Characterising spatial structure in climate model ensembles

Richard E. Chandler, a  Clair R. Barnes, a  and Chris M. Brierley b

a Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK
b Department of Geography, University College London, Gower Street, London WC1E 6BT

Corresponding author: Richard Chandler, r.chandler@ucl.ac.uk
ABSTRACT: This paper presents a methodology that is designed for rapid exploratory analysis of the outputs from ensembles of climate models, especially when these outputs consist of maps. The approach formalises and extends the technique of ‘inter-model empirical orthogonal function’ analysis, combining multivariate analysis of variance techniques with singular value decompositions (SVDs) of structured components of the ensemble data matrix. The SVDs yield spatial patterns associated with these components, that we call ensemble principal patterns (EPPs). A unique hierarchical partitioning of variation is obtained for balanced ensembles in which all combinations of factors, such as GCM:RCM pairs in a regional ensemble, appear with equal frequency: suggestions are also proposed to handle unbalanced ensembles without imputing missing values or discarding runs. Applications include the selection of ensemble members to propagate uncertainty into subsequent analyses; and diagnosis of modes of variation associated with specific model variants or parameter perturbations. The approach is illustrated using outputs from the EuroCORDEX regional ensemble over the UK.
1. Introduction

Much of our understanding of the climate system is derived from climate model experiments, which also underpin most projections of future climate based on specified scenarios of greenhouse gas emissions and socioeconomic development (Chen et al. 2021). Climate models may be used individually, as when using general circulation models (GCMs) or earth system models (ESMs) to provide global simulations at a relatively coarse spatial resolution; or in combination, as when using GCMs or ESMs to provide boundary conditions for high-resolution regional climate model (RCM) simulations.

Climate model experiments are associated with multiple sources of uncertainty (Rougier and Goldstein 2014), which are often studied using collections (‘ensembles’) of model runs. These include ensembles that explore internal variability, often by varying initial conditions in a single model (e.g. Rodgers et al. 2021; Maher et al. 2021; von Trentini et al. 2020); perturbed physics ensembles (Collins et al. 2006) that vary the representation of key processes, again within a single model; and multimodel ensembles (Meehl et al. 2007; von Trentini et al. 2019) that can be regarded as attempting to characterise the range of outcomes that could be anticipated on the basis of current scientific knowledge, as represented by the models considered.

In any such ensemble, the sampled uncertainties induce variation between ensemble members. It is therefore of interest to characterise this variation: for example, by identifying sources of uncertainty to which the outputs are most sensitive and hence suggesting priorities for future model improvements; or by identifying representative subsets of members for use in applications where it is not feasible to process the entire ensemble. The latter situation can arise when climate projections are used as inputs to other models of climate change impacts: if the impacts models themselves are computationally expensive or if resources are otherwise limited, then this limits the number of ensemble members that can be used (Cannon 2015).

A substantial body of literature is devoted to quantifying the relative importance of different sources of variation within an ensemble. For example, Hawkins and Sutton (2009) applied a heuristic approach to projections from the World Climate Research Programme’s Coupled Model Intercomparison Project phase 3 (CMIP3) multimodel ensemble, to partition the variation into components attributable respectively to GCMs, emissions scenarios and internal variability of the climate system. This approach was placed on a more formal footing using Analysis of Variance
(ANOVA) techniques by Yip et al. (2011). A drawback of these authors’ fixed-effects ANOVA methodology, however, is that the unambiguous partitioning of variation requires a balanced ensemble which, in this case, means that each GCM is run the same number of times with each emissions scenario. The CMIP3 ensemble does not meet this condition, so Yip et al. (2011) discarded some ensemble members and applied their methodology to the largest possible balanced subset. Of course, discarding ensemble members incurs a loss of information: this can be avoided using a random-effects ANOVA (Northrop and Chandler 2014); or by considering the problem as one of a balanced ensemble with some values missing, and estimating (‘imputing’) these missing values (Evin et al. 2019, 2021). The latter approaches both use Bayesian methods and require the use of Markov Chain Monte Carlo (MCMC) techniques that are computationally intensive. This perhaps disincentivises their routine use in situations where the outputs of interest are high-dimensional (i.e. involve many quantities of interest), for example when analysing maps involving large numbers of spatial locations.

In high-dimensional settings, a further potential concern with ANOVA approaches is that they are designed to analyse scalar-valued quantities. Most applications of ANOVA in climate science have therefore considered each quantity separately: for example, Christensen and Kjellström (2020) applied fixed-effects ANOVAs to a subset of the EuroCORDEX regional model ensemble (Jacob et al. 2014), separately for each spatial location within the EuroCORDEX study region, and mapped the results. However, a comprehensive analysis of variation in high-dimensional ensemble outputs requires methods that are specifically designed for a high-dimensional setting. This has rarely been attempted to date, a notable exception being Sain et al. (2011) who used functional ANOVA methods to analyse ensembles of maps within a Bayesian framework which, as noted above, can be computationally expensive. Other exceptions include applications of clustering methods to find representative subsets of ensemble members for use in impacts studies (e.g. Cannon 2015; Casajus et al. 2016): these methods aim to produce a simplified representation of the variation by clustering the ensemble members into groups that are relatively homogeneous and distinct.

Alternative simplified representations can be obtained via dimension reduction, for example by applying principal components analysis (PCA) or empirical orthogonal function (EOF) analysis across ensemble members (e.g. Li and Xie 2012; Yim et al. 2016; Wang et al. 2020). Such ‘inter-model EOF’ analyses seek to identify dominant modes of spatial variation across the ensemble:
the cited references demonstrate how these modes may be of interest in their own right, as well as how the associated scores (see below) can be used to identify contrasting models for further study. A potential difficulty, however, is that ensembles often contain structure that EOF analyses are not designed to handle. In regional ensembles for example, structure is induced by the GCM:RCM combinations used to generate each member. Similarly, in the CMIP ensembles, direct EOF analyses will be dominated by the models with the most runs. To avoid this problem when analysing CMIP5 outputs, Yim et al. (2016) worked with the average of each model’s runs, while Wang et al. (2020) analysed a single run from each model. However, no formal justification was provided for either approach.

Against this background, the main contribution of the present paper is to formalise the application of dimension reduction techniques to climate model ensembles, while also accounting appropriately for ensemble structure. The resulting approach, which we call ensemble principal pattern (EPP) analysis, reduces to inter-model EOF analysis when applied to unstructured ensembles containing a single member per model. It is intended primarily as a tool to enable the rapid exploration of ensembles, either to select contrasting or representative members for use in impacts studies or to highlight distinctive modes of variation that may merit further investigation.

To motivate our proposal and fix ideas subsequently, Section 2 presents some outputs from the EuroCORDEX regional ensemble over the United Kingdom, and discusses some questions that could be asked about these outputs. The basic inter-model EOF methodology is described in Section 3 as a reference. In Section 4 the methodology is extended, in conjunction with multivariate ANOVA (MANOVA) techniques, to handle ensembles with more complex structures: the extension is illustrated using the EuroCORDEX example. Section 5 concludes and suggests further potential applications.

2. Motivating example: the EuroCORDEX regional ensemble

We consider the bias in simulated summer (JJA) mean daily maximum surface temperature (‘\texttt{tasmax}’) across the United Kingdom, over the period 1989–2008, for 64 EuroCORDEX ensemble members (Jacob et al. 2014) that were produced using combinations of ten RCMs and ten GCM runs from the CMIP5 experiment (Taylor et al. 2012). The GCM runs, from six unique GCMs, were conditioned on CMIP5 historical forcings until 2005, and on the RCP8.5 emissions scenario
from 2006 to 2008: over this limited period the RCP8.5 forcings deviate very little from the the 2005 levels (Van Vuuren et al. 2011) and hence can be considered as plausible proxies for the historical values.

To calculate the biases in JJA \( \text{tasmax} \), the 1989–2008 mean \( \text{tasmax} \) values were first calculated for each ensemble member on the native grid of the corresponding RCM. Next, these mean values were regridded to the 12 × 12km\(^2\) grid used by the HadUK gridded observational dataset (Perry et al. 2009), using a conservative area-weighting scheme (Jones 1999): the 12km grid resolution is similar to that of each EuroCORDEX RCM. Finally the biases for each ensemble member, shown in Figure 1, were computed by subtracting the 1989–2008 mean HadUK \( \text{tasmax} \) values.

The columns of Figure 1 represent GCM runs, and the rows represent RCMs. To simplify the subsequent presentation, the ten GCM runs will be considered as representing separate models throughout: this is not required for the methodology, however. With this convention, each GCM:RCM pair contributes at most a single member to the ensemble. The ensemble is unbalanced however, with 36 of the 100 possible pairs missing. The missing pairs include four that are available but have been excluded deliberately: 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.

As noted in Section 1, lack of balance causes problems when attributing variation to its potential sources, whence analyses are sometimes restricted to the largest balanced subset of an ensemble. Here, the largest such subset involves eight RCMs and four GCM runs, hence including just 32 of the 64 available ensemble members. It is clearly undesirable to restrict attention to such a small subset.

Apart from the ‘missing’ runs, the most obvious feature of Figure 1 is the sixth column, corresponding to the HadGEM2-ES GCM: all ensemble members driven by this GCM appear to have a warm bias. Closer inspection also suggests potential cool biases associated with the HIRHAM5, RACMO22E and RCA4 RCMs (fourth, sixth and seventh rows): however, it is hard to be confident about this based on a purely visual inspection. Moreover, it is hard to identify any systematic and detailed spatial structure from the maps. This creates difficulties for several potential users of the ensemble. Consider, for, example, a regional climate modeller who sees EuroCORDEX as an opportunity to explore the simulation of summer temperatures by different RCMs: apart from
Fig. 1. Bias in mean simulated surface temperature (°C) over the UK, for the period 1989–2008, from 64 members of the EuroCORDEX regional ensemble (see main text for details). Rows and columns correspond to RCMs and GCMs respectively.
the apparent cool biases noted above, Figure 1 does not obviously provide useful information to support such a comparison.

Another potential user is the climate impacts modeller with limited computational resources. Such a user might produce the equivalent of Figure 1, showing future changes in some impact-relevant quantity under a specified emissions scenario; and they may wish to choose a small number of ensemble members spanning the range of potential impacts. This could be done by computing spatial averages of relevant quantities for each ensemble member, and then choosing members spanning the range of the resulting averages. However, spatial averages are not necessarily the most relevant quantities when considering impacts that may be localised in space (e.g. associated with urban areas): it would therefore be helpful to find an alternative visualisation that allows users to identify detailed spatial structures in an ensemble.

Our two notional EuroCORDEX ‘users’ have different interests and priorities. Nonetheless, both would benefit from a visualisation that simultaneously (i) is more compact than Figure 1 and (ii) allows them to identify potentially interesting modes of variation. This is the aim of our EPP methodology — although, as discussed later, its use is not restricted to regional ensembles.

3. EPP analysis: the simplest case

To develop the methodology, we start by considering an unstructured ensemble — for example, a multimodel ensemble with a single member per model, or a perturbed physics ensemble with one member per model variant. The ensemble outputs are considered as a collection of vectors \( \{Y_i : i = 1, \ldots, n\} \) say, where \( n \) is the number of members and \( Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iS})' \) is a vector of length \( S \) containing a value for the \( i \)th member at each of \( S \) spatial locations. Moreover, let \( Y \) be the \( n \times S \) ‘ensemble data matrix’ with \( Y_i \) as its \( i \)th row; let \( \bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i \) denote the \( S \times 1 \) overall ensemble mean vector, with \( s \)th element \( \bar{Y}_s = n^{-1} \sum_{i=1}^{n} Y_{is} \); and let \( 1 \) be an \( S \times 1 \) vector of ones.

In this case, as noted earlier, the proposed methodology is equivalent to an inter-model EOF analysis which, for present purposes, is most conveniently presented by considering the Singular Value Decomposition (SVD) of the centred data matrix \( Y - \bar{Y} \bar{Y}' = \bar{Y} \) say, with \( Y_i - \bar{Y} \) as its \( i \)th row. Since the approach is widely used already (see Section 1), we merely sketch it here to prepare for the extension to structured ensembles in Section 4. For more details of the underpinning mathematics, see Krzanowski (1988, Section 4.1) or Gentle (2007, Section 3.10).
When \( n < S \), the centred matrix \( \tilde{Y} \) typically has rank \( n - 1 \) so that \( n \) separate maps are needed to visualise the complete ensemble (the ensemble mean, plus \( n - 1 \) maps to visualise the centred matrix — or, equivalently, separate maps of each ensemble member). Dimension reduction aims to find an accurate approximation of \( \tilde{Y} \) of lower rank, say \( d < n - 1 \), so that the information can be visualised in just \( d + 1 \) maps including the ensemble mean.

The SVD of the centred data matrix takes the form

\[
\tilde{Y} = U \Lambda V', \quad \text{say,}
\]

where \( U \) and \( V \) are orthogonal matrices with dimensions \( n \times n \) and \( S \times n \) respectively, and \( \Lambda \) is a \( n \times n \) diagonal matrix with non-negative elements sorted in decreasing order. Denote these elements by \( \lambda_1 > \lambda_2 > \ldots > \lambda_n = 0 \) (\( \lambda_n \) is zero because the rank of \( \tilde{Y} \) is at most \( n - 1 \)); then (1) can be rewritten as

\[
\tilde{Y} = \lambda_1 u_1 v_1' + \lambda_2 u_2 v_2' + \ldots + \lambda_n u_n v_n',
\]

where \( u_j \) and \( v_j \) are vectors, of length \( n \) and \( S \), containing the \( j \)th columns of \( U \) and \( V \) respectively. Denoting the \( i \)th element of \( u_j \) by \( u_{ij} \), the \( i \)th row of \( \tilde{Y} \) is thus

\[
\lambda_1 u_{i1} v_1' + \lambda_2 u_{i2} v_2' + \ldots + \lambda_n u_{in} v_n',
\]

so that each row of \( \tilde{Y} \) is a weighted sum of the \( n \) vectors \( v_1', v_2', \ldots, v_n' \). These vectors are each of length \( S \) and, in the current context, represent spatial patterns: moreover, the patterns are uncorrelated because \( V \) is orthogonal. In fact, these patterns are the principal components of \( \tilde{Y} \) and the weights \( \{ \lambda_j u_{ij} : i = 1, \ldots, n; j = 1, \ldots, n \} \) are the corresponding principal component scores.

Now consider truncating the right-hand side of equation (2) to retain just \( d < n \) terms. The result is an \( n \times S \) matrix, \( \tilde{Y}^* \) say, of rank \( d \) and with \( i \)th row \( \lambda_1 u_{i1} v_1' + \lambda_2 u_{i2} v_2' + \ldots + \lambda_d u_{id} v_d' \). This matrix is a reduced-rank approximation to \( \tilde{Y} \): the corresponding approximation to the original uncentred data matrix is \( \tilde{Y}^* + \mathbf{1}\tilde{Y}' = Y^* \), say. In fact, denoting by \( Y_{is}^* \) the \((i, s)\)th element of \( Y^* \), the sum of squared approximation errors \( \sum_{i=1}^n \sum_{s=1}^S (Y_{is} - Y_{is}^*)^2 \) is smaller than under any other rank
approximation of $Y$; and the $j$th spatial pattern $v_j$ accounts for a proportion $\lambda_j^2 / \sum_{i=1}^n \lambda_i^2$ of the ‘total variation’ $\sum_{i=1}^n \sum_{s=1}^S \left( Y_{is} - \bar{Y}_s \right)^2$.

As noted above, the convention in PCA is to define the principal components as the columns of $V$ and to subsume the singular values $\{\lambda_j\}$ into the corresponding scores. In the current context however, it is arguably more interpretable to multiply the $\{\lambda_j\}$ by the $\{v_j\}$ instead: the resulting spatial patterns $\{\lambda_j v_j\}$ have the same units of measurement as the original quantity of interest and hence can be interpreted directly as contributions to the overall variation of that quantity, while the scores $\{u_{ij}\}$ are dimensionless and hence can be interpreted as ‘weights’ attached to each pattern. We therefore define the EPPs as the patterns $\{\lambda_j v_j\}$; and the $\{u_{ij} : i = 1, \ldots, n; j = 1, \ldots, d\}$ are the corresponding EPP scores. An ensemble member with a positive (resp. negative) score for the $j$th EPP will tend to have positive (resp. negative) deviations from the ensemble mean in regions where the corresponding pattern $v_j$ is positive, and vice versa.

A further detail is that $U$ and $V$ in (1) are not uniquely defined, because replacing $U$ and $V$ with $-U$ and $-V$ respectively does not change the result. We remove this ambiguity by specifying that positive EPP scores are associated overall with above-average values of the quantity of interest. Operationally, this is achieved by multiplying each EPP by the sign of its average value: if this average is negative, the multiplication reverses the signs of the EPP and the corresponding scores.

### 4. EPPs for structured ensembles

As noted in Section 1, many ensembles contain structure that is not accounted for by the basic methodology described above. We now extend the methodology to address this, by combining multivariate Analysis of Variance (MANOVA) techniques with SVDs. For ease of exposition, we focus on regional ensembles involving $R$ RCMs and $G$ GCMs, in which each RCM:GCM combination has been run at most once. Extensions to more general structured ensembles are straightforward in principle: details can be found in the Supplement to the paper.

#### a. Regional ensembles: theory

We denote the ensemble data matrix by $Y$ as before. Now however, the individual ensemble members are denoted by $\{Y_{rg}\}$, with $r$ and $g$ indexing RCMs and GCMs respectively. The total
The number of runs is denoted by \( n \), so that the dimension of \( Y \) is \( n \times S \); and \( n_r \) and \( n_g \) denote the numbers of runs involving RCM \( r \) and GCM \( g \), respectively.

The ensemble structure can be represented using a statistical model:

\[
Y_{rg} = \mu + \alpha_g + \beta_r + \varepsilon_{rg} \quad (g = 1, \ldots, G; r = 1, \ldots, R),
\]

where \( \sum_{g=1}^{G} \alpha_g = \sum_{r=1}^{R} \beta_r = 0 \). In (3), \( \mu \) represents an overall mean; \( \alpha_g \) and \( \beta_r \) represent systematic departures from this overall mean for the \( g \)th GCM and \( r \)th RCM respectively; and \( \varepsilon_{rg} \) is a ‘residual’ representing variation that cannot be attributed to systematic contributions from either the GCM or the RCM (for example, arising from internal variability in the system), treated as though it is drawn independently for each member from a distribution with mean vector \( 0 \) and covariance matrix \( \Sigma \).

With the exception of the \( S \times S \) matrix \( \Sigma \), these quantities are all vectors of length \( S \).

Equation (3) defines an additive MANOVA model, which is the multivariate analogue of the two-way ANOVA model used in several references from Section 1. Least-squares estimates of the coefficient vectors \( \{\alpha_g\} \) and \( \{\beta_r\} \) can be obtained from the data matrix: the estimation is particularly simple for a balanced ensemble with a run for every RCM:GCM combination so that \( n_r = G \), \( n_g = R \) and \( n_\cdot = RG \). Let \( \bar{Y}_r = n_r^{-1} \sum_{g=1}^{G} Y_{rg} \) and \( \bar{Y}_g = n_g^{-1} \sum_{r=1}^{R} Y_{rg} \) denote the mean vectors over all available runs for the \( r \)th RCM and \( g \)th GCM respectively; and let \( \bar{Y}_\cdot = n^{-1} \sum_{r=1}^{R} \sum_{g=1}^{G} Y_{rg} \) denote the overall ensemble mean. Then, following standard derivations e.g. Davison (2003, Section 9.2.2), in the balanced case the least-squares estimates of \( \mu \), \( \alpha_g \) and \( \beta_r \) are respectively \( \hat{\mu} = \bar{Y}_\cdot \), \( \hat{\alpha}_g = \bar{Y}_g - \bar{Y}_\cdot \) and \( \hat{\beta}_r = \bar{Y}_r - \bar{Y}_\cdot \). These are natural summary measures: for example, the estimated mean field for GCM \( g \) is \( \hat{\mu} + \hat{\alpha}_g = \bar{Y}_\cdot + \bar{Y}_g - \bar{Y}_\cdot = \bar{Y}_g \), with a corresponding expression for RCM \( r \). However, the appropriate analogues of these quantities are less obvious in unbalanced ensembles with some missing RCM:GCM combinations. In this case, model (3) provides a natural generalisation as we now elaborate.

Modern treatments of unbalanced (M)ANOVA models (e.g. Faraway 2014, Section 13.2) exploit their alternative representation as linear models with dummy covariates defining group membership.
Specifically, (3) can be written as

\[
Y = 1\mu' + X_G\alpha + X_R\beta + \varepsilon = \begin{pmatrix} 1 & X_G & X_R \end{pmatrix} \begin{pmatrix} \mu' \\ \alpha \\ \beta \end{pmatrix} + \varepsilon = X \begin{pmatrix} \mu' \\ \alpha \\ \beta \end{pmatrix} + \varepsilon, \text{ say, (4)}
\]

where, in addition to the vector \( \mu \) defined already, \( 1 \) is now a vector containing \( n \) ones; \( X_G \) is an \( n \times (G - 1) \) matrix in which the \( g \)th column contains 1s in the rows corresponding to runs in which the \( g \)th GCM was used, −1s in the rows where the \( G \)th GCM was used, and zeroes everywhere else; \( X_R \) is an \( n \times (R - 1) \) matrix in which the \( r \)th column contains 1s in the rows corresponding to runs in which the \( r \)th RCM was used, −1s in the rows where the \( R \)th RCM was used, and zeroes everywhere else; \( \alpha \) is the \( (G - 1) \times S \) matrix in which the rows are the coefficient vectors \( \{\alpha_g : g = 1, \ldots, G - 1\} \); \( \beta \) is the corresponding \( (R - 1) \times S \) matrix containing the coefficient vectors \( \{\beta_r : r = 1, \ldots, R - 1\} \); \( \varepsilon \) is an \( n \times S \) matrix in which the rows are the vectors \( \{\varepsilon_{rg}\} \); and the matrix \( X = (1 \ X_G \ X_R) \), of dimension \( n \times (R + G - 1) \), is obtained by placing its component parts side by side. The accompanying python script (see “Data Availability” statement) demonstrates these constructions.

Under formulation (4), the least-squares coefficient estimates satisfy the equation (Rao 1973)

\[
(X'X) \begin{pmatrix} \hat{\mu}' \\ \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = X'Y, \tag{5}
\]

each side of which is an \( (R + G - 1) \times S \) matrix. The cost of solving this system increases linearly in \( S \) so that the solution is feasible on modern computers, even for large spatial domains.

Computing \( \hat{\alpha}_G = -\sum_{g=1}^{G-1} \hat{\alpha}_g \) and \( \hat{\beta}_R = -\sum_{r=1}^{R-1} \hat{\beta}_r \) now yields a complete set of effect estimates for each GCM and RCM in the ensemble, satisfying the constraints \( \sum_{g=1}^{G} \hat{\alpha}_g = \sum_{r=1}^{R} \hat{\beta}_r = 0 \). These estimates can be mapped individually or, to avoid having to inspect \( G + R \) maps, subjected to their own SVD decompositions by stacking them into matrices of dimension \( G \times S \) and \( R \times S \) respectively. These matrices are both centred by construction. The SVD decompositions yield two sets of EPPs, summarising the dominant patterns of variation among the GCMs and RCMs respectively. Moreover, SVDs of the ‘residual’ matrix \( Y - 1\hat{\mu}' - X_G\hat{\alpha} - X_R\hat{\beta} \) (which is also
centred) define spatial modes of variation that are not systematically related to either the GCMs or RCMs. We denote this residual matrix by $e(2)$: the superscript is for notational consistency with the Supplement.

A full EPP analysis thus consists of two steps: first, calculation of the GCM, RCM and residual effects based on (3); and second, application of SVD to each set of effects to obtain the corresponding spatial modes of variation.

**b. Regional ensembles: partitioning of variation**

The development above provides low-dimensional estimates of RCM, GCM and residual effects in a regional ensemble. To fully understand the ensemble structure however, it is also necessary to quantify the relative importance of these effects. If, for example, the RCMs contribute only a small proportion of the total variation, then the RCM EPPs themselves are arguably of limited interest. We now address this issue. The approach can be regarded either as an extension of the ANOVA methodology of Yip et al. (2011), to handle unbalanced ensembles with multivariate outputs; or as a computationally cheap alternative to the functional ANOVA methodology of Sain et al. (2011).

1) **Balanced ensembles**

Initially, it is again helpful to consider a balanced ensemble with no missing runs. In this case, the output vector for the $(r, g)$ run can be written as

$$Y_{rg} = \bar{Y} - \left( \bar{Y}_g - \bar{Y} \right) + \left( \bar{Y}_r - \bar{Y} \right) + \left( \bar{Y}_{rg} - \bar{Y}_r - \bar{Y}_g + \bar{Y} \right).$$ 

(6)

Consider now the total *sum of squares and cross-products* (SSCP)

$$\sum_{r=1}^{R} \sum_{g=1}^{G} \left( Y_{rg} - \bar{Y} \right) \left( Y_{rg} - \bar{Y} \right)' = T,$$ 

(7)

which is a matrix of dimension $S \times S$ that plays the same role as the total sum of squares in a standard univariate ANOVA (Krzanowski 1988, Section 13.3). To aid interpretation, note that $T/(RG - 1)$ is the sample covariance matrix between all pairs of locations across the ensemble.
Using (6), a little algebra shows that the total SSCP can be decomposed as

\[ T = T_G + T_R + T_E, \]

say, (8)

where

\[ T_G = R \sum_{g=1}^{G} \left( \bar{Y}_g - \bar{Y}_\cdot \right) \left( \bar{Y}_g - \bar{Y}_\cdot \right)' \]

and

\[ T_R = G \sum_{r=1}^{R} \left( \bar{Y}_r - \bar{Y}_\cdot \right) \left( \bar{Y}_r - \bar{Y}_\cdot \right)' \]

are SSCP matrices associated with the GCMs and RCMs respectively; and

\[ T_E = \sum_{g=1}^{G} \left( Y_{rg} - \bar{Y}_g \bar{Y}_r - \bar{Y}_\cdot \right) \left( Y_{rg} - \bar{Y}_g \bar{Y}_r - \bar{Y}_\cdot \right)' \]

represents residual unstructured variation. The relative magnitudes of these three SSCP matrices can therefore be used to summarise the relative importance of the three sources of variation within the ensemble. Note that for the balanced case under consideration, \( T_G \) and \( T_R \) are formed respectively from the least-squares estimates of the GCM and RCM effects in model (3): \( T_G = R \sum_{g=1}^{G} \hat{\alpha}_g \hat{\alpha}'_g \) and \( T_R = G \sum_{r=1}^{R} \hat{\beta}_r \hat{\beta}'_r \). Note also that under (3), an unbiased estimator of the residual covariance matrix \( \Sigma \) is \( \hat{\Sigma} = T_E / (n_\cdot - R - G + 1) \):

the denominator here is the residual degrees of freedom after estimating an overall mean together with \( R - 1 \) independent RCM effects and \( G - 1 \) independent GCM effects. Moreover, in the balanced case with \( n_\cdot = RG \), \( n_\cdot - R - G + 1 \) is equal to \( (G - 1)(R - 1) \) so that \( T_E = (G - 1)(R - 1) \hat{\Sigma} \)

and the right-hand side of equation (8) can be rewritten as

\[ R \sum_{g=1}^{G} \hat{\alpha}_g \hat{\alpha}'_g + G \sum_{r=1}^{R} \hat{\beta}_r \hat{\beta}'_r + (G - 1)(R - 1) \hat{\Sigma}. \]

We will revisit this expression later, when considering unbalanced ensembles.

Unfortunately, if the number of locations \( S \) exceeds 1, there is no unique way to compare the magnitudes of \( T_G \), \( T_R \) and \( T_E \) in (8) (if \( S = 1 \), then each of these matrices is a single number and can be expressed as a proportion of the total variation \( T \)). One option is to produce maps showing the diagonal elements of \( T \), and the proportional contributions of \( T_G \), \( T_R \) and \( T_E \) to each of these diagonal elements: these are the “total variability partition” maps in, for example, Figure 5 of Christensen and Kjellström (2020). Importantly, the diagonal elements can all be calculated without having to compute the SSCP matrices themselves (see the Supplement for details), which is helpful because they are of dimension \( S \times S \) so that storage requirements can be excessive for large \( S \).

To supplement the maps just described, the diagonal elements of the matrices in (8) can be summed to obtain their traces. The trace operator is additive, so that \( \text{trace}(T) = \text{trace}(T_G) + \text{trace}(T_R) + \text{trace}(T_E) \) and the total trace is partitioned unambiguously into single-number sum-
maries representing contributions from each source of variation. A potential disadvantage is that the off-diagonal elements of the SSCP matrices do not influence the partitioning: we return to this point in Section 5. For the moment however, we highlight a useful alternative interpretation involving the centred data matrix $\tilde{Y}$. It is easy to verify that the total SSCP $T$ can alternatively be written as $T = \tilde{Y}'\tilde{Y}$; and that $\text{trace}(T)$ is the sum of the squared elements of $\tilde{Y}$, referred to as the ‘total variation’ in Section 3. This quantity is the squared Frobenius norm of $\tilde{Y}$ (Gentle 2007, Section 3.9), a standard measure of the overall magnitude of a matrix. Similarly, $\text{trace}(T_G)$ and $\text{trace}(T_R)$ are the squared Frobenius norms of $X_G\hat{\alpha}$ and $X_R\hat{\beta}$ respectively (the contributions of the GCMs and the RCMs to the centred data matrix, according to (4)); and $\text{trace}(T_E)$ is that of the residual matrix $e^{(2)}$. The trace-based partitioning of variation therefore provides an interpretable decomposition focused on the data matrix itself, rather than the SSCP.

Finally, we note that the traces of the components of (8) are related to the SVDs of the components of (4): if the singular values of $\tilde{Y} = Y - 1\hat{\mu}'$ are $\lambda_1 \geq \ldots \geq \lambda_n = 0$ as in Section 3, then $\text{trace}(T) = \sum \lambda_i$, with corresponding expressions for the remaining traces. The singular values of $X_G\hat{\alpha}$ are just $\sqrt{R}$ times those for the GCM EPPs, while those of $X_R\hat{\beta}$ are $\sqrt{G}$ times those for the RCM EPPs and $\text{trace}(T_E)$ is related to the SVD of the residual matrix $e^{(2)}$. The proposed methodology therefore provides a hierarchical partitioning of the ensemble variation: the squared singular values of the EPPs for each component of (4) sum to the variation explained by that component and, in turn, these componentwise contributions sum to the total variation $\text{trace}(T)$.

2) **Unbalanced ensembles**

If some $(r, g)$ combinations are missing from a regional ensemble then equation (8) no longer holds and, as discussed in Section 1, the ensemble variation cannot be partitioned unambiguously. In this case, instead of discarding or imputing members to obtain a balanced ensemble, one alternative is to determine the range of variation (RoV) that is potentially attributable to each source. This is done by performing two separate analyses, each based on a sequence of statistical models as described in the Supplement. In the first analysis, the maximum possible variation is attributed to the GCMs, with the RCM information used only to account for variation that cannot otherwise be explained. In the second analysis, the roles of GCMs and RCMs are reversed. The
difference between the results indicates the extent to which the partitioning is affected by the lack of balance: in the balanced case, both analyses recover the unique decomposition (8).

An alternative to the RoV approach is to provide a partitioning of variation that maintains a hierarchical relationship with the RCM, GCM and residual EPPs, as found in a balanced ensemble. This can be done via the alternative representation (9) of (8), because in the unbalanced case we still have estimates of the \( \{\alpha_g\} \), the \( \{\beta_r\} \) and \( \Sigma \) as described earlier. We therefore denote the three terms in (9) as \( T_G^\dagger \), \( T_R^\dagger \) and \( T_E^\dagger \) respectively, and define the estimated SSCP for a complete ensemble as

\[
T^\dagger = T_G^\dagger + T_R^\dagger + T_E^\dagger. \tag{10}
\]

The traces of the components \( T_G^\dagger \), \( T_R^\dagger \) and \( T_E^\dagger \) can now be used to estimate the relative contributions of each source of variation in a complete ensemble; and the corresponding EPPs provide a hierarchical partitioning of (10).

A potential objection to this approach is that it no longer provides an exact decomposition of variation for the observed ensemble. We therefore propose to use it in conjunction with the RoVs, as illustrated in the example below. Importantly, the RoVs are not guaranteed to contain the relative proportions derived from (10): they provide information on potential partitionings of variation in the observed ensemble, whereas (10) aims to account additionally for variation associated with unobserved ensemble members. If the observed members are unrepresentative in some sense — for example, if GCMs with above-average responses tend to be paired with RCMs that have below-average responses — then this lack of representativeness will feed into the RoVs. If the results from (10) lie outside these ranges therefore, this suggests that the characteristics of the available and missing ensemble runs may differ.

The use of an estimated partitioning of variation can be regarded as a form of imputation (see Section 1). By contrast with other imputation schemes however, it does not require estimation of the missing ensemble members: rather, it reweights the contributions to each SSCP to account for undersampled parts of the complete ensemble. This is similar in spirit to the well-established approach, in survey sampling, of reweighting to handle situations in which subgroups of a population are over- or under-represented (e.g. Little and Rubin 2020, Chapter 3).
Fig. 2. Estimated decomposition of variation in the completed EuroCORDEX ensemble, for bias in summer mean daily maximum surface temperature over the UK from 1989 to 2008. Panel (a) shows the ensemble standard deviations at each grid cell (i.e. the square roots of the diagonal elements of $T^\dagger/(n_\cdot - 1)$, with $T^\dagger$ as in (10)), while panels (b)–(d) show the diagonal elements of $T^\dagger_G$, $T^\dagger_R$ and $T^\dagger_E$ respectively, as percentages of the corresponding elements of $T^\dagger$. The traces of the respective matrices are also quoted, as percentages of trace($T^\dagger$).

c. Regional ensembles: an unbalanced example

The EPP methodology is now applied to the EuroCORDEX UK temperature biases of Section 2. The data matrix has $n_\cdot = 64$ rows corresponding to the ensemble members, and $S = 1652$ columns corresponding to grid cells.

Figure 2 shows the estimated partitioning of variation for a complete ensemble based on equation (10). Focusing first on the estimated ensemble standard deviations in Figure 2(a), derived from $T^\dagger$, the most obvious features are two local areas of substantially enhanced variation. One of these is in the south-east, corresponding to the Greater London area; the other is in the English Midlands which is also a major conurbation. Other areas of enhanced variation are less pronounced, and include non-urban regions (e.g. the Scottish Highlands) as well as upland and industrial areas of northern England. This suggests that as far as summer maximum temperature biases are concerned, the ensemble indicates some uncertainty associated with large urban heat islands and also, to a lesser extent, with topography.
Figures 2(b) to 2(d) use equation (10) to probe the sources of variation in the ensemble. The RCMs account for the highest percentage (53%) of the estimated total variation, although the GCMs also account for 38%: unstructured residual variation is relatively unimportant. These figures can be compared with the RoVs derived from the available ensemble members: these analyses reveal that the RCMs contribute between 35% and 62% of the total variation while the GCMs contribute between 29% and 56%. The GCM and RCM estimates from (10) both fall well within the respective ranges from the Type I analyses, whence there is no evidence that the available ensemble members are unrepresentative with respect to biases in summer $\text{tasmax}$. A closer inspection of the maps also reveals that in the two urban areas where the ensemble variation is highest, the RCMs account for up to around 70% of it: this perhaps reinforces recent evidence (Lo et al. 2020) that uncertainties in urban heat island effects are primarily attributable to the RCMs.

To understand the variation in more detail, Figures 3 and 4 show the GCM and RCM EPPs respectively. In both cases, panel (a) is the estimate of $\mu$ in (3). The first GCM EPP accounts for 92% of the GCM-attributed variation and shows a gentle north-south gradient; the second EPP is much less important. By contrast, the first RCM EPP — accounting for 87% of the RCM-related variation — clearly picks out the pattern corresponding to the effects of urban heat islands and topography. This demonstrates that the uncertainty regarding these effects comes predominantly from the RCMs, and in fact that it is the dominant pattern of RCM-related variation in the ensemble. Moreover, the EPP scores associated with this pattern enable us to identify the RCMs with the smallest and largest such effects, which are RACMO22E and REMO2015 respectively. For users with a particular interest in UK summer maximum temperatures in urban heat islands, this analysis therefore helps to identify the EuroCORDEX runs spanning the range of relevant historical biases.

5. Discussion

a. Summary of the proposed methodology

EPPs are designed to enable rapid exploration of structured climate model ensembles, particularly when the outputs of interest are high-dimensional. They are descriptive rather than inferential: in particular, we do not attempt to quantify uncertainties about sources of variation or to assess their ‘statistical significance’, which require more time-consuming methods as reviewed in the Introduction. Nor do they aim to estimate the underlying properties of a model or system: this
Fig. 3. EPP analysis for the estimated GCM effects in the 64-member EuroCORDEX ensemble, for bias in summer mean daily maximum surface temperature from 1989 to 2008. Panel (a) represents the ensemble mean; panels (b) and (c) are maps of the first and second GCM EPPs respectively; and panel (d) shows the associated EPP scores for each GCM in the ensemble. Subtitles for panels (b) and (c) indicate the percentage contributions of each EPP to the GCM-attributed variation.

Fig. 4. EPP analysis for the estimated RCM effects in the 64-member EuroCORDEX ensemble, for bias in summer mean daily maximum surface temperature from 1989 to 2008. The panels are directly analogous to those in Figure 3.

contrasts, for example, with superficially similar techniques that have been used to study the dynamical properties of individual climate models (e.g. Maher et al. 2018; Haszpra et al. 2020; Bódai et al. 2021). Estimation of system properties requires appropriately-designed ensembles of sufficient size. By contrast, an EPP analysis is essentially an arithmetical decomposition of the
ensemble: for balanced ensembles the decomposition is exact, while for unbalanced ensembles it relies on minimal assumptions corresponding to representations such as equation (3).

For the descriptive analysis of regional ensembles, (3) itself is relatively uncontroversial: it makes no distributional assumptions, although one potential restriction is the additive structure in which the effect of a given RCM is the same regardless of which GCM is driving it. This assumption can be relaxed if the ensemble contains multiple runs of each GCM:RCM combination: the Supplement provides details. The model provides interpretable summaries of the GCM and RCM effects via the least-squares coefficient estimates, and provides a framework for reweighting contributions to the SSCP in unbalanced ensembles via equations (9) and (10). This reweighting does not require the estimation of missing ensemble members; nor does it require that members are discarded to obtain a balanced subset. An open question, however, is to assess the uncertainty associated with the use of $T^\dagger, T^\dagger_G, T^\dagger_R$ and $T^\dagger_E$ in place of $T, T_G, T_R$ and $T_E$. A related question is considered by Christensen and Kjellström (2022), who use heuristic arguments to examine the implications of missing ensemble members for estimation of the coefficient vectors ($\{\alpha_g\}$ and $\{\beta_r\}$ in the current context). It is not obvious, however, that these arguments can be extended to examine the effect on the partitioning of variation.

Although EPP analysis is designed primarily for descriptive purposes, it is natural to ask how it is affected by sources of variation that are not considered explicitly — for example, equation (3) contains no direct representation of internal variability, because the EuroCORDEX ensemble does not provide multiple runs of each GCM:RCM pair with varying initial conditions. In such cases the associated unmodelled or unrepresented variation is subsumed into the residual term $T_E$ in (8), or its estimate $T^\dagger_E$ in (10). A consequence is that if, as in Figure 2, the residuals account for only a small proportion of the total variation, then the unmodelled sources must themselves contribute relatively little.

b. How many EPPs?

In Figure 3, most variation in the GCM effects is dominated by the first two EPPs which represent respectively a spatial monopole and dipole. A reviewer has pointed out that this situation is common, and queried whether higher-order EPPs may reveal more nuances of structure. We have investigated this (details are in the accompanying code — see Data Availability Statement).
The higher-order EPPs are relatively unimportant (for example, GCM EPPs 3 and 4 contribute respectively 0.9% and 0.3% of the GCM-attributed variation) and, moreover, exhibit no interpretable spatial patterns. These results are typical for GCM EPPs in our experience — although the monopole / dipole pattern is not so typical for RCM EPPs as exemplified by Figure 4. At some level, a lack of ‘interesting’ higher-order structure is itself noteworthy: for example, it suggests that identification of representative ensemble members can be done using just the first two EPP scores for each source of variation. Further investigation is needed, however, to determine whether these findings can be replicated across different ensembles, regions, time periods and quantities of interest.

c. Alternative measures of variation

To summarise the overall magnitudes of the SSCP matrices involved in the partitioning of total variation, the development above uses their traces. Although this approach is appealing for its relationship with the Frobenius norm of the centred data matrix, it neglects the off-diagonal elements of the SSCPs. These elements are related to the correlations between neighbouring locations and thus, in principle, contain information about the spatial extent of ‘typical’ differences between ensemble members. Alternative summaries of the SSCPs have been considered in the MANOVA literature (e.g. Huberty and Olejnik 2006, Chapter 3), albeit justified by the assumption (not made in the development above) that the residual vectors (the \( \{ \varepsilon_{rg} \} \) in (3)) have multivariate normal (Gaussian) distributions. One such alternative derives from fact that in the Gaussian case the least-squares MANOVA coefficient estimates are also the maximum likelihood estimates (Krzanowski 1988, Section 15.2); hence fitted models can be summarised in terms of their maximised log-likelihoods. In particular, the scaled deviance for a model is defined as twice the difference between its maximised log-likelihood and the highest log-likelihood attainable (i.e. the log-likelihood from a model that fits the ensemble outputs perfectly). The scaled deviance is a measure of ‘lack of fit’: for example, in linear regression models it is proportional to the residual sum of squares (Davison 2003, Section 10.2).

For a two-way additive MANOVA such as equation (3), it can be shown (see the Supplement) that in a balanced ensemble the scaled deviance for Model 0 can be partitioned into contributions from the GCMs, RCMs and residuals in exactly the same way as the total trace, and that the proportions
of scaled deviance attributable to the GCMs and RCMs are \( \text{trace}[\Sigma^{-1}T_G]/\text{trace}[\Sigma^{-1}T] \) and \( \Sigma^{-1}T_R]/\text{trace}[\Sigma^{-1}T] \) respectively. Estimated extensions to unbalanced ensembles using (10) are immediate. Some care is needed with the calculations in practice however, as described in the Supplement. We have applied the approach to several quantities derived from the EuroCORDEX ensemble: in all cases, a high proportion of the deviance was attributable to the RCMs, even in situations where the GCMs dominated the trace-based partitioning of variation. The deviance-based partitioning accounts for the spatial correlation structure of the ensemble members via the \( \Sigma \) matrix: any differences compared with the trace-based partitioning must therefore be associated with this correlation structure, implying that the ensemble members contain differing spatial ‘patches’ of above- and below-average values that are associated with inter-RCM variation. This implication is unsurprising, and suggests that the trace-based partitioning of variation yields more useful insights than a partitioning based on deviances. Further work is needed to determine whether this conclusion holds in general, or whether it is specific to the ensembles and study region considered in this paper.

\textit{d. Extensions and other potential applications}

Our example considers maps of a single climate index (bias in \texttt{tasmax}). The methodology can also be applied to multiple indices simultaneously: all that is required is that the ensemble data matrix \( Y \) contains the quantities of interest. For example, if a user wishes to select regional ensemble members for use with an impacts model taking mean winter temperature and total winter precipitation as inputs, one option is to carry out an EPP analysis of the \( n \times 2S \) data matrix containing the relevant values of both temperature and precipitation. An inter-model EOF analysis along these lines is considered by Zhou et al. (2020). In cases involving variables with different units of measurement however, it is desirable to standardise the data for each variable prior to analysis so that the results are not dominated by the contributions from individual variables: see Krzanowski (1988, Section 2.2) for a discussion of this in the closely related context of principal components analysis. In climate science, it is also common to standardise each index on a per-grid-square basis before calculating SVDs, when performing PCA or similar. The appropriateness of this for an EPP analysis depends on the context. We did not do it, because our goal was to understand the spatial variation in the ensemble: for example, standardising the EuroCORDEX
temperature biases individually for each grid square would have removed the interesting excess variation associated with urban heat islands and topography in the first panel of Figure 2.

The potential applications of EPP analysis extend beyond the regional ensemble example considered above. Consider, for example, a CMIP ensemble with different numbers of runs for each model (see references in Section 1). The EPP analysis of such an ensemble starts from the representation

\[ Y_{gi} = \mu + \alpha_g + \varepsilon_{gi}, \]  

where \( Y_{gi} \) now denotes the output from the \( i \)th run of the \( g \)th CMIP GCM \((g = 1, \ldots, G)\); \( \mu \) represents an overall mean; \( \alpha_g \) represents a systematic departure from this mean for the \( g \)th GCM; and \( \varepsilon_{gi} \) represents residual variation. (11) is the direct analogue of (3): the least-squares estimates of \( \mu \) and \( \alpha_g \) are now \( \hat{\mu} = G^{-1} \sum_{g=1}^{G} Y_g \) and \( \hat{\alpha}_g = \bar{Y}_g - \hat{\mu} \) respectively, where \( \bar{Y}_g \) is the mean of the ensemble members from the \( g \)th GCM. An EPP analysis thus focuses on the SVDs of the \( \hat{\alpha}_g \) and, if relevant, the residuals \( Y_{igi} - \hat{\mu} - \hat{\alpha}_g \). This is essentially the inter-model EOF approach taken, without formal justification, by Yim et al. (2016) (see Section 1).

Moving beyond regional and single-scenario CMIP ensembles, the Supplement describes extensions to ensembles with more complex structures. For example, the approach could be used to explore the spatial structure of GCM- and scenario-specific variation in the CMIP ensembles which, as noted above, are often highly unbalanced: a simple application in this setting would use the GCM EPPs to characterise (dis)similarities between models in terms of their spatial patterns of projected future change. A more sophisticated analysis might focus on scenario effects: in such an analysis it may be reasonable to expect that the first scenario EPP will correspond to an overall pattern of change, and for the corresponding scenario-specific EPP scores to be related to some measure of net radiative forcing. Any departures from this expected pattern could yield interesting insights into the dynamics of the models. Other potential applications are to ensembles in which a single model is used to obtain projections for each of a set of emissions scenarios, starting from each of a common collection of carefully chosen initial conditions: here, an EPP analysis of the residual / interaction term in an additive representation of scenario and initial condition effects could potentially reveal information about the state-dependence of climate change signals.

EPPs can also be used to identify gaps in an existing structured ensemble, and hence to identify design priorities for additional runs. For example, in a regional ensemble each member can be
summarised using the first EPP scores for the corresponding RCM and driving GCM respectively: the ensemble structure can then be visualised as a scatterplot of the corresponding pairs of scores. Such a plot will reveal combinations of scores — and hence of characteristic modes of behaviour — that are not well represented and hence could be prioritised in subsequent ensemble updates.

Finally, we note that EPP analysis has potential applications beyond climate model ensembles. One such application is to gridded data products providing estimates of quantities that are not observed directly at the locations of interest: in such settings, one way to characterise the estimation uncertainty in the data product is to provide multiple samples from its joint uncertainty distribution. The uptake of such such techniques by data product providers is currently low, although they are likely to become more widely available in the future (Chandler et al. 2012). EPPs provide one possible route for data product users to choose informed and representative subsets of samples, enabling uncertainty to be propagated through their subsequent analyses.

Acknowledgements. This research was funded under the UK Climate Resilience programme, which is supported by the UKRI Strategic Priorities Fund. The programme is co-delivered by the Met Office and NERC on behalf of UKRI partners AHRC, EPSRC, ESRC.

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 (see column headings in Figure 1) for producing and making available their model output. We also acknowledge the Earth System Grid 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 thank the Editor and three reviewers for their careful reading and constructive comments on an earlier draft of the paper.

Data availability: All figures and analysis can be reproduced using python scripts and example data linked from https://www.ucl.ac.uk/statistics/research/eurocordex-uk. The scripts also include an additional example, demonstrating the EPP analysis of an unstructured ensemble.
References


