Dealing with trend uncertainty in empirical estimates of European rainfall climate for insurance risk management

Stephen Jewson 1 | Tanja Dallafior 2 | Francesco Comola 2

Abstract
The insurance industry uses mathematical models to estimate the risks due to future natural catastrophes. For climate-related risks, historical climate data are a key ingredient used in making the models. Historical data for temperature and sea level often show clear and readily quantified climate change driven trends, and these trends would typically be accounted for when building risk models by adjusting earlier values to render the earlier data relevant to the future climate. For other climate variables, such as rainfall in many parts of the world, the questions of whether there are climate change driven trends in the historical data, and how to quantify them if there are, are less simple to answer. We investigate these questions in the context of European rainfall with a specific focus on how to deal with the uncertainty around trend estimates. We compare 10 empirical methodologies that one might use to model and predict trends, including traditional statistical testing and alternatives to statistical testing based on standard methods from model selection and model averaging. We emphasize prediction and risk assessment, rather than detection of trends, as our goal. Viewed in terms of this goal, the methods we consider each have qualitative and quantitative advantages and disadvantages. Understanding these advantages and disadvantages can help risk modellers make a choice as to which method to use, and based on the results we present, we believe that in many common situations model averaging methods, as opposed to statistical testing or model selection, are the most appropriate.

Keywords
AIC, BIC, catastrophe model, climate change, cross-validation, insurance, model averaging, model selection, rainfall, statistical significance, trends

1 | INTRODUCTION
Roughly one-third of the insurance industry relates to property insurance, and this part of the industry charges roughly $1.6 trillion per year in premiums globally (McKinsey, 2020). Much of this is then paid out in response to insurance claims, and one of the main causes of large payouts is natural catastrophes. In order to quantify the risks of having to make large payouts due to natural catastrophes, property insurers use mathematical models known as catastrophe (cat) models. Descriptions of how cat models are constructed are given in the review...
paper by Friedman (1972) and the text books by Mitchell-Wallace et al. (2017) and Michel (2018). The results from cat models are used as inputs for the pricing of insurance contracts, the management of insurance portfolios and the determination of appropriate levels of reserves for insurance companies. Most insurance contracts start within a few weeks or months of being priced and run for a year. We will call this period of 1 year the contract period, and cat models often focus principally on estimating the risk during this contract period. For climate-related risks, there are some parts of the world and some perils for which seasonal forecasts (see, e.g., the review by Troccoli, 2010) may contain relevant information for the first part of the contract period. In addition, there may be parts of the world and perils for which decadal climate forecasting (see, e.g., the review by Smith et al., 2012) might be relevant. However, in other cases, neither seasonal nor decadal forecasts are relevant. For these cases, risk estimation is often based on estimates of the climate during the contract period based on just long-term historical climatology and, if appropriate, estimates of climate change trends. Even in the cases where seasonal or decadal forecasts contain relevant information, initial estimates of risk are often also derived from just long-term climatology and climate change trends, to provide a baseline. This baseline can then be used as a starting point from which the impacts of the inclusion of different seasonal or decadal forecasts can be evaluated. We will call the estimation of climate risks during the contract period using just long-term climatology and climate change trends ‘year-ahead prediction’. Since climate change trends are generally sufficiently small that climate is only slightly different from one year to the next, year-ahead prediction is more or less equivalent to estimating current climate conditions. The main relevance of the climate change trends to the analysis is in how to interpret the historical data, not in how to predict how the climate will change over the next year. If climate change trends are clearly present in the historical data then the older data are adjusted to bring it to the level of year-ahead climate, a process often referred to as detrending (see, e.g., the discussion on detrending in Jewson et al., 2005). However, in many cases the trends are uncertain and obscured by climate variability, and it is difficult to know whether to apply detrending or not. Understanding how to adjust past climate data for trends and make year-ahead predictions when the trends are highly uncertain is the topic of this article.

To study uncertain trends, we will focus on the year-ahead prediction of winter and summer total rainfall amounts for various regions in Europe. Such predictions are one of the factors that feed into the flood and drought risk modelling used by insurers, along with estimates of and modelling of the variability of rainfall within the season, including the likelihood of extremes, and the correlations of rainfall in time and space (for a description of a flood cat model, see Kaczmarska et al., 2018). European rainfall over the last 50 years does not show the same highly statistically significant trends due to climate change that we see in temperature and sea-level over the same period (Maraun, 2013), and it is also strongly affected by decadal and longer time-scale variability. As a result, it is difficult to know how to, or whether to, apply detrending to the historical rainfall data. One can certainly calculate the trends in the observed rainfall data, and they are never precisely zero because of the statistical noise of climate variability. But as estimates of any real underlying climate change, the observed trends are highly uncertain: for instance, we shall see below that the standard errors on the trend estimates can be as large as the trend estimates themselves. There are various impacts of this uncertainty: trend estimates may vary materially according to the time period used to estimate the trend; trend slopes may or may not be statistically significant; the question of statistical significance is sensitive to the significance threshold chosen; and in some cases trend slopes flip between statistically significant and not statistically significant simply because of the addition of a single year of data. As a result of the uncertainty in estimating rainfall trends, it is still common practice in insurance risk modelling to perform year-ahead prediction of rainfall climate using an assumption of stationarity, that is, no trend.

However, in spite of the uncertainties involved in estimating trends in rainfall, there is increasing evidence that the observed trends in rainfall in certain regions and seasons in Europe may contain a real component as well as a component due to the statistical noise of climate variability. For instance, an increasing number of rainfall indices show statistically significant trends, and these trends are to some extent consistent with numerical model results (e.g., see the analysis of observed trends and comparison with climate model projections given by the European Environment Agency, 2017). In this article, we explore how insurance risk modellers might deal with this situation in which there is emerging evidence that many of the observed trends may contain a real climate change component, but in which the magnitude of the climate change component is still highly uncertain. For risk modelling, the appropriate goal is accurate prediction, rather than evaluation of whether trends are real or not. Ideally, we would find a single general method for trend estimation that can be applied automatically to many different variables and indices, from rainfall to storm numbers, and that will give reasonable results in all cases. The challenge in designing such a method is to
find the right balance between the risk of ignoring real trends (i.e., the risk of false negatives, or type 2 errors) and the risk of over-interpreting observed trends that might be caused by the statistical noise of climate variability (which is related to false positives, or type 1 errors). This is a difficult balance to find. A secondary goal, which may to some extent be particular to the needs of the insurance industry, is to use methods that minimize the volatility of predictions from year to year, since such volatility is disruptive to some practices in the industry.

Ultimately, we would like to base our estimates of trends on three sources of information: empirical trend estimates, an understanding of decadal and interdecadal climate variability, and information on trends and climate variability from numerical climate models. Using each of the above sources involves challenges: empirical trend estimates are often highly uncertain and long-term variability can be mistaken for trends; many aspects of climate variability are only poorly understood, and numerical climate models suffer from biases that may need correction (see, e.g., the detailed discussion of climate model biases in the context of the UKCP climate projections; Met Office Hadley Centre, 2019). In the present study, we focus on methods that address the first of these challenges: that of dealing with the uncertainty in empirical trend estimates. In particular, we analyse the performance of various statistical methods that can be used to deal with the uncertainties. In the future, we hope to extend the methodologies presented here to include a more appropriate consideration of decadal and interdecadal climate variability, and to merge the estimates with information that can be derived from climate models. We will focus on seasonal total rainfall amounts only, even though changes in the whole distribution of rainfall, including extremes, are also relevant for risk modelling. We focus on seasonal totals for three reasons. First, we need to introduce a concept that is more or less unknown in atmospheric sciences: the idea that there are ways to model trends that lie in-between ignoring a trend and modelling it in full. By focusing on seasonal total rainfall amounts, we can introduce and explore this unfamiliar idea in the context of familiar models (including ordinary least squares [OLS] trends, the normal distribution and the Akaike information criterion [AIC]: see below for details). Second, there is much to learn about this topic, and by starting with a relatively tractable case we believe we can build a foundation of understanding of the issues that will help when we tackle the more difficult problem of modelling the extremes. Third, the methods we describe in this article have general applicability for modelling climate change trends in other variables. In summary, considering seasonal totals is a first step towards developing a set of methodologies that can be applied to the whole distribution and a range of different climate variables.

Some academic studies of flood modelling and climate change (e.g., Alfieri et al., 2018) use rainfall information solely derived from numerical models, or from numerical models downscaled using bias-correction schemes. These methods are appropriate if the goal is to study possible future changes in flood risk, as a function of the changes in rainfall predicted by numerical models. This approach is less common in insurance industry flood risk modelling, which has the rather different goal of obtaining as accurate as possible high resolution predictions of the year-ahead climate and related impacts, and which is typically based on analysis of observed rainfall amounts (see, for instance, the model used in Zanardo et al., 2019). One would imagine that over time there will be a convergence between the use of numerical models and observed data for understanding past, present and future rainfall behaviours, in order to benefit from the best of both approaches.

To investigate how one might model trends in past climate data, we will consider 10 methods for estimating trends. They are all based on two simple models: OLS, which is a standard way to estimate a trend, and the even simpler model that we will call flat-line (FL), which assumes that the trend is flat, and so which models the series as stationary. The 10 methods are listed below. OLS itself is the first of the methods we consider. Of the remaining nine methods, five of them are ways to choose between OLS and FL: we refer to these as model selection (MS) decision rule methods. The remaining four are ways to merge, or blend, OLS with FL: we refer to these as model averaging (MA) methods. Further details of each method are given below and in Sections 2–4. We will evaluate the methods from a variety of perspectives, with an emphasis on their ability to predict future values rather than their ability to detect the existence of trends. The methods involve a certain level of mathematical complexity, which we briefly describe. However, the complexity of the internal workings of the methods should not distract from the fundamental simplicity of the questions they are being used to answer: methods 2–6 are simply rules that one might use to decide whether to model a trend or not, and methods 7–10 are simply methods that one might use to blend together an estimated trend with a trend of zero, which reduces the trend. Furthermore, the evaluation of the methods is ultimately based on their performance in the tests we present, rather than on the underlying theoretical derivations of the methods. As a result, the methods can, if desired, be simply considered as black-boxes that give candidate answers to these two questions of whether to use a trend and how to blend trends.
The 10 methods are as follows:

1. OLS.
2. OLS/FL with statistical significance model selection (SIG-MS).
3. OLS/FL with predictive error model selection (PRE-MS).
4. OLS/FL with cross-validation model selection (CV-MS).
5. OLS/FL with Akaike information criterion (AIC) model selection (AIC-MS).
6. OLS/FL with Bayesian information criterion (BIC) model selection (BIC-MS).
7. OLS/FL with predictive error model averaging (PRE-MA).
8. OLS/FL with cross-validation model averaging (CV-MA).
9. OLS/FL with AIC model averaging (AIC-MA).
10. OLS/FL with BIC model averaging (BIC-MA).

The first two methods (OLS and OLS with significance testing) are well known in atmospheric sciences and are thoroughly covered in standard climate science textbooks (von Storch & Zwiers, 1999; Wilks, 2011), as well as general statistics textbooks. The next four methods (3, 4, 5 and 6) are model selection decision rules that we use to ask the same question as we ask with statistical significance testing (whether to use OLS or FL) but using different criteria, which may give different results. Using either cross-validation (method 4), AIC (method 5) or BIC (method 6) to decide between models is well known in atmospheric sciences and is also described in the standard atmospheric science textbooks cited above, although has not, as far as we are aware, been applied to the particular question of how to model trends in climate data. Predictive error model selection (model 3) is less well known but has been used in applied meteorology for modelling temperature trends (Jewson & Penzer, 2006). The next four methods (7,8, 9 and 10) are based on the idea of model averaging, in which different models are combined, or blended, in an effort to improve predictions. Model averaging, and in particular a form of model averaging known as Bayesian model averaging, is well known in atmospheric sciences as a method for combining weather and climate forecasts (see Wilks, 2011), but again, as far as we are aware, has not been used to model climate trends. A general textbook on the topics of model selection and model averaging is Burnham and Anderson (2010). In our situation, the main effect of applying model averaging to the OLS and FL models is to reduce the magnitude of the slope of the OLS trend towards zero, as a function, in some way, of the uncertainty around the trend estimate. The four model averaging methods determine the level of trend reduction in different ways, and give different results, which we will compare.

In Section 2, we will describe the data we will use, and define a subset of it, that we refer to as the example data set, for use in our initial examples. We then apply the first two of the 10 methods to this example data set. In Section 3, we then apply the four model selection decision rule methods (methods 3, 4, 5 and 6) to the example set. In Section 4, we apply the four model averaging methods (methods 7, 8, 9 and 10) to the example data set. In Section 5, we present results from simulation experiments using synthetic data, which we use to help understand better the theoretical behaviour of the predictive skill and volatility of the different methods. In Section 6, we give results from applying all 10 models to data for a number of regions in Europe, and back-test (hindcast) the methods to compare their performance. Finally, in Section 7, we summarize our results, draw conclusions and describe some possible further work.

2 | DATA, OLS TRENDS AND STATISTICAL SIGNIFICANCE

2.1 | Rainfall data

The rainfall data we use are from the E-obs v19.1b data set (Cornes et al., 2018). This data consists of observed rainfall amounts, interpolated onto a grid using a statistical model. This data is often used as the basis for building the statistical rainfall models used in insurance flood risk modelling and so understanding how to model the trends in this data is highly relevant. The data extends from 1950 to 2018. We will use different subsets of this data for different purposes.

We create seasonal total rainfall indices for summer and winter for eight regions across Europe, creating 16 indices in total. The eight regions are United Kingdom (UK), Spain (ES), France North (FRn), France South (FRs), Germany North (Dn), Germany South (Ds), Italy (IT) and Scandinavia (Denmark, Norway and Sweden, DKS). The regions were primarily determined using national boundaries, but France and Germany were split because of the clear differences in the rainfall trends between the north and south that are evident in the data. Aggregating into seasons and regions reduces statistical noise due to small-scale weather variability and helps estimate the trends.

In Sections 2–4, we introduce and discuss the 10 methods using an example data set consisting of UK summer total rainfall. Part of this data set (from 1979 to
2018) is illustrated by the red line in Figure 1. In Section 5, we then extend some aspects of the analysis to the other regions.

2.2 | OLS estimates of trends

We will only model linear trends. This is not because we believe that the real trends are exactly linear: they are undoubtedly not linear at some level. But as we see below, for the data we are analysing, even estimating the slope parameter in a linear trend model is in many cases quite difficult, and going beyond a linear trend model to a non-linear model would therefore be unlikely to be justifiable. Also, using non-linear trends to make sensible predictions is difficult. Furthermore, the issues around selection of detrending methodology in the presence of uncertainty, which are the main focus on this article, are most clearly illustrated using linear trends, even though they also apply to more complex models. We will model the residuals around the trends by making the commonly used assumptions that they are Gaussian, homoscedastic and uncorrelated in time. The assumptions of Gaussianity and lack of correlation in time have been verified as reasonable modelling assumptions for our data using autocorrelation and quantile–quantile plots. Although we will proceed on the basis of uncorrelated residuals, we do not believe the residuals are completely uncorrelated since there is good evidence from other sources (see, e.g., Kendon et al., 2019) that there are climate variations on interdecadal timescales that presumably do create some long time-scale autocorrelations. However, since these are not detectable in our data, they would be difficult to model within the context of this study. We will discuss the possible implications of this on our analysis in Section 7.

A standard way to estimate a linear trend from data is to use the well-known method of OLS, in which the trend is adjusted to fit the training data as closely as possible, where ‘as closely as possible’ is defined by minimizing the sum of the squared deviations between the trend and the data, calculated using the dependent variable. OLS is described in detail in textbooks such as von Storch and Zwiers (1999) and Wilks (2011). OLS gives the best linear unbiased estimate of trend slope, if the assumptions of lack of correlation of the residuals and homoscedasticity are correct. Strictly speaking, the OLS method returns estimates of the offset and slope of the trend, but not of the variance of the residuals. We add the estimate of the variance of the residuals using maximum likelihood (ML). Our parameter estimates for all three parameters (offset, slope and variance) are then equivalent to ML estimates, as we are assuming a Gaussian model with independent residuals. The equivalence with ML estimates is necessary for the use of the methods based on AIC and BIC (methods 5, 6, 9 and 10).

OLS is not the only commonly used method for estimating trends. For instance, some authors have reported that the Theil-Sen, M and MM estimators are more robust to outliers (Finger, 2013), and quantile regression can be used to estimate the median and other quantiles (Bucheli et al., 2020; Conradt et al., 2015; Dalhaus et al., 2018). However, these methods are more difficult to include in model averaging frameworks, which is the main focus of this study.

The OLS trend for the most recent 40-year period (1979–2018) from our example data set of UK summer total rainfall is illustrated as the thick black straight line in Figure 1. The trend slope is 0.877 mm/year, with standard error of 0.796 mm/year, and Figure 1 includes trend lines (blue dashed lines) at plus and minus two standard errors. We see that the standard error is roughly the same size as the trend slope, and we will see later that this trend is not statistically significant at the 5% level. Also, the trend is the opposite sign to the long-term trend predicted by some climate models for changes in UK summer rainfall (see the European Environment Agency report, 2017). There are several possible explanations for this difference. The observed trend may be due to interdecadal climate variability and may reverse in the future; or it may be simply an artefact of unpredictable shorter timescale climate variability; or it may be a genuine
anthropogenic climate change signal. Most likely, it is a combination of all of these in some unknown proportions.

These large uncertainties, the lack of statistical significance and disagreement with at least some climate models raises questions as to whether it makes sense to model the trend or not to make good predictions and risk estimates. They also make this dataset a perfect candidate for our analysis, as analyzing and understanding methodologies that could be used to deal with such difficult cases and still produce reasonable results that are not highly sensitive to the uncertainties is the goal of this article.

Although OLS is used commonly in atmospheric sciences, and science in general, for the present purpose it has some particular shortcomings, related to the large uncertainty of the trend slope estimate for the datasets we are considering. This uncertainty propagates into year-ahead predictions made using this model, degrading the quality of those predictions. When the trend slope uncertainty is large, this can lead to predictions that are less good than predictions made by ignoring the trend, even if part of the observed trend is due to a real physical phenomenon with a linear trend. This is an example of overfitting: the use of OLS may be overfitted relative to the use of the FL model with no trend. A statistical insight into this issue is that although OLS is the best linear unbiased estimator, it is not the best linear estimator. In fact, there is always an equal or more accurate (but biased) linear estimator with a smaller trend. Estimating this best linear estimator is effectively the goal of the four model averaging methods (methods 7, 8, 9 and 10).

To understand this overfitting problem intuitively, we can use an extreme case: imagine attempting to fit a trend to just 3 years of the summer total rainfall data given in Figure 1. In this case, because of the likely small size of the real trend, if one even exists (of the order of 1 mm per year) and the large size of the variability (of the order of 100 mm per year) the OLS trend slope, and any predictions made using the trend slope, would be almost entirely determined by the random noise due to climate variability in the data, rather than capturing any real underlying trend. Predictions would be wildly inaccurate and would be highly volatile from year to year as more data is added. Although the OLS trend estimate would be unbiased as an estimate of the real trend, we see from this example that unbiasedness is not sufficient to guarantee that a trend estimate is sensible to use: the uncertainty of the trend estimate is also important. This is an extreme case, but even in non-extreme cases based on more data this effect is still present, to a lesser degree, and always reduces the predictive ability of OLS trends to some extent, depending on the slope of the trend and the size of the noise. The subsequent methods we discuss are different ways for dealing with this issue of uncertainty on the trend parameter in such a way as to improve predictions and reduce volatility from year to year. Another perspective is that the model averaging methods are examples of the standard idea of ‘bias-variance trade-off’ (see, e.g., textbooks such as Wasserman, 2004) as applied to this problem, in which the bias of an estimate is increased in order that the variance decreases, with the hope of increasing the overall predictive skill.

2.3 | OLS estimates with SIG-MS

A standard method for addressing the uncertainty in trend estimates is to ask whether the trend is statistically significant or not, use the trend if it is, and ignore it and revert to the FL model if it is not. Often, a p value threshold of 5% is used to make the decision. If the trend slope parameter has a p value greater than 5% then the trend is considered not significant, while if it has a value less than 5% it is considered significant. There has been much discussion about the validity of this method, how to interpret it, and the potential shortcomings (for a fairly definitive discussion by statisticians, see Figure 2).
Wasserstein, 2016), but it remains a useful quick indicator of whether we should consider a trend likely to be real or not (i.e., whether we can detect the trend). The red line in Figure 2 shows 30 versions of year-ahead predictions, which use statistical significance at 5% to decide whether to model the trend or not in sliding-window subsets of our example data set. The first prediction is based on data from the 40-year period from 1950 to 1989, and is a prediction for 1990; the second prediction is based on data from the 40-year period from 1951 to 1990 and is a prediction for 1991, and so on. The horizontal axis on the graph shows the year being predicted. The black dots show predictions made using the same 40-year periods using FL.

We see that the prediction from OLS with statistical significance testing is rather stable from 1990 to 2012. This is because the trend is never statistically significant during this period, and so all the predictions are based on the FL model using 40 years of data. Then, in 2013, the prediction jumps up by around 15%. This is because the trend from 1973 to 2012 is positive and statistically significant, and the prediction for 2013 is based on that trend. In 2014, however, the prediction jumps back down. This is because the trend for 1974–2013 is not statistically significant. From 2014 to the end of the series, the trends are not statistically significant and so the predictions are again based on the FL model. Ignoring for a moment the 1 year where OLS is used, the FL predictions are generally decreasing up to 2007, and increasing thereafter. This is perhaps an indication of the presence of part of a cycle of interdecadal variability.

With our focus on year-ahead prediction, rather than the detection of trends, this example highlights that significance testing has certain disadvantages, as follows:

a. The use of a p value of 5% risks ignoring emerging evidence of the existence of a trend (a false negative, or type 2, error). Five percent can be appropriate for making scientific claims, in the sense that one would not want to assert that there is a trend in a dataset that has previously been considered stationary unless the evidence is sufficiently strong. But it is less appropriate for insurance risk modelling, or risk modelling in general. If risk modellers use this method when a trend is real and gradually emerging in the data, then they risk ignoring the trend for too long. The value of 5% is also arbitrary, rather than based on any kind of modelling principle.

b. Risk modelling is, by definition, about identifying the range of possibilities, which is a different goal from that of trend detection and scientific proof. If there is evidence for a trend, whether statistically significant or not, then the trend should be considered as a possibility and included in the risk analysis in some appropriate way with an appropriate weighting.

c. Applied every year to decide whether to model a trend or not, significance testing can lead to predictions and risk estimates that are highly volatile from year to year, since a trend might be not significant (and hence considered zero) in one year, and then significant the next, as we see in Figure 2. This volatility is simply an artefact of the decision to use a method that switches the trend on or off in a binary fashion at a certain arbitrary level, and for some applications, including insurance risk modelling, could be highly disruptive.

d. We have additional information that there are likely trends in the data, above and beyond the information contained in the data itself. A basic argument for the existence of trends in any climate data is that since major aspects of the climate are changing, and everything in the climate is highly interconnected, it is implausible that any aspect of the climate would have precisely no trend. Rainfall specific arguments would be that increasing temperatures may lead to increased water vapour in the atmosphere, which may lead to greater rainfall (see, e.g., the discussion in Skliris et al., 2016), and also that numerical model simulations show trends in rainfall in many regions.

To overcome the first objection above, one could consider using a higher threshold for the p value, such as 10% or more. This would partly resolve the concern that the 5% level typically used for scientific proof may be too strict, although the choice of threshold would still be arbitrary. It would not, however, overcome the other objections.

3 | DECISION RULE METHODs

Significance testing can be described as a model selection decision rule that determines whether one should model a trend or not. We now consider four other decision rules that can be used as a way to make this decision, but based on different criteria, and that avoid the arbitrary selection of a threshold. Since we consider four such methods, the choice of which method to use is still to some extent arbitrary, but now only over a small and finite set of choices, and other considerations can be brought to bear that can help choose between the methods (including the results shown later in Sections 5 and 6). The methods involve some detailed statistical analysis but in the end are all simply methods for answering the question: should we model the trend or not?
3.1 | OLS with PRE-MS

The first alternative decision rule we consider is OLS with predictive error model selection. The mathematical derivation of statistical significance testing of a linear trend (as covered, e.g., in Wasserman, 2004) does not take into consideration the need to predict future data values, and as a result when statistical significance is used as a decision rule for whether to model a trend or not there is no reason to think it will lead to a choice which is optimal in terms of predictive skill. An alternative approach is to compare the skill of the OLS and FL predictions mathematically. This leads to the simple rule that predictions made by modelling the trend will be more accurate (lower predictive mean squared error [PMSE]) on average than predictions made by ignoring the trend if and only if the absolute value of the real trend is larger than the standard error on the trend (Jewson & Penzer, 2006). In practice one only has estimates of the trend and the standard error on the trend, and so a practical decision rule can only be based on comparing the estimate of the absolute value of the real trend and the estimate of the standard error on the trend.

Figure 3a shows 30 predictions made by applying this simple predictive error decision rule to our example data set. The black dots show predictions from the FL model (as in Figure 2) and the black crosses show predictions from the OLS model. The OLS predictions are below the FL predictions from 1990 to 2005, and then above from 2006 onwards. This reflects a negative trend in the early part of the data, and a positive trend in the latter part of the data. The red line shows predictions from the predictive error model selection decision rule, which either takes the same value as the FL prediction, or the same value as the OLS prediction. The first six predictions, from 1990 to 1995, use FL. The predictions for 1996 and 1997 switch to using OLS with a negative trend. The predictions from 1998 to 2008 use FL, and the predictions from 2009 to 2019 use OLS with a positive trend. We see that relative to using statistical significance testing with a 5% confidence level as a decision rule (Figure 2), this method chooses the OLS model more often. In other words, there is a range of intermediate trend values that are not statistically significant at the 5% level, and in which the trend would be ignored by statistical testing, but which would lead to this method choosing the trend model for prediction.

3.2 | OLS with CV-MS

The second alternative decision rule we consider is OLS with cross-validation model selection. Cross-validation is a commonly used method for the evaluation of models (see textbooks such as Wilks, 2011; von Storch &
Zwiers, 1999; Wasserman, 2004). It compares models in terms of how well they can predict the past data in an out-of-sample sense. For our decision as to whether to model a trend or not, one can use leave-one-out cross-validation as follows. Given the N data points that we are using to model the trend, we split them into N-1 training points and 1 validation point. We then fit both the OLS trend model and the FL model to the N-1 training points and use both to make predictions of the validation point. We calculate errors for both. We then repeat this N times for the N different ways of splitting the data and calculate the PMSE of the N predictions from both models. The model with the lower PMSE is chosen as the better model and is used for prediction.

Figure 3b shows 30 predictions made using the cross-validation model selection decision rule. The picture is somewhat similar to that shown in Figure 3a, except that cross-validation does not choose OLS in 1997, 2009 or 2019. We see that relative to using the predictive error criterion, cross-validation appears to choose OLS slightly less frequently.

### 3.3 | OLS with AIC-MS

Another method for the evaluation and selection of models is the AIC (see explanations in textbooks such as Wilks, 2011; von Storch & Zwiers, 1999; Wasserman, 2004, and a more thorough treatment in Burnham & Anderson, 2010). AIC is a scoring system for models that have been fitted using ML, in which different models fitted to the same dataset are compared according to the value of likelihood achieved at the maximum, penalized by the number of parameters to compensate for the fact that models with more parameters will inevitably achieve higher likelihoods. The formula for the basic AIC for one model is:

$$AIC = 2k - 2\ln L$$

where AIC is the AIC score, $L$ is the likelihood achieved at the maximum, and $k$ is the number of parameters that were adjusted to maximize the likelihood in that model. The sign convention is such that lower values indicate better models.

We can calculate AIC values for fitting OLS (which has three parameters: offset, trend and variance of residuals) and for fitting FL (which has two parameters: offset and variance of residuals). The linear trend model would always be expected to achieve a higher value of likelihood at the maximum, because it has more parameters, but in the AIC score it is penalized for the extra parameter and as a result which model achieves the best (lowest) AIC score depends on the data-set.

For univariate linear Gaussian models, of which OLS is an example, there is a higher-order approximation to AIC known as AICC:

$$\text{AICC} = \text{AIC} + \frac{2k^2 + k}{n - k - 1}$$

where $n$ is the number of data points.

Since we know $k$ and $n$ for our two models, we can calculate the penalty terms used in AICC (i.e., the part of the AICC that penalizes for the number of parameters) as $2k + (2k^2 + k)/(n - k - 1)$. For FL ($k = 2, n = 40$) and OLS ($k = 3, n = 40$) this gives 4.27 and 6.58, respectively. The difference is 2.31, which is the additional penalty applied to the OLS model because it has one more parameter.

Figure 3c shows 30 predictions from the AICC-based model selection decision rule. This method gives similar results to the previous two methods but chooses OLS slightly less often than both. Relative to cross-validation, it differs by not choosing OLS in 2017.

### 3.4 | OLS with BIC-MS

The final method for the evaluation and selection of models that we consider is the BIC. Like AIC, BIC is a scoring system for models that have been fitted using ML. It is an approximation to the Bayesian probability that the model is correct, given the observations. The AIC can also be considered as an approximation to the same probability, but using different Bayesian priors, and neither AIC nor BIC is more Bayesian than the other in spite of the name BIC (see the discussion in Burnham & Anderson, 2010).

The formula for the BIC is:

$$\text{BIC} = \ln(n) k - 2\ln L$$

For our FL and OLS models, the penalty terms (given in this case by $\ln(n) k$) are 7.38 and 11.1 respectively. The difference is 3.69, which is the additional penalty applied to the OLS model because it has one more parameter. That the absolute size of BIC penalty terms is larger than the AIC penalty terms is irrelevant in terms of which model is selected, since AIC and BIC scores are never compared with each other. But that the difference in the penalties for the two models is larger for the BIC model is relevant and indicates that BIC is more likely to choose the FL model than the OLS model, relative to AIC, in this case.

Figure 3d shows 30 predictions from the BIC-based model selection decision rule. This method chooses OLS
markedly less often than the previous three decision rules, and only uses OLS in 2013 and 2015. It is thus choosing OLS only once more than statistical testing, which chose OLS only in 2013.

3.5 Discussion of model selection methods

The four model selection decision rule methods described above ask whether we should be modelling a trend or not. They have the advantage, over using statistical significance testing as a decision rule, that they do not involve an arbitrary threshold of significance. The four methods have similar intentions, but realize them in different ways, and give different results. Comparing Figures 2 and 3 we see that all four model selection methods choose the OLS model more frequently than statistical testing. The predictive error, cross-validation and AICC methods give relatively similar results. The BIC method stands out from the other methods as choosing the OLS model the least often by a wide margin and is closer to statistical testing than it is to the other models.

These four methods go some way to improving on statistical testing. However, these methods are still not ideal from our point of view. In any of the four methods if OLS is selected by only a small margin then the estimates from FL are completely ignored (and vice versa). In that situation, good risk management should instead merge the results of FL into the analysis, but with a lower weighting, and widen the uncertainty. Also, by using a binary switch between models, the decision rules introduce volatility into predictions and create situations in which one additional data point can cause sudden changes in predictions, as we see very clearly in Figure 3 in several places. For many applications, including insurance pricing and risk management, this is undesirable.

Based on these considerations, in Section 4 below we consider model averaging methods that avoid using a binary switch or decision rule altogether, and instead create predictions that interpolate smoothly between trend and no-trend models with appropriate weights on each. The models differ in terms of how they do that interpolation.

3.6 Bayesian model selection methods

The analysis in this article takes place in the context of OLS and ML, which are non-Bayesian methods. Similar analysis can be done using Bayesian methods, with all the usual advantages and disadvantages that brings (see Wasserman (2004) or Efron and Hastie (2016) for a discussion of the issues). In a Bayesian analysis, model selection (and model averaging) can be performed using Bayes’ Factors (Burnham & Anderson, 2010) or the WAIC (Watanabe, 2010). We have not extended our study to include Bayesian methods for the specific reasons that (a) the issues we are addressing can be understood and resolved within the OLS and ML framework, that many will be familiar with and (b) Bayesian analysis is more complex, and that complexity might obscure the essential simplicity of the problems we are trying to solve, and the simplicity of the solutions we are exploring.

4 MODEL AVERAGING METHODS

We now consider four methods for model averaging. Once again the methods involve some detailed statistical analysis, but in the end are all simply methods for coming up with a trend slope which is a compromise, or blend, between the OLS trend and no trend. These methods correspond to the four model selection decision rules discussed above, but now repurposed for model averaging. All four methods model the distribution of future seasonal total rainfall as a mixture distribution created by combining the distributions estimated from the OLS and FL models. The mixture is defined by the mixing weights \( w_1 \) and \( w_2 \), and the only differences between the four models are the different values for these weights.

If the means and variances of the two predictions being combined are \( m_1, m_2, v_1 \) and \( v_2 \), then, using the standard expressions for the mean and variance of a mixture distribution, the mean \( m \) of the prediction is given by

\[
m = w_1 m_1 + w_2 m_2
\]

and the variance \( v \) of the prediction is given by

\[
v = w_1 v_1 + w_2 v_2 + w_1 (m_1 - m)^2 + w_2 (m_2 - m)^2
\]

where the last two terms on the right-hand side represent the additional uncertainty in the predictions which is created by the uncertainty of which of the two models to use, which increases to the extent that the two models give different predictions.

4.1 OLS estimates with PRE-MA

The first model averaging method that we will consider uses a method known as parameter shrinkage. The
intuitive idea behind parameter shrinkage is that if there is a parameter in a statistical model that is too important not to include, but at the same time too uncertain to actually use for prediction, then the estimate of the parameter can be reduced towards zero, using some criterion to define an optimal reduction. This reduction allows the parameter to be included, in some partial form, and so is intended to be preferable to either including or not including the parameter. Parameter shrinkage has a long history in statistics (see, e.g., Copas, 1983) but the only example we are aware of in climate science is that given in Jewson and Penzer (2006), which we use below.

In our case, the criterion for parameter reduction will be to minimize the PMSE of the year-ahead prediction, under the assumption that there is a linear trend. Minimizing the PMSE of the year-ahead prediction is minimization of an out-of-sample prediction and is hence different from the in-sample minimization of MSE used to fit the OLS trend line in the first place. Following Jewson and Penzer (2006), for the OLS model, minimum PMSE predictions are achieved by reducing the trend slope by a factor of

$$r = \frac{\beta^2}{\beta^2 + \sigma^2_{\beta}}$$

where $\beta$ is the trend slope and $\sigma^2_{\beta}$ is the square of the standard error on the trend slope. From this formula, we can see that when the standard error term $\sigma^2_{\beta}$ is small relative to the trend slope $\beta$, $r$ is close to 1 and the OLS trend is scarcely reduced. When the standard error is large relative to the trend slope, $r$ is close to 0 and the reduction in the OLS trend is large. If the optimal reduction factor $r$ could be calculated exactly, then the resulting predictions would always on average be better than OLS predictions. In practice, however, the trend slope and the standard error of the trend slope are only estimated from the data, and we have to use estimates in the above expression. As a result, the optimal reduction itself is only estimated.

Figure 4a shows 30 year-ahead predictions made using predictive error model averaging (solid red line). All four of the model averaging prediction methods generate predictions that lie on or between the FL and OLS predictions, by definition. In the case of predictive error model averaging, we see that the model averaged predictions vary from being close to FL to being close to OLS, for different years being predicted. This is because the $r$ shrinkage factor varies as the estimated slope, and uncertainty on the estimated slope, vary. For instance, for the latter part of the period, the predictions are closer to OLS, reflecting a larger and more certain positive trend near the end of the series. Comparing with Figure 3a we can see that model averaging gives less volatile predictions than model selection, as expected. The dramatic jumps between 1995/1996, 1997/
1998 and 2008/2009 are smoothed out, but not eliminated completely.

### 4.2 OLS estimates with CV-MA

The second model averaging method we consider also uses parameter shrinkage for the trend slope, as in the previous section, but this time we estimate the magnitude of the trend reduction not from theoretical considerations but by using cross-validation. This is more laborious but perhaps has the advantage that it avoids abstract mathematical assumptions. We calculate the reduction of the trend slope as follows. Given the N data points we are using to estimate the trend, we split them into N-1 training points and 1 validation point. We then fit a trend to the N-1 training points, reduce it by a factor, and predict the validation point. We repeat this N times, for the N different ways of splitting the data, and calculate the PMSE of the N predictions. That PMSE is a function of the reduction factor. We then vary the reduction factor, repeat the whole process many times, and find the value of the reduction factor that gives the lowest PMSE.

Figure 4b shows 30 predictions from cross-validation-based model averaging. Comparing with Figure 4a, we see that cross-validation gives somewhat similar results, although the predictions lie closer to those from the FL model. The three jumps referred to above are also present in Figure 4b but are a little smaller.

### 4.3 OLS estimates with AIC-MA

The third model averaging method that we will consider is based on the AICC score as used above for model selection, but now converted into weights for model averaging. This use of the AICC scores to create weights is well known generally in statistics (see, e.g., Burnham and Anderson, 2010) but we are not aware of it having ever been used in atmospheric sciences. The AICC scores are converted to weights on the models in the following steps:

a. Define the minimum of the two AICC scores as \( \text{AICC}_{\text{min}} \).

b. Calculate the deviations of each of the two scores from this minimum, \( DAIC = \text{AICC} - \text{AICC}_{\text{min}} \).

c. Calculate the relative likelihood of each of the two models as \( RL = \exp(-DAIC/2) \).

d. Normalize the two relative likelihoods so that they sum to one, to create weights. These weights can be interpreted as estimates of the probability that each model is the model that minimizes the Kullback–Leibler metric that AICC is based on (Burnham & Anderson, 2010).

Figure 4c shows 30 predictions made using AICC-based model averaging. Comparison with the previous two model averaging methods shows that the jumps are further reduced, and the curve is further smoothed.

### 4.4 ML estimates with BIC-MA

The fourth and final model averaging method we consider is the BIC analogue of AIC-MA. Weights are derived from the BIC values in exactly the same way that the weights for AIC were derived above.

Figure 4d shows 30 predictions from BIC-based model averaging. The predictions tend to be closer to those from the FL model than those of the other three model averaging methods. Of the four model averaging methods, BIC clearly gives the least volatile, smoothest results. The large jumps are very much smoothed out.

### 4.5 Discussion of model averaging methods

To illustrate the model averaging methods more clearly in practice, the modelled trends for our example data set for the period 1979 to 2018 are shown in Figure 5. This figure shows the same rainfall data as Figure 1 (thin red line) and the OLS trend (thick black line) is also the same as in Figure 1. The additional coloured lines now show the trend estimates from the four model averaging methods: PRE-MA (purple), CV-MA (dark blue), AIC-MA (light blue), BIC-MA (red). The vertical axis has been stretched relative to Figure 1 to show clearly the differences between these lines. We see that all four methods reduce the OLS trend significantly towards zero, because of the large uncertainty. The relativities between the methods in this single example are not the same as the relativities that we see in Figure 4, or that we will see later in the simulation tests. This highlights that the relativities between the results from the methods may be different when considered on average and for individual cases.

The model averaging methods improve over model selection in various ways. They incorporate aspects of both FL and OLS models in each prediction, which is appropriate for good risk management. They also reduce the volatility of predictions as we have seen in Figure 4. Whether they improve predictive performance on average, and when, is investigated in Sections 5 and 6 below.
In the previous sections, we have described 10 methods for making year-ahead predictions of seasonal total rainfall and applied them to an example dataset. This has shed light on the characteristics of the predictions generated by the different methods, particularly in terms of (a) their volatility from year to year and (b) the similarities and differences of their predictions relative to FL and OLS. However, thus far we have learnt nothing about when the predictions are accurate in terms of PMSE. This will be addressed in this and the next section.

5 | SIMULATION TEST RESULTS

In the previous sections, we have described 10 methods for making year-ahead predictions of seasonal total rainfall and applied them to an example dataset. This has shed light on the characteristics of the predictions generated by the different methods, particularly in terms of (a) their volatility from year to year and (b) the similarities and differences of their predictions relative to FL and OLS. However, thus far we have learnt nothing about when the predictions are accurate in terms of PMSE. This will be addressed in this and the next section.

5.1 | Simulation setup

In this section, we describe the results of simulation tests that allow us to compare our 10 prediction methods in a controlled environment in which we can generate as much data as is needed to distinguish between the methods and to understand their potential predictive skill. Our tests work as follows. We simulate 10,000 synthetic time series of 41 data points, using a linear trend plus white noise. The first 40 years are analogous to our example data set of 40 years of UK summer total rainfall. The 41st year in each series corresponds to the next year, which we are trying to predict. We then use the first 40 data points to predict the 41st year using the 10 prediction methods described above and analyse the resulting predictions in terms of PMSE and volatility. We repeat this exercise for various levels of real trend, ranging from 0% to 6% of the standard deviation of the residuals.

Simulation exercises such as this require careful interpretation for two reasons. First, the results cannot be used as a look-up table to decide which trend method to use for different sizes of observed trend, since the trend used in the simulation study is the real (usually unknown) trend, rather than the observed trend. Second, they are all in the context of a perfectly linear trend with perfectly uncorrelated Gaussian residuals: they contain no interdecadal climate variability (except as a residual of the year-to-year white noise variability) and the trend is constant from the start to the end of the 40 years. Nevertheless, these tests do shed some light on the general properties of the different prediction methods tested, as we will see below.

The results from these simulation experiments are a function of the number of years of data used to fit the trends: in our case 40 years. If this is changed, the results change, although only qualitatively: the relativities of the different methods remain the same.

5.2 | General results

Figure 6, parts (a) and (b), shows the predictive root mean square error (PRMSE) values for the 10 different prediction methods versus the size of the real trend expressed as a percentage of the noise around the trend. PRMSE values are also expressed relative to the noise in the trend model, which is 100. We will explain each line in this figure in turn, starting with the general patterns.

The steep dashed black lines are the same in both panels of Figure 6 and show the PRMSE for the FL model. For very small trends of up to around 1% this model works well and is the best of the models tested. However, for larger trends, the performance of the model deteriorates rapidly, and beyond around 1.8% it becomes the worst model. One can conclude that it can only be used in situations in which one is absolutely confident that the real trend is either zero or very small. It is therefore not suitable as a general method that could be applied to a range of datasets with uncertain trends, since some may indeed contain a real trend larger than 1.8%.

The horizontal thick black lines are the same in both panels of Figure 6 and show the PRMSE for the OLS model. For very small trends of up to around 1% this model works well and is the best of the models tested. However, for larger trends, the performance of the model deteriorates rapidly, and beyond around 1.8% it becomes the worst model. One can conclude that it can only be used in situations in which one is absolutely confident that the real trend is either zero or very small. It is therefore not suitable as a general method that could be applied to a range of datasets with uncertain trends, since some may indeed contain a real trend larger than 1.8%.

The horizontal thick black lines are the same in both panels of Figure 6 and show the PRMSE for the OLS model. The PRMSE is constant as a function of the slope of the trend, and is determined by the noise in the model, plus a
small factor due to parameter uncertainty on the regression parameters. For very small trends, up to around 1%, OLS is the worst of the models considered, while beyond around 2% it is the best. One can conclude that it can only be used in situations in which one is absolutely confident that the real trend is reasonably large. Like FL, it is therefore also not suitable as a general method that could be applied to a range of datasets, since some may not contain a real trend, or may contain only a small real trend.

The thin black lines are the same in both panels of Figure 6 and show PRMSE results for statistical significance testing. The coloured lines show results for four model selection methods. In panel (a), the coloured lines show results for four model averaging methods. In both panels, the model selection and model averaging models are as follows: Predictive error (purple), cross-validation (dark blue), AIC (light blue), BIC (red) they all give results that are worse than OLS but better than FL. Between 1.2% and 1.4%, there is a region in which the model selection methods (the coloured lines in Figure 6a) all perform worse than both OLS and FL. Over the same region the model averaging methods (the coloured lines in Figure 6b) all perform better than both OLS and FL.

5.3 | Results for individual methods

We now consider the performance of the individual model selection and model averaging methods, starting with statistical significance testing. For small trends, up to around 1%, this method works well, but for larger trends, in the range 2%–4%, it performs badly relative to other model selection and model averaging methods because it is reluctant to select OLS.

The purple lines show the PRMSE for predictions from the predictive error model selection (panel (a)) and predictive error model averaging (panel (b)) models. Between 1.2% and 1.9%, predictive error model averaging is the best of all the models. Up to around 3% predictive error model averaging beats predictive error model selection. Neither predictive error model selection nor predictive error model averaging performs particularly well for very small trends, because they select OLS too often.

The dark blue lines show the PRMSE for predictions from cross-validation model selection and cross-validation model averaging, while the light blue lines show the PRMSE for predictions from AIC model selection and AIC model averaging. These two methods show almost exactly the same predictive performance and so we describe their results together. Up to around 3% model averaging beats model selection for both models. Model averaging for both models beats predictive error model averaging for the smallest trends up to around 0.5%, but predictive error model averaging beats both models thereafter.

The red lines show the PRMSE for predictions from BIC model selection and BIC model averaging. Up to around 4% BIC model averaging beats BIC model selection. For small trends up to around 1% BIC model averaging is the best of the compromise methods (the model selection and model averaging methods) but it does less well for larger trends of around 3%. BIC model selection follows the statistical significance testing results very closely.

Based on the above results, we can make the following general comments (all specific to fitting trends to 40 years of data):

1. For datasets with small real trends, definitely less than 1%, FL will work best.
2. For datasets with large real trends, definitely greater than 2%, OLS will work best.
3. It is impossible to say which method is best overall: how well the methods will perform on real datasets depends on the range of real trend values in those datasets. For datasets with a wide range of real trend values the compromise methods may work better than either FL or OLS because they avoid the poor performance of FL and OLS at one or other of the extremes.

4. For small trends, up to around 3%, model averaging beats the corresponding model selection method in each case (and vice versa for large trends).

5. The results from the different methods form a clear sequence: the results from statistical testing and BIC are closest to the results from the FL model, because they do not choose OLS very often, or do not put much weight on OLS; the results from the predictive error methods are closest to the results from the OLS model because they choose OLS more often or put more weight on OLS; the results from the AIC and cross-validation models are in-between.

Figure 7 shows the corresponding volatilities of each method, measured as the standard deviation of the prediction (as opposed to standard deviation of the prediction error, which is given by the PRMSE values in Figure 6). OLS and FL show constant volatilities determined simply by the noise in the data. The other patterns are somewhat similar to the patterns in PRMSE, although shifted to higher values of trend. Overall, we see that the model averaging methods give markedly lower volatilities than the model selection methods, for trends up to around 3%. BIC model averaging has the lowest volatility for small trends from 0% to 2%.

6 | EUROPE WIDE RESULTS

We now apply the 10 prediction methods described above to our 16 rainfall data indices. Applying 10 methods to 16 data sets gives 160 different combinations. For each of these 160 combinations, we back-test the method using the period 1950–2018, for summer and winter separately, as follows. We fit each model to 40 years of data and then predict 1 year ahead, using the 29 40-year fitting periods that start with 1950–1989, and that finish with 1978–2017. We compare the predictions with the real outcomes, and we then calculate a single PRMSE for each of the 160 combinations based on the 29 tests.

This method for testing the models has two clear limitations. First, the number of tests is small (only 29 per index) and strongly influenced by noise in the data. Second, the trends over the testing period are assumed constant, and hence we are averaging over situations in which for example FL might do well in the earlier period and OLS might do well in the later period as the effects of climate change get stronger. Nevertheless, it is a simple way to assess the different methods in a real situation.

The results for the 160 tests are shown in Figure 8 for summer and Figure 9 for winter. The horizontal axis shows the eight regions, ranked by increasing magnitude.
of the trend calculated over the entire 69-year data period, from small trends to large trends. The values of the magnitude of the trend over the entire period are shown at the top of the plot. This ranking means that the horizontal axis is roughly consistent with the horizontal axes in Figures 5 and 6. The vertical axis shows the PRMSE of predictions expressed relative to the PRMSE of FL and as a percentage of the variance of the residuals. The results for the FL model are given by the horizontal black line. The results for the other 10 models are given by the various circles, according to the key in the figure. Model selection methods have open circles, and model averaging methods have closed circles.

We now discuss the results, method by method, in both Figures 8 and 9. FL performs the best in 5 out of 16 of the cases (Dn summer, Dn winter, FRn winter, ES winter, IT winter). We can see this in the figures when the circles for the other 10 methods are all above the horizontal line that represents the FL results. FL does well for these regions because the trends in these cases are close to zero. However, in a number of other cases FL does not perform particularly well and is among the worst performing (UK summer, IT summer, DKSN summer, UK winter, DKSN winter), because in these cases there are larger trends. Overall, FL is one of the best methods.

OLS results are shown as black dots. In one case, OLS is the best performing method (UK summer), while in another case, OLS is the worst (FRn winter). In many cases it is one of the less well performing methods. This is because in these datasets the trends, where they exist, are generally rather weak. We can conclude that OLS is not a good general method that can be applied across the board for these datasets.

The results from the four model selection methods are shown as open circles. They are typically among the worst of the methods, although not in every case. It is difficult to see any particular pattern with respect to relativities between the four methods.

The results from the four model averaging methods are shown as filled circles. They perform well in all cases and perform rather similarly to each other. In cases with weak trends, BIC-MA performs the best of the model averaging methods. Overall, FL and BIC-MA are the best performing methods. BIC-MA performs better than FL in summer, and FL performs better than BIC-MA in winter.

The results are consistent with the results from the simulation tests discussed in the previous section: (a) FL and OLS work well in the two extreme situations of small and large trends; (b) for small trends, model averaging beats model selection; and (c) model averaging does reasonably well in all cases.

As we learnt from the simulation results, the relative performance of the different methods is very much driven by the range of underlying real trends in the data. Most of the trends in the datasets we are using are in the range in which FL and model averaging, and in particular BIC-MA, do well. If the underlying real trends were larger, then other models would do better. These results are based on data from 1950 onwards. The effects of climate change would be expected to be larger more recently, which is perhaps a reason to expect that BIC-MA would outperform FL on more recent data.

7 | SUMMARY AND CONCLUSIONS

Climate change is affecting different climate variables in different ways. Temperature and sea level show clear and unambiguous upward trends in many parts of the world, and in many cases the trends easily pass tests of statistical significance. For climate risk modellers, there is no doubt that such trends should be accounted for when adjusting past data to make it relevant for the present or near future. Other climate variables, such as rainfall, show more ambiguous trends. In some regions and seasons there are statistically significant trends in rainfall, while in others there are not. Additional information about rainfall trends comes from numerical models. Such information is valuable, but it is hard to know exactly how much weight to place on the information given the possibility of model biases, and the disagreements between different models. For these reasons, climate risk modellers face difficult decisions when deciding what rules to apply.
to decide whether to and how to account for climate change trends in historical rainfall data in their analyses.

To start exploring this topic, we have considered a number of methods for dealing with uncertain trends, in the context of European rainfall. We have focused on empirical methods for estimating linear trends and ignored both numerical model output and the issues created by long-term climate variability, as a first step. Starting with the well-known ordinary least squares (OLS) model for fitting trends, we have explored the properties of statistical significance testing, which can be used to decide whether to model a trend or revert to the simpler flat-line (FL) model, which ignores the trend. We have then looked at four model selection decision rules that can be used in a similar way to determine whether to fit a trend or use the FL model. The models use criteria based on predictive skill, cross-validation, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). All four models choose to model the trend sooner (i.e., when the trend is smaller) than statistical significance testing at 5% would.

We have then explored four methods for combining the OLS and FL models to produce a single reduced-trend prediction. The four models we consider use different, but reasonable criteria for making the combination, again using predictive skill, cross-validation, the Akaiake information criterion and the Bayesian information criterion.

We have compared the resulting 10 methods in three ways: by looking at the qualitative character of their predictions in an example data set; by applying them to thousands of years of simulated data and evaluating their predictive skill; and by using out-of-sample back-testing on observed rainfall data.

The qualitative comparison shows that all five model selection methods (including regular statistical testing) produce predictions that are liable to make sudden jumps from year to year, and back again. This is because they make a binary choice between using FL and OLS. For many applications, including insurance risk modelling, this is undesirable. The model averaging methods avoid this problem.

The simulation tests generate a wealth of information about the relative performance of the methods as a function of the real (but usually unknown) trend. Based on the simulation results, one cannot say that one method is better than the others, since the methods perform differently for different sizes of the real trends. However, there are some specific conclusions from the simulation tests, most notably that (a) for trends that are known to be small, FL will work best, (b) for trends that are known to be large, OLS will work best, (c) for trends that are in-between, or uncertain, one of the compromise models based on model selection or model averaging may well work best, (d) for small trends model averaging beats model selection in terms of predictive error and volatility, (e) the predictive error methods give results closer to the OLS results, while the BIC methods give results closer to the FL results. The AIC and cross-validation methods give very similar results, and these results lie in-between those of the predictive error and BIC methods, and finally (f) statistical testing with 5% level of significance gives almost identical results to BIC model selection.

The results from back-testing on real rainfall data indicate that no single method performs best or worst, because of the short testing period, and because the underlying trends are different across our different datasets, and different methods will perform well or badly depending on the sizes of those trends. However, certain patterns are apparent, which are consistent with the results from the simulation experiments, and the fact that the trends in the datasets being tested are relatively small (all less than 2%). FL and OLS do well in certain cases (when the trends are small and large respectively) but fail in other cases (when the trends are large and small respectively). Model averaging methods beat model selection methods and perform reasonably well in all cases.

Our overall conclusions from the various analyses are as follows:

1. Risk modellers should carefully consider their strategy for how they deal with emerging trends in historical climate data. Various strategies are possible, and some have better properties than the obvious strategy of ignoring trends until they become statistically significant, which can lead to large year-to-year changes in forecasts, late identification of emerging trends and low predictive skill. Different strategies have different advantages and disadvantages that should be understood before making a choice.
2. One aspect that should affect the decision of which strategy to use is the likely range of trends in the data being considered, which strongly influences the performance of different methods.
3. Model averaging methods, in particular, perform well across a range of trend sizes, and produce accurate predictions with low volatility.

There are a number of shortcomings in our analysis. The first is the likely presence of climate variability on interdecadal timescales. This has not been accounted for in our modelling. In our European rainfall example, we
did not model residuals as auto-correlated, because the observed auto-correlations did not indicate the presence of significant non-zero values. However, we know from longer records of European rainfall, and other studies, that interdecadal climate variability does affect European rainfall. Understanding the impact on recent observed trends is difficult, but in general we know that not modelling it likely overestimates the sizes of trends and underestimates their uncertainty.

Another shortcoming is that we have only considered total rainfall, and only aggregate regions and seasons. Risk modellers would ideally have models of rainfall trends for both the total, and other aspects of the distribution of variability, that vary continuously in space and by season. This is a significantly more challenging task than the modelling we have attempted above. Previous authors have used quantile regression to estimate trends in quantiles (e.g., to estimate trends in drought indices [Bucheli et al., 2020] and agricultural yield [Bucheli et al., 2020; Conradt et al., 2015]), and developing methodologies that combine the ideas of model averaging presented here with quantile regression would seem like a promising approach to address the challenge of modelling changes in the distribution.

Our future plans are to extend the ideas described above to the entire rainfall distribution, including rainfall extremes, to methods that incorporate interdecadal variability, potentially using non-linear trends or numerical model output, and to more complex models that can capture continuous variation of trends in space and time. We believe that the general concepts we have presented, with the emphasis on considering the level of evidence for trends, and focusing on predictive skill and volatility, rather than attempting to prove or disprove the existence of trends using statistical testing, may also be helpful for the modelling of many of the other climate indices used in climate risk modelling. Statistical testing remains important, of course, for the purpose for which it was designed.

ACKNOWLEDGEMENTS
The authors would like to thank Enrica Bellone, Jo Kaczmarska and Christos Mitas for interesting discussions related to this topic, and two anonymous reviewers and the editor for helpful suggestions that have improved the article.

AUTHOR CONTRIBUTIONS
Stephen Jewson: Conceptualization (lead); formal analysis (lead); investigation (lead); methodology (lead); project administration (lead); software (lead); supervision (lead); visualization (lead); writing – original draft (lead); writing – review and editing (lead). Tanja Dallafior: Formal analysis (supporting); investigation (supporting); visualization (supporting); writing – review and editing (supporting). Francesco Comola: Data curation (lead); writing – review and editing (supporting).

CONFLICT OF INTEREST
Tanja Dallafior and Francesco Comola work for RMS Ltd., a company that builds catastrophe risk models for the insurance industry.

ORCID
Stephen Jewson https://orcid.org/0000-0002-6011-6262

REFERENCES
Alfieri, L., Dottori, F., Betts, R., Salamon, P. & Feyen, L. (2018) Multi-model projections of river flood risk in Europe under global warming. Climate, 6(1), 6.
Bucheli, J., Dalhaus, T. & Finger, R. (2020) The optimal drought index for designing weather index insurance. European Review of Agricultural Economics, 48, 573–597.
Burnham, K. & Anderson, D. (2010) Model selection and multimodel inference. New York: Springer-Verlag.
Conradt, S., Finger, R. & Bokusheva, R. (2015) Tailored to the extremes: quantile regression for index-based insurance contract design. Agricultural Economics, 46, 537–547.
Copas, J. (1983) Regression, prediction and shrinkage. Journal of the Royal Statistical Society, Series B, 45(3), 311–354.
Cornes, R., van der Schrier, G., van den Besselaar, E. & Jones, P. (2018) An ensemble version of the E-OBS temperature and precipitation datasets. Journal of Geophysical Research – Atmospheres, 123(17), 9391–9409.
Dalhaus, T., Musshoff, O. & Finger, R. (2018) Phenology information contributes to reduce temporal basis risk in agricultural weather index insurance. Scientific Reports, 8, 1–10.
Efron, B. & Hastie, T. (2016) Computer age statistical inference: algorithms, evidence, and data science. New York: CUP.
European Environment Agency. (2017) Mean precipitation. [Online]. Available at: https://www.eea.europa.eu/data-and-maps/indicators/european-precipitation-2/assessment [Accessed 15th April 2020].
Finger, R. (2013) Investigating the performance of different estimation techniques for crop yield data analysis in crop insurance applications. Agricultural Economics, 44, 217–230.
Friedman, D. (1972) Insurance and the natural hazards. ASTIN, 7, 4–58.
Jewson, S. & Penzer, J. (2006) Estimating trends in weather series: consequences for pricing derivatives. Studies in Nonlinear Dynamics & Econometrics, 10(3), 1–10.
Jewson, S., Brix, A. & Ziehmann, C. (2005) Weather derivative valuation. Cambridge: CUP.
Kaczmarska, J., Jewson, S. & Bellone, E. (2018) Quantifying the sources of simulation uncertainty in natural catastrophe models. SERRA, 32(3), 591–605.
Kendon, M., McCarthy, M., Jevrejeva, S., Matthews, A. & Legg, T. (2019) State of the UK climate 2018. International Journal of Climatology, 39(S1), 1–55.
Marau, D. (2013) When will trends in European mean and heavy daily precipitation emerge? Environmental Research Letters, 8 (1), 1–7.
McKinsey. (2020) State of property and casualty insurance 2020. [Online]. Available at: https://www.mckinsey.com/industries/financial-services/our-insights/state-of-property-and-casualty-insurance-2020 [Accessed 26th March 2021].

Met Office Hadley Centre. (2019) UKCP18 land projections: science report. Exeter: Met Office.

Michel, G. (2018) Risk modelling for hazards and disasters. Oxford: Elsevier.

Mitchell-Wallace, K., Jones, M., Hillier, J. & Foote, M. (2017) Natural catastrophe risk management and modelling. Chichester: Wiley Blackwell.

Skliris, N., Zika, J., Nurser, G., Josey, S. & Marsh, R. (2016) Global water cycle amplifying at less than the Clausius-Clapeyron rate. Scientific Reports, 6(1), 38752.

Smith, D., Scaife, A. & Kirtman, B. (2012) What is the current state of scientific knowledge with regard to seasonal and decadal forecasting? Environmental Research Letters, 7(1), 1–11.

Von Storch, H. & Zwiers, F. (1999) Statistical analysis in climate research. Cambridge: CUP.

Troccoli, A. (2010) Seasonal climate forecasting. Meteorological Applications, 17(3), 251–268.

Wasserman, L. (2004) All of statistics. New York: Springer.

Wasserstein, L. (2016) ASA statement on statistical significance and p-values. The American Statistician, 70(2), 129–133.

Watanabe, S. (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. Journal of Machine Learning Research, 11(1), 3571–3594.

Wilks, D. (2011) Statistical methods in the atmospheric sciences, 3rd edition. Oxford: Academic Press.

Zanardo, S., Nicotina, L., Hilberts, A. & Jewson, S. (2019) Modulation of economic losses from European floods by the North Atlantic oscillation. Geophysical Research Letters, 46(5), 2563–2572.

How to cite this article: Jewson, S., Dallafior, T., & Comola, F. (2021). Dealing with trend uncertainty in empirical estimates of European rainfall climate for insurance risk management. Meteorological Applications, 28(4), e2008. https://doi.org/10.1002/met.2008