg}$ in place of $\tilde{Z}_{rg}(t)$ throughout; and define $S_{r\cdot} = \sum_{g=1}^{G}\tilde{Z}_{rg}$ $(r = 1, \ldots, R)$ and $S_{\cdot g} = \sum_{r=1}^{R}\mathbb{Z}_{rg}$ $(g = 1, \ldots, G-1)$. Then, since the $\left\{\tilde{Z}_{rg}\right\}$ are distributed independently as $N(0, \tau_1^2)$ in the absence of constraints, we have $\mathbb{E}\left(S_{r\cdot}\right) = \mathbb{E}\left(S_{r\cdot}\right) =$

0; $\mathsf{Var}\left(S_{r\cdot}\right) = G\tau_1^2$; $\mathsf{Var}\left(S_{\cdot g}\right) = R\tau_1^2$;

$$\begin{aligned}
\mathsf{Cov}\left(\tilde{Z}_{rg}, S_{r^*\cdot}\right) &= \mathsf{Cov}\left(\tilde{Z}_{rg}, \sum_{g=1}^{G} \tilde{Z}_{r^*g}\right) = \begin{cases} \tau_1^2 & r^* = r \\ 0 & \text{otherwise;} \end{cases} \\
\mathsf{Cov}\left(\tilde{Z}_{rg}, S_{\cdot g^*}\right) &= \begin{cases} \tau_1^2 & g^* = g \\ 0 & \text{otherwise;} \end{cases} \\
\text{and} \qquad \mathsf{Cov}\left(S_{r^*\cdot}, S_{\cdot g^*}\right) &= \mathsf{Cov}\left(\sum_{g=1}^{G} \tilde{Z}_{rg}, \sum_{r=1}^{R} \tilde{Z}_{rg^*}\right) = \tau_1^2 \ .
\end{aligned}$$

The augmented vector

$$\left(\tilde{Z}_{11} \ \ldots \ \tilde{Z}_{1G} \ \tilde{Z}_{21} \ \ldots \ \tilde{Z}_{2G} \ \ldots \ \tilde{Z}_{R1} \ \ldots \ \tilde{Z}_{RG} S_{1\cdot} \ \ldots \ S_{R\cdot} \ S_{\cdot 1} \ \ldots \ S_{\cdot(G-1)}\right)$$

thus has mean $\mathbf{0}$ and covariance matrix

	$\tilde{Z}_{11}$	$\tilde{Z}_{12}$	$\ldots$	$\tilde{Z}_{1(G-1)}$	$\tilde{Z}_{1G}$	$\cdots$	$\tilde{Z}_{R1}$	$\tilde{Z}_{R2}$	$\ldots$	$\tilde{Z}_{R(G-1)}$	$\tilde{Z}_{RG}$	$S_{1\cdot}$	$S_{2\cdot}$	$\ldots$	$S_{R\cdot}$	$S_{\cdot 1}$	$S_{\cdot 2}$	$\ldots$	$S_{\cdot(G-1)}$
$\tilde{Z}_{11}$	τ_1^2	0	$\cdots$	0	0		0	0	$\cdots$	0	0	τ_1^2	0	$\cdots$	0	τ_1^2	0	$\cdots$	0
$\tilde{Z}_{12}$	0	τ_1^2	$\cdots$	0	0		0	0	$\cdots$	0	0	τ_1^2	0	$\cdots$	0	0	τ_1^2	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\tilde{Z}_{1(G-1)}$	0	0	$\cdots$	τ_1^2	0		0	0	$\cdots$	0	0	τ_1^2	0	$\cdots$	0	0	0	$\cdots$	τ_1^2
$\tilde{Z}_{1G}$	0	0	$\cdots$	0	τ_1^2		0	0	$\cdots$	0	0	τ_1^2	0	$\cdots$	0	0	0	$\cdots$	0
$\vdots$			$\vdots$			$\ddots$			$\vdots$				$\vdots$				$\vdots$		
$\tilde{Z}_{R1}$	0	0	$\cdots$	0	0		τ_1^2	0	$\cdots$	0	0	0	0	$\cdots$	τ_1^2	τ_1^2	0	$\cdots$	0
$\tilde{Z}_{R2}$	0	0	$\cdots$	0	0		0	τ_1^2	$\cdots$	0	0	0	0	$\cdots$	τ_1^2	0	τ_1^2	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\tilde{Z}_{R(G-1)}$	0	0	$\cdots$	0	0		0	0	$\cdots$	τ_1^2	0	0	0	$\cdots$	τ_1^2	0	0	$\cdots$	τ_1^2
$\tilde{Z}_{RG}$	0	0	$\cdots$	0	0		0	0	$\cdots$	0	τ_1^2	0	0	$\cdots$	τ_1^2	0	0	$\cdots$	0
$S_{1\cdot}$	τ_1^2	τ_1^2	$\cdots$	τ_1^2	τ_1^2		0	0	$\cdots$	0	0	$G\tau_1^2$	0	$\cdots$	0	τ_1^2	τ_1^2	$\cdots$	τ_1^2
$S_{2\cdot}$	0	0	$\cdots$	0	0		0	0	$\cdots$	0	0	0	$G\tau_1^2$	$\cdots$	0	τ_1^2	τ_1^2	$\cdots$	τ_1^2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$S_{R\cdot}$	0	0	$\cdots$	0	0		τ_1^2	τ_1^2	$\cdots$	τ_1^2	τ_1^2	0	0	$\cdots$	$G\tau_1^2$	τ_1^2	τ_1^2	$\cdots$	τ_1^2
$S_{\cdot 1}$	τ_1^2	0	$\cdots$	0	0		τ_1^2	0	$\cdots$	0	0	τ_1^2	τ_1^2	$\cdots$	τ_1^2	$R\tau_1^2$	0	$\cdots$	0
$S_{\cdot 2}$	0	τ_1^2	$\cdots$	0	0		0	τ_1^2	$\cdots$	0	0	τ_1^2	τ_1^2	$\cdots$	τ_1^2	0	$R\tau_1^2$	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$S_{\cdot(G-1)}$	0	0	$\cdots$	τ_1^2	0		0	0	$\cdots$	τ_1^2	0	τ_1^2	τ_1^2	$\cdots$	τ_1^2	0	0	$\cdots$	$R\tau_1^2$

Conditioning on $S_{1\cdot} = S_{2\cdot} = \ldots = S_{R\cdot} = S_{\cdot 1} = S_{\cdot 2} = \ldots = S_{\cdot(G-1)} = 0$ and using the standard

result given in Appendix A4 again, the conditional mean of $\tilde{Z}_{11}, \ldots, \tilde{Z}_{RG}$ is $\mathbf{0}$ and the conditional covariance matrix is $\tau_1^2 \left(\boldsymbol{I}_{RG \times RG} - \boldsymbol{C}\boldsymbol{D}^{-1}\boldsymbol{C}' \right)$ where $\boldsymbol{I}_{RG \times RG}$ is the $RG \times RG$ identity matrix and where

$$\boldsymbol{C} = \begin{array}{c} \\ \tilde{Z}_{11} \\ \tilde{Z}_{12} \\ \vdots \\ \tilde{Z}_{1(G-1)} \\ \tilde{Z}_{1G} \\ \vdots \\ \tilde{Z}_{R1} \\ \tilde{Z}_{R2} \\ \vdots \\ \tilde{Z}_{R(G-1)} \\ \tilde{Z}_{RG} \end{array} \begin{array}{c} \begin{array}{cccc} S_{1\cdot} & S_{2\cdot} & \ldots & S_{R\cdot} \end{array} \\ \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & & \vdots & \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \end{pmatrix} \end{array} \begin{array}{c} \begin{array}{cccc} S_{\cdot 1} & S_{\cdot 2} & \ldots & S_{\cdot(G-1)} \end{array} \\ \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ & \vdots & & \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 1 \end{pmatrix} \end{array}$$

and

$$\boldsymbol{D} = \begin{array}{c} \\ S_{1\cdot} \\ S_{2\cdot} \\ \vdots \\ S_{R\cdot} \\ S_{\cdot 1} \\ S_{\cdot 2} \\ \vdots \\ S_{\cdot(G-1)} \end{array} \begin{array}{c} \begin{array}{cccc} S_{1\cdot} & S_{2\cdot} & \ldots & S_{R\cdot} \end{array} \\ \begin{pmatrix} G & 0 & \cdots & 0 \\ 0 & G & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & G \\ 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{pmatrix} \end{array} \begin{array}{c} \begin{array}{cccc} S_{\cdot 1} & S_{\cdot 2} & \ldots & S_{\cdot(G-1)} \end{array} \\ \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \\ R & 0 & \cdots & 0 \\ 0 & R & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & R \end{pmatrix} \end{array}.$$

It is straightforward to show, using the standard formula for the inverse of a partitioned matrix (Gentle, 2007, Section 3.4), that

$$\boldsymbol{D}^{-1} = \begin{array}{c} \\ S_{1\cdot} \\ S_{2\cdot} \\ \vdots \\ S_{R\cdot} \\ S_{\cdot 1} \\ S_{\cdot 2} \\ \vdots \\ S_{\cdot(G-1)} \end{array} \begin{array}{c} \begin{array}{cccc} S_{1\cdot} & S_{2\cdot} & \ldots & S_{R\cdot} \end{array} \\ \begin{pmatrix} d_{11}^{(0)} & d_{11}^{(1)} & \cdots & d_{11}^{(1)} \\ d_{11}^{(1)} & d_{11}^{(0)} & \cdots & d_{11}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ d_{11}^{(1)} & d_{11}^{(1)} & \cdots & d_{11}^{(0)} \\ d_{12} & d_{12} & \cdots & d_{12} \\ d_{12} & d_{12} & \cdots & d_{12} \\ \vdots & \vdots & \ddots & \vdots \\ d_{12} & d_{12} & \cdots & d_{12} \end{pmatrix} \end{array} \begin{array}{c} \begin{array}{cccc} S_{\cdot 1} & S_{\cdot 2} & \ldots & S_{\cdot(G-1)} \end{array} \\ \begin{pmatrix} d_{12} & d_{12} & \cdots & d_{12} \\ d_{12} & d_{12} & \cdots & d_{12} \\ \vdots & \vdots & \ddots & \vdots \\ d_{12} & d_{12} & \cdots & d_{12} \\ d_{22}^{(0)} & d_{22}^{(1)} & \cdots & d_{22}^{(1)} \\ d_{22}^{(1)} & d_{22}^{(0)} & \cdots & d_{22}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ d_{22}^{(1)} & d_{22}^{(1)} & \cdots & d_{22}^{(0)} \end{pmatrix} \end{array},$$

where $d_{11}^{(0)} = (R + G - 1)/RG$, $d_{11}^{(1)} = (G - 1)/RG$, $d_{12} = -1/R$, $d_{22}^{(0)} = 2/R$ and $d_{22}^{(1)} = 1/R$.

Writing $\mathbf{1}$ for the $RG \times 1$ vector of ones, the matrix $\boldsymbol{C}\boldsymbol{D}^{-1}\boldsymbol{C}'$ is now given by

$$
2d_{12}\mathbf{1}\mathbf{1}' + \begin{array}{c} \\ \tilde{Z}_{11} \\ \tilde{Z}_{12} \\ \vdots \\ \tilde{Z}_{1G} \\ \vdots \\ \tilde{Z}_{R1} \\ \tilde{Z}_{R2} \\ \vdots \\ \tilde{Z}_{RG} \end{array} \begin{array}{c} \begin{array}{cccc} \tilde{Z}_{11} & \tilde{Z}_{12} & \cdots & \tilde{Z}_{1G} \end{array} \cdots \begin{array}{cccc} \tilde{Z}_{R1} & \tilde{Z}_{R2} & \cdots & \tilde{Z}_{RG} \end{array} \\ \begin{pmatrix} d_{11}^{(0)}+d_{22}^{(0)} & d_{11}^{(0)}+d_{22}^{(1)} & \cdots & d_{11}^{(0)}+d_{22}^{(1)} & \cdots & d_{11}^{(1)}+d_{22}^{(0)} & d_{11}^{(1)}+d_{22}^{(1)} & \cdots & d_{11}^{(1)}+d_{22}^{(1)} \\ d_{11}^{(0)}+d_{22}^{(1)} & d_{11}^{(0)}+d_{22}^{(0)} & \cdots & d_{11}^{(0)}+d_{22}^{(1)} & \cdots & d_{11}^{(1)}+d_{22}^{(1)} & d_{11}^{(1)}+d_{22}^{(0)} & \cdots & d_{11}^{(1)}+d_{22}^{(1)} \\ \vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ d_{11}^{(0)}+d_{22}^{(1)} & d_{11}^{(0)}+d_{22}^{(1)} & \cdots & d_{11}^{(0)}+d_{22}^{(0)} & \cdots & d_{11}^{(1)}+d_{22}^{(1)} & d_{11}^{(1)}+d_{22}^{(1)} & \cdots & d_{11}^{(1)}+d_{22}^{(0)} \\ \vdots & \vdots & & \vdots & \ddots & \vdots & \vdots & & \vdots \\ d_{11}^{(1)}+d_{22}^{(0)} & d_{11}^{(1)}+d_{22}^{(1)} & \cdots & d_{11}^{(1)}+d_{22}^{(1)} & \cdots & d_{11}^{(0)}+d_{22}^{(0)} & d_{11}^{(0)}+d_{22}^{(1)} & \cdots & d_{11}^{(0)}+d_{22}^{(1)} \\ d_{11}^{(1)}+d_{22}^{(1)} & d_{11}^{(1)}+d_{22}^{(0)} & \cdots & d_{11}^{(1)}+d_{22}^{(1)} & \cdots & d_{11}^{(0)}+d_{22}^{(1)} & d_{11}^{(0)}+d_{22}^{(0)} & \cdots & d_{11}^{(0)}+d_{22}^{(1)} \\ \vdots & \vdots & \ddots & \vdots & & \vdots & \vdots & \ddots & \vdots \\ d_{11}^{(1)}+d_{22}^{(1)} & d_{11}^{(1)}+d_{22}^{(1)} & \cdots & d_{11}^{(1)}+d_{22}^{(0)} & \cdots & d_{11}^{(0)}+d_{22}^{(1)} & d_{11}^{(0)}+d_{22}^{(1)} & \cdots & d_{11}^{(0)}+d_{22}^{(0)} \end{pmatrix} \end{array},
$$

the elements of which are either $2d_{12} + d_{11}^{(0)} + d_{22}^{(0)} = (R + G - 1)/RG$, $2d_{12} + d_{11}^{(0)} + d_{22}^{(1)} = (R-1)/RG$, $2d_{12}+d_{11}^{(1)}+d_{22}^{(0)} = (G-1)/RG$ or $2d_{12}+d_{11}^{(1)}+d_{22}^{(1)} = -1/RG$. The corresponding elements of the required conditional covariance matrix are then $\tau_1^2(R-1)(G-1)/RG$, $-\tau_1^2(R-1)/RG$, $-\tau_1^2(G-1)/RG$ and τ_1^2/RG. ■

Bibliography

Allen, M. R., Stott, P. A., Mitchell, J. F. B., Schnur, R., and Delworth, T. L. (2000). Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature*, 407:617–620. doi:10.1038/35036559.

Arnell, N. W., Kay, A. L., Freeman, A., Rudd, A. C., and Lowe, J. A. (2021). Changing climate risk in the UK: a multi-sectoral analysis using policy-relevant indicators. *Climate Risk Management*, 31:100265.

Barnes, C. R., Chandler, R. E., and Brierley, C. M. (2022a). Comparison of EuroCORDEX output with UKCP18 regional ensemble. Technical report, UK Climate Resilience Programme project CR20-3 'Enabling the use and producing improved understanding of EuroCORDEX data over the UK'. Available from https://www.ucl.ac.uk/statistics/research/ukcordex.

Barnes, C. R., Brierley, C. M., and Chandler, R. E. (2022b). A comparison of regional climate projections with a range of climate sensitivities. *Under review*.

Bates, D., Maechler, M., and Jagan, M. (2023). *Matrix: Sparse and Dense Matrix Classes and Methods*. R package version 1.5-4.

Bates, J. (2007). Some considerations of the concept of climate feedback. *Q. J. R. Meteorol. Soc.*, 133:545–560. doi:10.1002/qj.62.

Bowman, A. W. (2019). Graphics for uncertainty. *Appl. Statist.*, 182(2):403–418. doi: 10.1111/rssa.12379.

Brunner, L., McSweeney, C., Ballinger, A. P., Befort, D. J., Benassi, M., Booth, B., Coppola, E., de Vries, H., Harris, G., Hegerl, G. C., Knutti, R., Lenderink, G., Lowe, J., Nogherotto, R., O'Reilly, C., Qasmi, S., Ribes, A., Stocchi, P., and Undorf, S. (2020). Comparing methods to constrain future European climate projections using a consistent framework. *Journal of Climate*, 33(20):8671–8692. doi: 10.1175/JCLI-D-19-0953.1.

Buser, C. M., Künsch, H. R., Lüthi, D., Wild, M., and Schär, C. (2009). Bayesian multi-model projection of climate: bias assumptions and interannual variability. *Climate Dynamics*, 33:849–868.

Buser, C. M., Künsch, H. R., and Weber, A. (2010). Biases and uncertainty in climate projections. *Scandinavian Journal of Statistics*, 37:179–199.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Chandler, R. E. (2013). Exploiting strength, discounting weakness: combining information from multiple climate simulators. *Phil. Trans. Roy. Soc. A*, 371:20120388. doi:10.1098/rsta.2012.0388.

Chandler, R. E. (2023). *TimSPEC: Time Series postprocessing of ensembles for climate*. R package version 0.0-3. Available from `https://github.com/Richard-Chandler/TimSPEC`.

Chandler, R. E. and Scott, E. M. (2011). *Statistical methods for trend detection and analysis in the environmental sciences*. John Wiley & Sons, Chichester.

Chen, D., Rojas, M., Samset, B., Cobb, K., Niang, A. D., Edwards, P., Emori, S., Faria, S., Hawkins, E., Hope, P., Huybrechts, P., Meinshausen, M., Mustafa, S., Plattner, G.-K., and

TrÃ©guier, A.-M. (2021). Framing, context, and methods. In Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J., Maycock, T., Waterfield, T., Yelekçi, O., Yu, R., and Zhou, B., editors, *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, pages 147–286. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. doi:10.1017/9781009157896.003.

Cox, D. R. (2006). *Principles of Statistical Inference*. Cambridge University Press, Cambridge.

Cox, P. M., Huntingford, C., and Williamson, M. S. (2018). Emergent constraint on equilibrium climate sensitivity from global temperature variability. *Nature*, 553:319–322. doi: 10.1038/nature25450.

Davison, A. C. (2003). *Statistical Models*. Cambridge University Press, Cambridge.

Durbin, J. and Koopman, S. J. (2002). A simple and efficient simulation smoother for state space time series analysis. *Biometrika*, 89(3):603–615.

Durbin, J. and Koopman, S. J. (2012). *Time series analysis by state space methods (second edition)*. Oxford University Press, Oxford.

Economou, T., Stephenson, D. B., Rougier, J. C., Neal, R. A., and Mylne, K. R. (2016). On the use of Bayesian decision theory for issuing natural hazard warnings. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 472(2194):20160295. doi: 10.1098/rspa.2016.0295.

Frühwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *J. Time Series Analysis*, 15(2):183–202.

Furrer, R., Sain, S. R., Nychka, D., and Meehl, G. A. (2007). Multivariate Bayesian analysis of atmosphere-ocean general circulation models. *Environ. Ecol. Stat.*, 14:249–266. doi:10.1007/s10651-007-0018-z.

Gentle, J. E. (2007). *Matrix Algebra: Theory, Computations, and Applications in Statistics*. Springer, New York.

Geoffroy, O., Saint-Martin, D., Olivié, D., Voldoire, A., Bellon, G., and Tytéca, S. (2013). Transient climate response in a two-layer energy-balance model. Part I: Analytical solution

and parameter calibration using CMIP5 AOGCM experiments. *J. Climate*, 26:1841–1857. doi:10.1175/JCLI-D-12-00195.1.

Goldstein, M. and Rougier, J. (2005). Probabilistic formulations for transferring inferences from mathematical models to physical systems. *SIAM J. Sci. Comput.*, 26(2):467–487. doi: 10.1137/S106482750342670X.

Greene, A. M., Goddard, L., and Lall, U. (2006). Probabilistic multimodel regional temperature change projections. *J. Climate*, 19:4326–4343.

Hansen, J., Johnson, D., Lacis, A., Lebedeff, S., Lee, P., Rind, D., and Russell, G. (1981). Climate impact of increasing atmospheric carbon dioxide. *Science*, 213:957–966.

Hegerl, G. C., Ballinger, A. P., Booth, B. B. B., Borchert, L. F., Brunner, L., Donat, M. G., Doblas-Reyes, F. J., Harris, G. R., Lowe, J., Mahmood, R., Murphy, J. M. J. M., Swingedouw, D., and Weisheimer, A. (2021). Toward consistent observational constraints in climate predictions and projections. *Frontiers in Climate*, 3:678109. doi:10.3389/fclim.2021.678109.

Jackson, L. S., Maycock, A. C., Andrews, T., Fredriksen, H.-B., Smith, C. J., and Forster, P. M. (2022). Errors in simple climate model emulations of past and future global temperature change. *Geophysical Research Letters*, 49:e2022GL098808. doi: 10.1029/2022GL098808.

Jacob, D., Petersen, J., Eggert, B., Alias, A., Christensen, O. B., Bouwer, L. M., Braun, A., Colette, A., Déqué, M., Georgievski, G., Georgopoulou, E., Gobiet, A., Menut, L., Nikulin, G., Haensler, A., Hempelmann, N., Jones, C., Keuler, K., Kovats, S., Kröner, N., Kotlarski, S., Kriegsmann, A., Martin, E., van Meijgaard, E., Moseley, C., Pfeifer, S., Preuschmann, S., Radermacher, C., Radtke, K., Rechid, D., Rounsevell, M., Samuelsson, P., Somot, S., Soussana, J.-F., Teichmann, C., Valentini, R., Vautard, R., Weber, B., and Yiou, P. (2014). EURO-CORDEX: new high-resolution climate change projections for European impact research. *Reg. Environ. Change*, 14:563–578. doi:10.1007/s10113-013-0499-2.

Kang, E. L., Cressie, N., and Sain, S. R. (2012). Combining outputs from the North American Regional Climate Change Assessment Program by using a Bayesian hierarchical model. *J. Roy. Stat. Soc., Series C*, 61(2):291–313.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795. doi: 10.1080/01621459.1995.10476572.

Kennedy-Asser, A. T., Andrews, O., Mitchell, D. M., and Warren, R. F. (2021). Evaluating heat extremes in the UK Climate Projections (UKCP18). *Environmental Research Letters*, 16(1):014039.

Kettleborough, J. A., Booth, B. B. B., Stott, P. A., and Allen, M. R. (2007). Estimates of uncertainty in predictions of global mean surface temperature. *Journal of Climate*, 20(5):843–855. doi:10.1175/JCLI4012.1.

Kjellström, E. and Giorgi, F. (2010). Introduction to special issue 'Regional Climate Model evaluation and weighting'. *Clim. Res.*, 44:117–119.

Knutti, R. (2010). The end of model democracy? *Climatic Change*, 102:395–404. doi:10.1007/s10584-010-9800-2.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A. (2010). Challenges in combining projections from multiple climate models. *J. Climate*, 23:2739–2758. doi:10.1175/2009JCLI3361.1.

Krzanowski, W. (1988). *Principles of Multivariate Analysis*. Oxford University Press.

Leith, N. A. and Chandler, R. E. (2010). A framework for interpreting climate model outputs. *J. Roy. Stat. Soc., Series C*, 59(2):279–296.

Monahan, J. F. (2001). *Numerical Methods of Statistics*. Cambridge University Press, Cambridge.

Murphy, J. M., Harris, G. R., Sexton, D. M. H., Kendon, E. J., Bett, P. E., Clark, R. T., and Yamazaki, K. (2019). UKCP18 land projections: Science report. Technical report, Met Office. Available from https://www.metoffice.gov.uk/pub/data/weather/uk/ukcp18/science-reports/UKCP18-Land-report.pdf.

Myhre, G., Shindell, D., Bréon, F.-M., Collins, W., Fuglestvedt, J., Huang, J., Koch, D., Lamarque, J.-F., Lee, D., Mendoza, B., Nakajima, T., Robock, A., Stephens, G., Takemura, T., and Zhang, H. (2013). Anthropogenic and natural radiative forcing supplementary material. In T.F. Stocker, T., Qin, D., Plattner, G.-K., Tignor, M., Allen, S., Boschung, J., Nauels, A., Xia, Y., Bex, V., and Midgley, P., editors, *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Available from www.ipcc.ch.

Nicholls, Z., Meinshausen, M., Lewis, J., Smith, C. J., Forster, P., Fuglestvedt, J. S., Rogelj, J., Kikstra, J. S., Riahi, K., and Byers, E. (2022). Changes in IPCC scenario assessment emulators between SR1.5 and AR6 unraveled. *Geophysical Research Letters*, 49:e2022GL099788. doi: 10.1029/2022GL099788.

Ollila, E., Palomar, D., and Pascal, F. (2021). Shrinking the eigenvalues of m-estimators of covariance matrix. *IEEE Transactions on Signal Processing*, 69:256–269. doi: 10.1109/TSP.2020.3043952.

Perry, M., Hollis, D., and Elms, M. (2009). The generation of daily gridded datasets of temperature and rainfall for the UK. Technical report, National Climate Information Centre, Met Office, Exeter.

Petris, G. (2010). An R package for dynamic linear models. *Journal of Statistical Software*, 36(12):1–16.

Qasmi, S. and Ribes, A. (2022). Reducing uncertainty in local temperature projections. *Science Advances*, 8:eabo6872. doi: 10.1126/sciadv.abo6872.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reis, E. A., Salazar, E., and Gamerman, D. (2006). Comparison of sampling schemes for dynamic linear models. *International Statistical Review*, 74(2):203–214.

Robert, C. and Casella, G. (2011). A short history of Markov Chain Monte Carlo: Subjective recollections from incomplete data. *Statistical Science*, 26(1):102–115. doi: 10.1214/10-STS351.

Rougier, J., Goldstein, M., and House, L. (2013). Second-order exchangeability analysis for multimodel ensembles. *J. Amer. Statist. Assoc.*, 108(503):852–863. doi: 10.1080/01621459.2013.802963.

Rougier, J. and Sexton, D. M. (2007). Inference in ensemble experiments. *Phil. Trans. Roy. Soc. A: Mathematical, Physical and Engineering Sciences*, 365(1857):2133–2143. doi: 10.1098/rsta.2007.2071.

Sansom, P. G., Cummins, D., Siegert, S., and Stephenson, D. B. (2021). Towards reliable projections of global mean surface temperature.

Sexton, D. M. H., Murphy, J. M., Collins, M., and Webb, M. J. (2012). Multivariate probabilistic projections using imperfect climate models part I: Outline of methodology. *Climate Dynamics*, 38:2513–2542. doi: 10.1007/s00382-011-1208-9.

Shiogama, H., Stone, D., Emori, S., Takahashi, K., Mori, S., Maeda, A., Ishizaki, Y., and Allen, M. R. (2016). Predicting future uncertainty constraints on global warming projections. *Scientific Reports*, 6:18903.

Smith, C. J., Kramer, R. J., Myhre, G., Alterskjær, K., Collins, W., Sima, A., Boucher, O., Dufresne, J.-L., Nabat, P., Michou, M., Yukimoto, S., Cole, J., Paynter, D., Shiogama, H., O'Connor, F. M., Robertson, E., Wiltshire, A., Andrews, T., Hannay, C., Miller, R., Nazarenko, L., Kirkevåg, A., Olivié, D., Fiedler, S., Lewinschal, A., Mackallah, C., Dix, M., Pincus, R., and Forster, P. M. (2020). Effective radiative forcing and adjustments in CMIP6 models. *Atmospheric Chemistry and Physics*, 20(16):9591–9618.

Smith, R. L., Tebaldi, C., Nychka, D., and Mearns, L. O. (2009). Bayesian modeling of uncertainty in ensembles of climate models. *J. Amer. Statist. Assoc.*, 104 (485):97–116. doi:10.1198/jasa.2009.0007.

Stainforth, D. A., Allen, M. R., Tredger, E. R., and Smith, L. A. (2007). Confidence, uncertainty and decision-support relevance in climate predictions. *Phil. Trans. R. Soc. Lond.*, A365:2145–2161. doi: 10.1098/rsta.2007.2074.

Stott, P. and Kettleborough, J. (2002). Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature*, 416:723–726. doi: 10.1038/416723a.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. *Bulletin of the American meteorological Society*, 93(4):485–498.

Tebaldi, C. and Sansó, B. (2009). Joint projections of temperature and precipitation change from multiple climate models: a hierarchical Bayesian approach. *J. Roy. Stat. Soc., Series A*, 172:83–106.

Tebaldi, C., Smith, R. L., Nychka, D., and Mearns, L. O. (2005). Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multi-model ensembles. *Journal of Climate*, 18:1524–1540.

Tucker, S. O., Kendon, E. J., Bellouin, N., Buonomo, E., Johnson, B., and Murphy, J. M. (2021). Evaluation of a new 12 km regional perturbed parameter ensemble over Europe. *Climate Dynamics*, pages 1–25.

von Storch, H. and Zwiers, F. W. (1999). *Statistical analysis in climate research*. Cambridge University Press, Cambridge.

Williamson, D. B. and Sansom, P. G. (2019). How are emergent constraints quantifying uncertainty and what do they leave behind? *Bull. Amer. Meteor. Soc.*, 100:2571–2588. doi: 10.1175/BAMS-D-19-0131.1.

Wood, S. N. (2006). *Generalized Additive Models: an Introduction with R*. Chapman & Hall / CRC, Boca Raton.