ESAIM: PROCEEDINGS AND SURVEYS, 2018, Vol. 60, p. 104-113
Jean-François Coeurjolly & Adeline Leclercq-Samson

IS FUTURE CLIMATE PREDICTABLE WITH STATISTICS?

Aurélien Ribes[1]

Abstract. The purpose of this note is to briefly introduce the statistical models and methods used in climate sciences to estimate, from observations, the sensitivity of the Earth's climate to Greenhouse Gases. First the context of climatology is described with an explanation of how statistics can interact with the use of climate models. A description of the main models used, which are original variants of Error-in-Variables models, follows. Then a few issues for which methodological progresses would be helpful are mentioned. This includes the inference of large covariance matrices and uncertainty quantification.

Résumé. Cette note a pour but d'introduire brièvement les modèles et outils statistiques utilisés en sciences du climat afin de quantifier, à partir d'observations, la sensibilité du climat à l'augmentation de l'effet de serre. Je commence par présenter le contexte actuel des sciences du climat, et la façon dont la statistique interagit avec l'utilisation des modèles de climat. Je décris ensuite les principaux modèles utilisés, qui sont des variantes originales des modèles à erreur. Enfin je mentionne quelques-uns des problèmes méthodologiques rencontrés, liés notamment à l'estimation de grandes matrices de covariance et à la quantification d'incertitudes.

Introduction

Climate sciences have been receiving increased attention over the last four decades, as a direct consequence of the threat implied by human influence on climate. This short note will focus on how to predict future climate, combining the statistical analysis of observed emerging climate change with current climate change knowledge.

The problem is as follows. Given a scenario describing the time evolution of anthropogenic emissions of greenhouse gases (GHGs: CO_2, CH_4, N_2O and others), e.g. from now to 2100, can we estimate the response of the climate system in terms of global mean temperature and/or other (more specific) variables? Previous attempts to address this question, most notably IPCC assessment reports[2] [11], have resulted in estimates with substantial uncertainties. For instance, in response to a *business as usual* emission scenario[3] the global mean temperature was assessed to increase by 3.2 – 5.4°C by the end of the 21st century (2081-2100 relative to 1850-1900). Uncertainty in this range of values is large: almost a factor of 2. Note however that the overall uncertainty regarding future climate is much larger, e.g. 0.9 – 5.4°C for the same period, if the (large) uncertainty related to the emission scenario is included. Behind this, there are large uncertainties on how sensitive the Earth's climate is to an increase in GHG concentrations. To speak about this *sensitivity*, two theoretical quantities have

[1] CNRM, Météo France and CNRS, 42 Avenue Gaspard Coriolis, 31057 France

[2] The Intergovernmental Panel on Climate Change (IPCC) is dedicated to providing objective, scientific views on climate change and its impacts. IPCC periodically produces Assessment Reports – the last such synthesis was published in 2013-2014 under the title of IPCC 5th Assessment Report.

[3] *Business As Usual* (BAU) means that no efforts are made to reduce global GHG emissions. Here we refer to the RCP8.5 scenario used in the IPCC AR5 ([11]) as a BAU scenario

Article published online by EDP Sciences and available at http://www.esaim-proc.org or https://doi.org/10.1051/proc/201760104

been introduced. The Equilibrium Climate Sensitivity (ECS) is the increase of global mean temperature after a doubling of the atmospheric concentration of CO_2, when the climate system reaches a new equilibrium. The Transient Climate Response (TCR) is the increase of global mean temperature at the time of CO_2 concentration doubling, after an exponential increase of 1%/yr – doubling is reached after 70 years. The likely ranges assessed for these two variables are 1.5–4.5°C and 1–2.5°C respectively ([11]), reflecting current uncertainties.

To derive such likely ranges, two main strategies can be said to have been competing in recent decades.

The first approach involves the use of numerical climate models, which are able to simulate the state and the dynamics of the climate system based on well-known physical equations (e.g. radiative physics, laws of thermodynamics, conservation of mass and energy, etc). This requires modelling the entire climate system[4]. The development of such models, with increasing levels of complexity, has been a central activity in climate sciences over the last few decades. Since the first global models were developed about 40 years ago, models have greatly improved their ability to reproduce the main features of the climate system. However, they still partly disagree when estimating the sensitivity to GHGs. As an exemple of model discrepancies, the ECS simulated by the last generation of models was 1.9–4.7°C. This range is very close the IPCC likely range reported above, which has remained remarkably unchanged since the first estimates in the late 1970's [3].

On the other hand, a second approach involves the estimation of Earth's sensitivity from available observations, as observations already provide some emerging information about climate change. As an illustration, the warming observed since the mid-19th century is about 0.8°C, while the CO_2 concentration has increased from $\sim$280 to $\sim$400ppm. Atmospheric CO_2 concentration is now increasing steadily, suggesting that observations could provide a narrower constraint range than climate models. The main issue with this approach is that several competing human activities have been influencing climate in the past decades. In addition to GHGs, anthropogenic aerosols[5] have induced some cooling, partially offsetting the GHG-induced warming[6]. As long-term changes will be largely dominated by the response to GHGs, past observations cannot be simply extrapolated. Instead, there is a need to disentangle the contributions of GHGs from those of other anthropogenic forcings[7] if we are to accurately predict the future. In order to distinguish between these influences, the main information available concerns the features of each response, in terms of time-series, spatial pattern, or both. In our case, this distinction is difficult to make as the responses to GHGs and other anthropogenic forcings share many common features over the observed period.

This paper is intended to provide a brief overview of the statistical techniques predominantly used in the climate community to address this issue of separating responses to different external forcings in observed data. Note that the same techniques were used to distinguish climate change from internal variability[8], and to investigate its causes – an area called *detection and attribution of climate change*. We focus on this specific topic, not because it involves more sophisticated or attractive statistics, but rather because it is critical with respect to current climate change knowledge. Therefore this paper is really driven by application, and aims to give a picture of current practice. There are also many other climate-oriented topics in which (sometimes more advanced) statistics are being used.

We will introduce the main statistical models in Section 1, together with the inference methods involved. We will briefly discuss a few specific issues and open challenges in Section 2 – these are topics in which, at least in the opinion of the author, enhanced relationships between the two communities might lead to improve procedures.

[4] Climate system is an interactive system consisting of five major components: the atmosphere; the hydrosphere, including the ocean; the cryosphere; the land surface and the biosphere

[5] Aerosols are particles suspended in the air. Anthropogenic aerosols are aerosols induced by human activities. These tend to cool the climate by reflection of incoming solar radiation and other physical processes.

[6] To illustrate this: the estimated 1951-2010 warming was +0.65±0.07°C, mostly explained by human influence (+0.7±0.1°C). However, the uncertainty on the GHG-induced warming (+0.9±0.4°C) and the cooling induced by other anthropogenic factors (-0.25±0.35°C) over the same period are much larger. See also Figure 1 (right panel).

[7] *External forcings* are agents outside the climate system which can cause a change in the climate system. They can be either natural (e.g. volcanic eruptions, solar activity) or anthropogenic (e.g. GHG emission).

[8] *Internal variability* refers to the variability of the climate system that would be observed without any change in external forcings. This variability is related to the chaotic dynamics of the system and is usually treated as random.

1. Statistical models and methods

The choice of the statistical models used in climate sciences is strongly related to climate models, how they are used, and the type of reliable information they can provide [9]. This interaction is discussed in Section 1.1. Two related models are discussed in Sections 1.2 and 1.3.

1.1. Choice of the statistical models

The first statistical models introduced to deal with the above mentioned topic were linear regression models where observations were regressed onto the expected (i.e. simulated) response to each forcing. These models can be written as follows:

$$Y = \sum_{i=1}^{k} \beta_i X_i + \varepsilon, \qquad \varepsilon \sim N(0, \Sigma), \tag{1}$$

where Y stands for the observations, X_i is the expected response to forcing i (simulated by a climate model), β_i is an unknown scaling factor, and ε is a random term describing internal climate variability. Y, X_i and ε are all vectors of size n. Note that this framework was first introduced in *detection and attribution* studies in order to assess the causal relationship between anthropogenic forcings and the observed warming.

Several different variables can be included into Y. Most studies focus on one physical variable such as the atmospheric near-surface temperature, or precipitation, or others, and consider this variable at different locations and times. Then Y is a spatio-temporal vector that provides a picture of the observed changes over space and time. The construction of X and Σ is similar, but the way they are estimated deserves some discussion.

Each X_i is constructed from a specific model simulation – which is sampled in space and time like Y. More precisely, in order to determine the expected response to a forcing i, the common practice is to perform a specific numerical experiment in which only forcing i varies over time – all other forcings are held constant. In that way, the discrimination between the different forcings is based on their respective spatio-temporal properties. This statistical approach is only based on climate model outputs, which means that the transfer function from, e.g., GHG concentrations to the climate system response, is mainly not estimated using statistics – this task is really left to physical understanding through climate model integrations. Note that there is a long tradition of considering spatio-temporal features to distinguish among forcings. Pioneering studies in the mid 1990's only focused on spatial correlation of observed and simulated trends to discuss human influence. This choice was motivated by the fact that different forcings typically involve different physical processes, resulting in contrasted responses[9].

The last quantity involved in (1) is the covariance matrix Σ. In the rest of this section, Σ is assumed to be known, so that estimation within model (1) is very well-known. In practice, Σ has to be inferred somehow. This raises several issues which are discussed in Sections 2.2 and 2.3.

A simple illustration of such data is provided in Figure 1 for global mean temperature. Note that this example is restricted to time-series while most studies use spatio-temporal vectors as for X_i and Y.

This statistical framework being posed, inference is performed to estimate β, i.e. the magnitude of each forced response, assuming that the spatio-temporal response pattern is known. In this way, the statistical analysis can correct any over- or under-estimation of the response by climate models. β is estimated with Generalised Least Squares

$$\widehat{\beta} = (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y \sim N\big(\beta, (X'\Sigma^{-1}X)^{-1}\big), \tag{2}$$

which coincides with the Maximum Likelihood Estimator (MLE) and has well-known optimal properties. As its distribution is also known, exact confidence intervals can be derived. The main point to notice in equation (2) is that Σ^{-1} is involved, which is usually much more difficult to estimate than Σ (see Section 2.3).

[9] Among other famous examples, an increase of greenhouse gas concentration implies a surface warming but also a cooling of the upper atmosphere, while an increase of solar irradiance, which could lead to a similar surface warming, will imply a warming of the upper atmosphere.

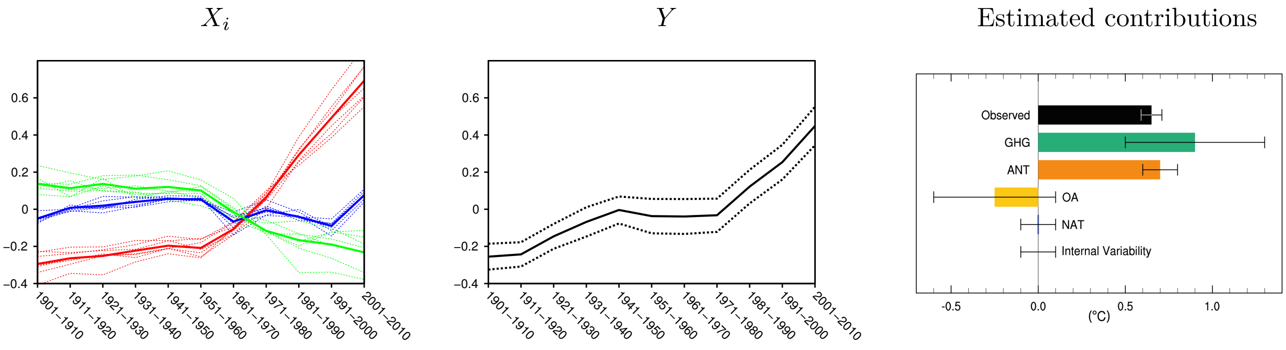

Figure 1. **Illustration based on real data.** The data involved in statistical models (1) or (3)-(5) are illustrated with global mean temperature time-series (decadal averages) over the period 1901-2010. **Left:** The expected responses to external forcings, X_i, as simulated by 7 individual climate models (dotted lines), and their multi-model average (solid lines). Blue: Natural forcings (NAT); Red: Greenhouse gases (GHG); Green: Other Anthropogenic forcings (OA). **Middle:** Historical observations Y (solid lines) taken from the HadCRUT4 dataset [16], with an indication of the magnitude of internal variability (dotted lines, corresponding to $\pm 2\sigma$), as estimated from models (Σ in (1), Σ_Y in (4)). **Right:** Assessed *likely* contributions of various groups of external forcings to the 1951-2010 linear trend on global mean temperature, as reported in the last IPCC report [2,11] (Reproduction of Fig. 10.5 from this report). ANT denotes the total Anthropogenic forcings, i.e. GHG+OA. Uncertainty on the "Observed" value (black) corresponds to the observational uncertainty, as estimated in HadCRUT4.

The estimation of the sensitivity to greenhouse gases – the main issue discussed in this paper – is directly related to the estimation of the β_{GHG} coefficient. Then, the estimate $\widehat{\beta}_{GHG}$ (with its uncertainty) can be used to rescale/correct the model response to greenhouse gases in order to provide a revised (observationally constrained) estimate. It is usually hoped that the spread in climate models will be reduced by this correction step, i.e. that more accurate estimates will be obtained thanks to the statistical analysis.

1.2. Error in Variables Models

Statistical model (1) was slightly complexified in the early 2000's in order to relax the assumption that X is known. In fact, each X_i comes from a climate model simulation, in which internal variability is at play – just like in the real world. Let X_i^* be the true (unknown) response of the model – which would be obtained by averaging over an infinite ensemble of independent realisations. Allowing for noise ε_{X_i} in X_i leads to the revised statistical model:

$$Y^* = \sum_{i=1}^{k} \beta_i X_i^*, \tag{3}$$

$$\left\{ \begin{array}{lcll} Y & = & Y^* + \varepsilon_Y, & \varepsilon_Y \sim N(0, \Sigma_Y), \\ X_i & = & X_i^* + \varepsilon_{X_i}, & \varepsilon_{X_i} \sim N(0, \Sigma_{X_i}), \quad i = 1, \ldots, k, \end{array} \right. \begin{array}{r} (4) \\ (5) \end{array}$$

In this framework, each X_i^* is an unknown latent variable which has to be estimated in addition to β (Y^* cannot be considered as another parameter, given (3)). The values fitted by the model are then $\widehat{X}^* \widehat{\beta}$. Models of this type are usually called *Error in Variables* or *Measurement error models*, as the predictors X^* are not exactly known [5]. They were first introduced in the climate literature by [1], although not in this general form.

The (-2 log-)likelihood can be written

$$\ell_{\mathrm{EIV}}(\beta, X^*) = (Y - X^*\beta)'\Sigma_Y^{-1}(Y - X^*\beta) + \sum_{i=1}^{k}(X_i - X_i^*)'\Sigma_{X_i}^{-1}(X_i - X_i^*) + cst. \tag{6}$$

For any given β, maximization of ℓ_{EIV} with respect to X^* is explicit. Let $\widehat{X}_\beta^*$ be the value at which the maximum is reached. This can be used to derive a profile (or concentrated) likelihood

$$\begin{aligned} \ell_{\mathrm{EIV}}^p(\beta) &= \ell_{\mathrm{EIV}}(\beta, \widehat{X}_\beta^*), & (7)\\ &= (X\beta - Y)'\left(\Sigma_Y + \sum_{i=1}^{k}\beta_i^2\Sigma_{X_i}\right)^{-1}(X\beta - Y) + cst. & (8)\end{aligned}$$

Note that the likelihood of the standard Gaussian linear model (i.e. Ordinary Least Square or OLS, see Section 1.1) is obtained as a limiting case, when $\Sigma_{X,1}, \dots, \Sigma_{X,k} \to 0$.

1.2.1. *Case $\Sigma_Y = \lambda\Sigma_{X_i}$ (TLS)*

If internal variability is the only source of uncertainty in ε_Y and each ε_{X_i}, it is reasonable to assume that $\Sigma_Y = \lambda\Sigma_{X_i}$. If only one simulation is used to estimate X_i, then $\lambda = 1$. In this way, it is assumed that the model's internal variability (Σ_{X_i}) is equal to that of the real world (Σ_Y). If an average over n_i independent realisations is computed to estimate X_i, then $\lambda = 1/n_i$ – this is the most common situation.

Under this assumption, multiplying (4)–(5) by $\Sigma_Y^{-1/2}$ preserves the linear relationship in (3) between Y and the X_i^*, and makes each ε a white noise. Then, the maximum likelihood estimate is derived from the Singular Value Decomposition (SVD) of the $n \times (k+1)$ matrix $[Y, X_1, \dots, X_k]$[10].

Such models are sometimes called Total Least Square (TLS) models. The TLS name comes directly from the geometric interpretation of the univariate case: the TLS fit minimizes the distance of data points from their orthogonal projections onto the regression line, as opposed to the usual OLS (Ordinary Least Square) which minimizes the distance along the y-axis.

The TLS approach has several specific features. First, unlike in OLS, no exact formula is available to compute confidence intervals. These are then computed from asymptotic techniques, and are usually found to be permissive (i.e. over-confident). Note also that computing confidence regions for the fitted values $X^*\beta$ (instead of either X^* or β) is a bit challenging. Second, and most importantly, the profile likelihood $\ell^p(\beta)$ is bounded[11]. In practice, $\ell^p(\beta)$ is sometimes very flat, which makes the TLS estimate of β somewhat unstable – the estimated slope can take very large values. This flat likelihood can also lead to open-ended confidence regions (e.g. regions which include $\pm\infty$, or even all of $\mathbb{R}$), which are difficult to interpret physically.

The use of OLS or TLS has long been a matter of debate in statistics. They both correspond to MLEs, but in different statistical models: (1) for OLS vs (3)-(5) for TLS. Therefore, the use of one or the other technique has to be related to whether or not X contains random uncertainties. We can also wonder what happens if the wrong model is used. If the data come from an OLS model, but the TLS estimate is used, it is non-optimal, mainly due to a larger variance. If the data come from a TLS model, but the OLS estimate is used, it is biased

[10] Minimizing the likelihood is equivalent to minimizing $\left\|\left[\varepsilon_Y, \varepsilon_{X_1}, \dots, \varepsilon_{X_k}\right]\right\|_2^2$. As

$$\left[Y, X_1, \dots, X_k\right] = \left[\sum_{i=1}^{k}\beta_i X_i^*, X_1^*, \dots, X_k^*\right] + \left[\varepsilon_Y, \varepsilon_{X,1}, \dots, \varepsilon_{X_k}\right],$$

we look for the best approximation of $Z = [Y, X_1, \dots, X_k]$ by a rank k matrix (the rank of Z is $(k+1)$ a.s.). If $Z = \sum_{i=1}^{k+1}\lambda_i u_i v_i'$ is the SVD of Z, with $\lambda_1 \geq \dots \geq \lambda_{k+1} > 0$, then the solution is given by $\widehat{Z}^* = \sum_{i=1}^{k}\lambda_i u_i v_i'$, and $\widehat{\beta}$ is derived from v_{k+1}.

[11] This is clear using the geometric point of view: the distance to orthogonal projections is always finite – even for a vertical regression line.

towards 0. Additionally, in this case, the upper bound of confidence intervals will be particularly underestimated (i.e. will be smaller than β far too often, assuming $\beta > 0$). This is the main reason why the TLS approach has become popular in the climate field: authors did not want to underestimate the possibility of a large sensitivity to GHGs.

In spite of the above mentioned (and somewhat inconvenient) features of the TLS technique, it has become a standard approach in the field, and is currently the *state-of-the-art* method. The main results reported in [11] (see in particular [2]) were based on this technique.

1.2.2. *Case* $\Sigma_Y \neq \lambda \Sigma_{X_i}$ *(General EIV)*

If additional sources of uncertainty are taken into account (i.e. other than internal variability), the relation $\Sigma_Y = \lambda \Sigma_{X_i}$ might not hold. Such additional sources can come from measurement uncertainty on Y (as on any physical measurement), or climate modelling uncertainty in X. The latter is required to account for the fact that models are not able to simulate the spatio-temporal response patterns perfectly. As is to be expected, climate models do exhibit some spread in this respect, as illustrated in Figure 1 (left panel). Such general models were first introduced in the field in [10]. Statistical inference within such a model was discussed in [8]. Surprisingly, given how simple this statistical model is, almost no statistical literature is available for this case.

In the case $\Sigma_Y \neq \lambda \Sigma_{X_i}$, (3)-(5) cannot be simultaneously pre-whitened by preserving the linear relationship between Y and the X_i. As no other technique is known to maximize the likelihood, the MLE has no closed-form expression. In order to derive a numerical approximation of MLE, a numerical iterative algorithm was proposed in [8]. As mentioned above, maximizing the likelihood in X^* given β is easy; maximizing in β given X^* is similarly easy (closed-forms in both cases). The algorithm proposed therefore involves iterative partial maximizations of the likelihood in β and X^*. Each step increases the likelihood, until convergence (as the likelihood is bounded). However, there is no guarantee that the sequence converges towards the global maximum of the likelihood. Then, it is possible to derive confidence regions on β from asymptotic properties of the MLE. As in the previous case, such confidence regions have been reported to be too permissive (i.e. over-confident).

In spite of some attractiveness, this general EIV framework has been very little used with real climate data. One reason is that the inference technique proposed is somewhat complicated, and was introduced quite recently. Another reason – possibly the main one – is that estimating the input matrices Σ_{X_i}, in particular the contribution related to climate modelling uncertainty, is very challenging (see 2.4). The current practice is therefore somewhat unsatisfactory, as error-in-variable models are being used (involving some complexity), but climate modelling uncertainty, which is likely to make the largest contribution to the error in X, is neglected.

1.3. **Alternative**

Recently, a simple alternative to the regression-based approach has been proposed in [19]. In this approach, it is assumed that climate modelling uncertainty applies consistently to both response magnitude and pattern. In this way, the asymmetry discussed in 1.1 between these two features of the response is broken, and the introduction of unknown coefficients β is no longer required. Other assumptions are unchanged, and climate model outputs are used for similar purposes.

The statistical model can be written as follows:

$$Y^* = \sum_{i=1}^{k} X_i^*, \tag{9}$$

$$\left\{ \begin{array}{rcll} Y &=& Y^* + \varepsilon_Y, & \varepsilon_Y \sim N(0, \Sigma_Y), \quad (4) \\ X_i &=& X_i^* + \varepsilon_{X_i}, & \varepsilon_{X_i} \sim N(0, \Sigma_{X_i}), \quad i = 1, \ldots, k, \quad (5) \end{array} \right.$$

with notation consistent with the previous subsection. The only difference with respect to (3)–(5) comes from (9). It is basically assumed that observations Y provide information on the sum of the forced responses involved.

The main parameter of interest is then X^* (as opposed to $X^*\beta$). As all errors ε are assumed to follow Gaussian distributions, both the MLE of X^* and its distribution are obtained in closed-form formulas:

$$\widehat{X}_i^* \;=\; X_i + \Sigma_{X_i}\Big(\Sigma_Y + \textstyle\sum_{i=1}^{k}\Sigma_{X_i}\Big)^{-1}(Y - X) \;\sim\; N\Big(X_i^*, \Big(\Sigma_{X_i}^{-1} + \big(\Sigma_Y + \textstyle\sum_{j\neq i}\Sigma_{X_j}\big)^{-1}\Big)^{-1}\Big). \qquad (10)$$

Inference is therefore much easier (and more accurate) than in the general EIV case, while identical assumptions are used regarding the sources of uncertainty involved.

Two remarks can be made about this model. First, the linear regression case (Section 1.1) can be obtained as a limiting case, when the error in X (i.e. Σ_X) is restricted to its amplitude, and becomes arbitrarily large, i.e. $\Sigma_X = \lambda X X'$, and $\lambda \to \infty$. So, if uncertainty is really concentrated on the magnitude of the response, the two approaches should lead to similar results. Second, this approach can be interpreted in a Bayesian perspective. Relation (5) provides a prior on each X_i^* (derived from the information provided by climate models), and the observations Y provide some additional information which can be used to derive a posterior. This Bayesian point of view could be used to extend this approach to non-Gaussian distributions quite easily.

Like EIV, this approach requires each Σ_{X_i} to be estimated, which is very challenging (see Section 2.4). In addition, this method has been proposed only very recently. These are two reasons why it has been little used up to now.

2. A few open challenges

This section briefly presents a few statistical issues for which current practice in climate sciences might be improved, at least in my opinion. There is no claim of completeness with respect to the selected issues nor with respect to the existing climate literature.

2.1. **Large dimension in climate datasets**

As in many other fields, typical datasets in climate are very large. For example, observations of near-surface atmospheric temperature cover the period from 1850 to present, with a 5°×5°resolution at the annual (or even monthly) time-step. This means about 2600 points in space, and about 160 points in time. The resulting observation vector Y (used in Section 1) is typically of dimension 10^5, when only one variable (temperature), one level (near-surface) and annual means are considered. The large-dimension issue is even more obvious when we consider climate model outputs, as climate models are continuously increasing their resolution (both in space and time), always remaining near the boundary of computational capacity.

Climate scientists routinely use *pre-processing* techniques to reduce and/or summaries this large amount of data. Among the most popular techniques are the computation of averages (either over time, e.g. annual averages, or over space), the projection onto principal components (usually referred to as Empirical Orthogonal Functions by climatologists), the projection onto spherical harmonics (which provide a very convenient basis for numerical calculations in climate models), or a combination of these. For the specific question discussed throughout this paper, the size of the observed vector Y is (after pre-processing) typically about a few tens – say 30. The extreme compression that is applied in this way is designed to focus on the largest space-time scales, but remains largely arbitrary.

Finally, it is quite difficult to determine which variables, which regions, which periods, or which combinations of these could provide most information with respect to the initial question. Statistical methods able to tackle this issue could be of great interest.

2.2. **Estimating internal climate variability**

Current state-of-the-art approaches (Sections 1.1 and 1.2.1) need the covariance matrix of internal climate variability, say Σ. In practice, this has to be estimated somehow, which is usually a challenging task.

Several issues arise when estimating Σ. First, there is no observed data of internal variability only - internal and externally forced variations are always mixed in the observed records. Second, the number, the quality and the location of observations have varied substantially over time. A direct consequence of this is that time-series like the global mean temperature are usually not assumed to be stationary over time (e.g. the *mean* is computed over a much smaller sample for 1850 than for recent years). Third, internal variability exhibits complex dependence structures in both space (e.g. the El Niño phenomenon) and time (e.g. low-frequency variability over oceanic basins). In particular, the assumption that observations for one year are independent of those of the previous year is not suitable. Simple parametric models like auto-regressive models have also proved to be too limited. There is a substantial body of literature describing low-frequency variability on decadal or even longer scales. It is common practice to take this low-frequency into account, although observations give only limited information on it.

A consequence of the above-mentioned issues is that statistical tools and models that are commonly used by statisticians to describe spatial or spatio-temporal processes have been little used to model internal climate variability at the global scale. As a main alternative, many authors estimate Σ from a sample of *control* numerical experiments performed by climate models. These are idealised simulations where all external forcings are set constant – so the resulting variability is necessarily *internal*. Statistical issues arising from this technique are discussed below.

2.3. **Large covariance matrices**

In this subsection, we assume that an i.i.d. sample $\varepsilon_1, \dots, \varepsilon_p \sim N(0, \Sigma)$ is available from numerical *control* experiments to estimate Σ. We further consider the case where the dimension of Σ, say n, is close to p – this is a high-dimension framework[12].

Let $\varepsilon = [\varepsilon_1, \dots, \varepsilon_p]$ be the $n \times p$ matrix from which Σ can be estimated. The sample estimate $\widehat{\Sigma} = \frac{1}{p}\varepsilon'\varepsilon$ is a natural estimator of Σ; it coincides with the MLE, it is unbiased and asymptotically optimal. However, its spectrum is biased: largest eigenvalues are over-estimated while smallest eigenvalues are under-estimated. This becomes an issue in high-dimension, as the smallest eigenvalue of $\widehat{\Sigma}$ comes very close to 0 if $n = p$ ([15]). This is even more critical in our case as we need an estimate of Σ^{-1} (rather than Σ) to estimate β and/or X^* – see e.g. (2). And $\widehat{\Sigma}^{-1}$ has very poor properties.

In order to deal with this issue, regularised estimators have been introduced. The study of regularised estimators of covariance matrices is an active field of research in random matrix theory. However, the problem of finding an estimator of Σ to plug into (2) in order to provide an efficient estimate of β seems quite original.

The first attempt made ([17, 18]) was based on a regularisation with the identity matrix I, following [14]. In short, let us consider estimators $\widetilde{\Sigma}$ such that $\widetilde{\Sigma} = \gamma\widehat{\Sigma} + \rho I$, where γ and ρ are real coefficients. Among this family, Ledoit and Wolf [14] investigated those that were more accurate[13] than $\widehat{\Sigma}$, and then derived estimators of γ and ρ. The resulting estimator $\widehat{\Sigma}_I$ can be plugged into (2) (or used in the other techniques described in Section 1), i.e.

$$\widehat{\Sigma}_I = \widehat{\gamma}\,\widehat{\Sigma} + \widehat{\rho}\, I, \qquad \widehat{\beta}_I = (X'\widehat{\Sigma}_I^{-1}X)^{-1}X'\widehat{\Sigma}_I^{-1}Y. \tag{11}$$

Such a regularised estimate is helpful because it is much better conditioned than $\widehat{\Sigma}$ – the smallest eigenvalue in $\widehat{\Sigma}_I$ cannot be smaller than $\widehat{\rho}$. In [18], the estimator $\widehat{\beta}_I$ was shown to be more accurate than those obtained by using any More-Penrose pseudo-inverse of $\widehat{\Sigma}$ in (2) (formerly the most widespread approach for dealing with this issue).

More recently, other approaches have been introduced. Hannart and Naveau [7] introduced a regularisation technique towards any target Δ – i.e., Δ can be different from I. The proposed construction is based on a Bayesian prior $\Sigma \sim \mathcal{W}^{-1}(\Delta, \alpha)$, from which a new estimator $\widehat{\Sigma}_\Delta = \widehat{\rho}_1\widehat{\Sigma} + \widehat{\rho}_2\Delta$ is derived. Hannart [6] tested the resulting estimator $\widehat{\beta}_\Delta = (X'\widehat{\Sigma}_\Delta^{-1}X)^{-1}X'\widehat{\Sigma}_\Delta^{-1}Y$ and proposed to better account for the uncertainty on $\widehat{\Sigma}_\Delta$ when computing confidence intervals on β.

[12] Numerical *control* experiments currently available worldwide allows p near a few hundreds. As discussed in Section 2.1, n would be much larger than this if no *pre-processing* was involved. In practice, this *pre-processing* is designed to ensure that n is smaller than, or close to, p.

[13] Accuracy is meant with respect to the mean square error in $\mathcal{M}_n$.

Finally, some connections have been made between statistical climatology and available statistical literature on this topic. The resulting techniques, however, are not yet widely used, and could certainly be further improved.

2.4. Climate modelling uncertainty

The last topic to be addressed regards the estimation of climate modelling uncertainty. As mentioned before, the use of climate models is critical in climate sciences, and the question arises as to how far they should be trusted / what the associated uncertainties are. Note that these questions are central, not only to an understanding of past variations (as is proposed here), but also to describe uncertainty on future projections. For this reason, efforts have been made to propose objective techniques to quantify uncertainty, although some limitations apply.

Current techniques to quantify uncertainties rely on some exchangeability assumption, such as the "models are statistically indistinguishable from the truth" paradigm. It is then assumed that the departure of any model m_i from the truth m^* follows the same distribution as the difference between any pair of models (m_i, m_j). This is equivalent[14] to assuming that models share some common error, with a particular type of positive dependence among models:

$$m_i \sim N(m^*, 2\,\Sigma_m) \quad \text{and} \quad \mathrm{Cov}(m_i, m_j) = \Sigma_m. \tag{12}$$

However, this is a strong assumption, which can hardly be proven. Only a few studies where past observations were compared to the spread of model outputs provide limited and empirical evidence to support such an assumption.

Furthermore, even if this paradigm is assumed to be true, several issues arise. First, about 30-40 climate models are currently being used worldwide, and can be compared thanks to coordinated numerical experiments [4, 20]. This number remains very small, in particular in comparison with n (the size of Y). Second, these climate models should not be considered as independent. Most of them share common ideas, parametrizations, or even large pieces of code (e.g. the same atmospheric model, [13]). Consequently, this ensemble is not designed to explore uncertainty and instead has been described as an *ensemble of opportunity* [12].

The climate community is aware of these limitations. A few attempts have been made to comprehensively explore and quantify uncertainties. These were based on Perturbed Physics Ensembles (PPEs), where a subset of the (unknown) physical parameters used in the climate model are perturbed. So far, however, few connections have been established with the "analysis of computer experiments" community, either to calibrate (i.e. tune) models or to quantify uncertainties. As a result, the estimation of climate modelling uncertainty remains very challenging; this places a strong constraint on the statistical techniques to be used to infer future climate.

3. Conclusion

This note briefly reviews how statistical techniques are being used to infer the sensitivity of the Earth's climate to Greenhouse Gases. As the observed warming continues to strengthen, we are now at a point where statistical approaches are becoming competitive with climate model uncertainties. Improving current methods is then critical for deriving refined estimates of past and future changes. Given the limited interaction of the two communities over recent decades, there is certainly room for improvement in current practices and knowledge concerning climate change. Many other statistical challenges arising in other topics of climate sciences could also have been mentioned, such as the analysis of extreme events, the evaluation, emulation and tuning of numerical models (either in climate or weather forecasting), the refined description of observed or expected changes, and others. As a result, a large variety of statistical techniques might be of interest in this field.

[14] Here we assume that all random variables follow Gaussian distributions

References

[1] M.R. Allen and P.A. Stott. Estimating signal amplitudes in optimal fingerprinting, part i: theory. *Climate Dynamics*, 21:477–491, 2003.

[2] N.L. Bindoff, P.A. Stott, K.M. AchutaRao, M.R. Allen, N. Gillett, D. D. Gutzler, K. K. Hansingo, G. Hegerl, Y. Hu, S. Jain, I.I. Mokhov, J. Overland, J. Perlwitz, R. Sebbari, and X. Zhang. *Detection and Attribution of Climate Change: from Global to Regional*. In: *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA., 2013.

[3] J.G. Charney, A. Arakawa, D.J. Baker, B. Bolin, R.E. Dickinson, R.M. Goody, C.E. Leith, H.M. Stommel, and C.I. Wunsch. Carbon dioxide and climate: a scientific assessment, 1979.

[4] V. Eyring, S. Bony, G.A. Meehl, C.A. Senior, B. Stevens, R.J. Stouffer, and K.E. Taylor. Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016.

[5] W.A. Fuller. *Measurement error models*. John Wiley & Sons, 1987.

[6] A. Hannart. Integrated optimal fingerprinting: method description and illustration. *Journal of Climate*, 29(6):1977–1998, 2016.

[7] A. Hannart and P. Naveau. Estimating high dimensional covariance matrices: A new look at the gaussian conjugate framework. *Journal of Multivariate Analysis*, 131:149–162, 2014.

[8] A. Hannart, A. Ribes, and P. Naveau. Optimal fingerprinting under multiple sources of uncertainty. *Geophysical Research Letters*, 41:1261–1268, 2014.

[9] G. Hegerl and F. Zwiers. Use of models in detection and attribution of climate change. *Wiley interdisciplinary reviews: climate change*, 2(4):570–591, 2011.

[10] C. Huntingford, P.A. Stott, M.R. Allen, and F.H. Lambert. Incorporating model uncertainty into attribution of observed temperature change. *Geophysical Research Letters*, 33, 2006.

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

[12] R. Knutti, R. Furrer, C. Tebaldi, J. Cermak, and G.A. Meehl. Challenges in combining projections from multiple climate models. *Journal of Climate*, 23(10):2739–2758, 2010.

[13] R Knutti, D Masson, and A Gettelman. Climate model genealogy: Generation cmip5 and how we got there. *Geophys. Res. Lett*, 40:1194–1199, 2013.

[14] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[15] V.A. Marčenko and L.A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR - Sbornik*, 1(4):457–483, 1967.

[16] C.P. Morice, J.J. Kennedy, N.A. Rayner, and P.D. Jones. Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The hadcrut4 data set. *Journal of Geophysical Research*, 117(D8), 2012.

[17] A. Ribes, J.-M. Azaïs, and S. Planton. Adaptation of the optimal fingerprint method for climate change detection using a well-conditioned covariance matrix estimate. *Climate Dynamics*, 33(5):707–722, 2009.

[18] A. Ribes, L. Terray, and S. Planton. Application of regularised optimal fingerprinting to attribution. part i: method, properties and idealised analysis. *Climate Dynamics*, 41(11-12):2817–2836, 2013.

[19] A. Ribes, F.W. Zwiers, J.-M. Azaïs, and P. Naveau. A new statistical approach to climate change detection and attribution. *Climate Dynamics*, 48:367–386, 2017.

[20] K.E. Taylor, R.J. Stouffer, and G.A. Meehl. An overview of cmip5 and the experiment design. *Bulletin of the American Meteorological Society*, 93(4):485–498, 2012.