ENVIRONMENTAL RESEARCH
LETTERS

LETTER • OPEN ACCESS

Interpreting extreme climate impacts from large ensemble simulations—are they unseen or unrealistic?

To cite this article: T Kelder *et al* 2022 *Environ. Res. Lett.* **17** 044052

IOP Publishing Environ. Res. Lett. 17 (2022) 044052 https://doi.org/10.1088/1748-9326/ac5cf4

ENVIRONMENTAL RESEARCH
LETTERS

OPEN ACCESS

RECEIVED
10 March 2021

REVISED
3 February 2022

ACCEPTED FOR PUBLICATION
11 March 2022

PUBLISHED
29 March 2022

LETTER

Interpreting extreme climate impacts from large ensemble simulations—are they unseen or unrealistic?

T Kelder[1,*], **N Wanders**[2], **K van der Wiel**[3], **T I Marjoribanks**[4], **L J Slater**[5], **R l Wilby**[1] and **C Prudhomme**[1,6,7]

[1] Geography and Environment, Loughborough University, Loughborough, United Kingdom
[2] Department of Physical Geography, Utrecht University, Utrecht, The Netherlands
[3] Royal Netherlands Meteorological Institute (KNMI), De Bilt, The Netherlands
[4] School of Architecture, Building and Civil Engineering, Loughborough, United Kingdom
[5] School of Geography and the Environment, University of Oxford, Oxford, United Kingdom
[6] European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, United Kingdom
[7] UK Centre for Ecology and Hydrology, Wallingford, United Kingdom
[*] Author to whom any correspondence should be addressed.

E-mail: T.Kelder@lboro.ac.uk

Keywords: UNSEEN, large ensembles, climate extremes, impacts, bias correction

Supplementary material for this article is available online

Abstract

Large-ensemble climate model simulations can provide deeper understanding of the characteristics and causes of extreme events than historical observations, due to their larger sample size. However, adequate evaluation of simulated 'unseen' events that are more extreme than those seen in historical records is complicated by observational uncertainties and natural variability. Consequently, conventional evaluation and correction methods cannot determine whether simulations outside observed variability are correct for the right physical reasons. Here, we introduce a three-step procedure to assess the realism of simulated extreme events based on the model properties (step 1), statistical features (step 2), and physical credibility of the extreme events (step 3). We illustrate these steps for a 2000 year Amazon monthly flood ensemble simulated by the global climate model EC-Earth and global hydrological model PCR-GLOBWB. EC-Earth and PCR-GLOBWB are adequate for large-scale catchments like the Amazon, and have simulated 'unseen' monthly floods far outside observed variability. We find that the realism of these simulations cannot be statistically explained. For example, there could be legitimate discrepancies between simulations and observations resulting from infrequent temporal compounding of multiple flood peaks, rarely seen in observations. Physical credibility checks are crucial to assessing their realism and show that the unseen Amazon monthly floods were generated by an unrealistic bias correction of precipitation. We conclude that there is high sensitivity of simulations outside observed variability to the bias correction method, and that physical credibility checks are crucial to understanding what is driving the simulated extreme events. Understanding the driving mechanisms of unseen events may guide future research by uncovering key climate model deficiencies. They may also play a vital role in helping decision makers to anticipate unseen impacts by detecting plausible drivers.

1. Introduction

Weather extremes such as floods, droughts, heatwaves and cyclones can have major societal impacts including mortality and morbidity (Gasparrini *et al* 2015, Raymond *et al* 2020), and economic damages (Felbermayr and Gröschl 2014, Klomp and Valckx 2014, Kousky 2014). Weather extremes can also increase inequality (Dell *et al* 2012, Hallegatte and Rozenberg 2017). In risk analyses, the full range of impacts that may arise from climate and weather extremes must be evaluated (Sutton 2019). For example, the credible maximum extreme event is important for risk estimates of potentially disruptive

impacts (Wilby *et al* 2011), such as mortality, morbidity, and damage from floods in large river systems and from dam failures (e.g. Vano *et al* 2019), or for climate-related shocks to food security (Kent *et al* 2017). However, brevity and sparsity of historical records are well known constraints that confound likelihood estimation of extreme events (Alexander 2016, Wilby *et al* 2017). Climate model projections reduce this limitation but may not capture the full range of extreme events that can arise from climate variability when just a few ensemble members are used (Van der Wiel *et al* 2019b, Mankin *et al* 2020). However, large ensemble simulations from seasonal to multi-decadal prediction systems offer a solution to the estimation of rare events due to their multiple realizations (Allen 2003, van den Brink *et al* 2005, Thompson *et al* 2017, Van der Wiel *et al* 2019b, Mankin *et al* 2020, Brunner and Slater 2022).

Traditionally, large ensembles have been generated by stochastic weather generators trained on the historical record (e.g. Wilks and Wilby 1999, Brunner and Gilleland 2020). However, advances in supercomputing and the physical realism of climate models have facilitated the exploitation of large ensemble simulations for the emulation of events with physically plausible drivers that have not yet been observed (Coumou and Rahmstorf 2012, Stevenson *et al* 2015, Stott *et al* 2016, Kent *et al* 2019, Thompson *et al* 2019, Deser *et al* 2020, Kay *et al* 2020, Swain *et al* 2020, Brunner and Slater 2022). Following Thompson *et al* (2017), we define the use of large ensemble simulations to estimate 'unseen' events more severe than those seen in the historical record as the Unprecedented Simulated Extremes using Ensembles (UNSEEN) approach.

One drawback of using model simulations is that biases are likely to exist, which may occasionally produce unrealistic extreme events. Many techniques have been developed to uncover potential systematic climate model biases (Eyring *et al* 2016, 2019), compare simulated extreme indices with observations (Weigel *et al* 2021), and to evaluate the consistency between simulated and observed distributions of extreme events (Thompson *et al* 2017, 2019, Kelder *et al* 2020, Suarez-Gutierrez *et al* 2021). However, none of these procedures can determine whether the models are correct for the right physical reasons.

Bias correction (or data adjustment) methods are widely used to reduce model discrepancies, especially when coupling climate model simulations with impact models (Warszawski *et al* 2014), but do not necessarily correct the simulations for the right physical reasons (Maraun *et al* 2017). For example, a mismatch between simulations and observations may be caused by observational uncertainties and natural variability, rather than by model biases (Addor and Fischer 2015, Casanueva *et al* 2020). Existing evaluation and correction methods are thus not designed for simulated unseen events. As a consequence, large ensemble simulations with extreme events outside the range of observed variability raise an important question: to what extent can such outliers be trusted? Are the events *unseen* or *unrealistic*?

In this paper, we demonstrate a framework to check that the conclusions about unseen events obtained from large ensemble analyses are sound. Our three steps for assessing the realism of simulated events outside the range of observed variability (figure 1) are inspired by the protocol for event attribution to climate change (Philip *et al* 2020). Step 1 is to review model properties and assess whether the system representation has the capability to represent relevant processes leading to extreme events. Step 2 is to evaluate the statistical features of the large ensemble of simulations (whether from global climate models or regional climate models) by evaluating the consistency of simulated distributions with observations. Bias correction is an integral part of assessing statistical features because it is common practice (e.g. Warszawski *et al* 2014) but may influence the simulated distribution of extreme events and impacts. We, therefore, evaluate the statistical features for both raw and bias corrected values. Step 3 is to assess the physical credibility of the model simulations. Although some studies check the physical processes leading to extreme events—such as teleconnections and land–atmosphere interactions (Van der Wiel *et al* 2017, Thompson *et al* 2019, Vautard *et al* 2019, Kay *et al* 2020)—establishing physical credibility is not straightforward (Philip *et al* 2020), especially for unseen events.

We demonstrate our framework using a case study of Amazon floods. In 2009 and 2012, floods in the Amazon led to the spread of disease, food, and water insecurity (Davidson *et al* 2012, Hofmeijer *et al* 2013, Marengo and Espinoza 2016, Bauer *et al* 2018). At that time, the 2009 flood was the most extreme in 107 years of records, yet three years later it became the second highest in 110 years, drastically altering likelihood estimates. Despite the Amazon stage record being one of the longest in the world, the ∼100 year series is still too short for estimating credible, worst-case events.

To sample more flood events than those available from the historical record, we use EC-Earth large ensemble global climate model simulations coupled with the PCR-GLOBWB global hydrological (water balance) model from an earlier study (Van der Wiel *et al* 2019b). EC-Earth and PCR-GLOBWB are state-of-the-art global models that have been applied in numerous multi-model intercomparison studies, such as within the Coupled Model Intercomparison Project (e.g. Taylor *et al* 2012, Samaniego *et al* 2019, Wanders *et al* 2019), and have been validated globally (Hazeleger *et al* 2012, Sutanudjaja *et al* 2018), including for Amazon streamflow (van Schaik *et al* 2018). Here, we extend previous studies by evaluating whether simulated extremes that exceed the historical

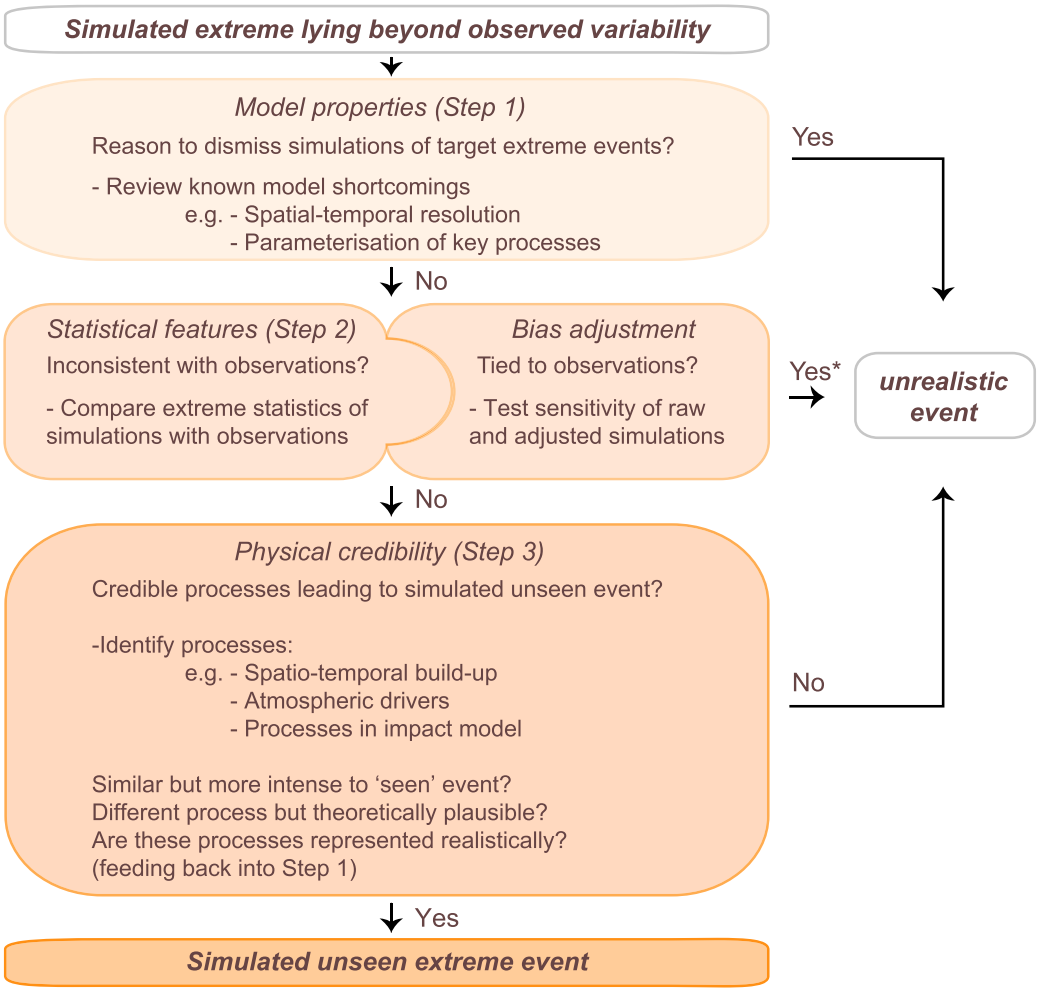

Figure 1. A three-step procedure for evaluating the realism of large ensemble simulations lying outside observed variability. Step 1 is to assess whether the model properties are fit for purpose. Step 2 is to statistically evaluate the simulations, then apply bias correction as required. Step 3 is to evaluate the credibility of the processes within the models leading to the simulation of an unseen event. The orange colour gradient indicates the increasing confidence in the simulation of unseen events throughout the framework.

record are likely to be unseen events or simply unrealistic. We do this by: reviewing the ability of EC-Earth and PCR-GLOBWB to simulate extreme Amazon floods (Step 1); assessing the statistical consistency of these large ensemble simulations with observations using raw data or bias corrected simulations (Step 2) then; exploring the physical drivers behind the largest simulated floods (Step 3).

2. Data

2.1. Study area

The Amazon basin contains the largest contiguous tropical forests in the world, covering an area of 6.5 million km^2. The Amazon river is an important but vulnerable freshwater ecosystem (Castello *et al* 2013), and a key source of food for local communities. Annual high and low flows in this river system are part of a seasonal regime, referred to as the flood pulse. Local livelihoods are adapted to 'normal' levels of inter-annual variability (Pinho *et al* 2012), such that annual floods are not necessarily perceived as 'bad' (Langill and Abizaid 2020). However, occasionally, climate variability can lead to extreme flows (Schöngart and Junk 2007, Towner *et al* 2020) that exceed coping capacities of local communities by impacting transportation, interrupting education and trade, and causing health problems, such as food insecurity (through agricultural losses), water insecurity, and vector-borne diseases (Hofmeijer *et al* 2013, Pinho *et al* 2015, Bauer *et al* 2018). The 2009 flood lasted over two months, destroyed half of the agricultural production, and affected over 20 000 families in the Amazon (Sena *et al* 2012). These events underline the socio-economic importance of estimating plausible Amazon extreme floods. Here,

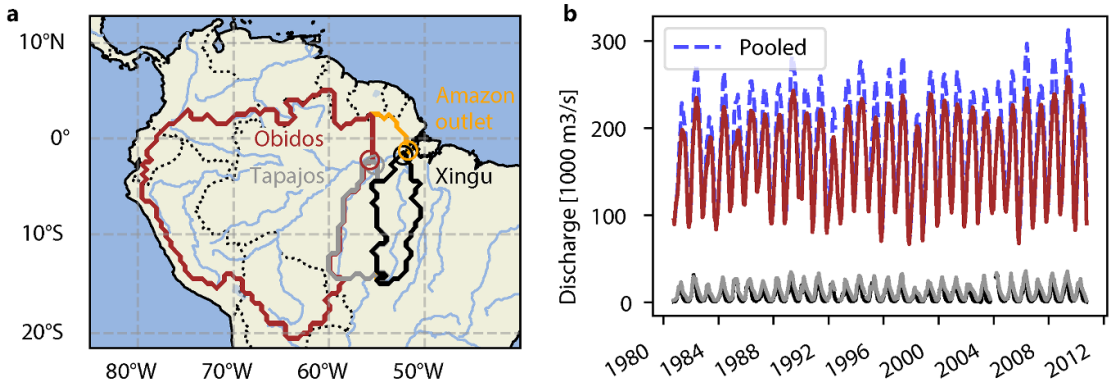

Figure 2. The Amazon basin with selected sub-catchments. (a) Circles indicate the location of the Amazon outlet (orange) and observation stations at the main Amazon River in Obidos (brown), and two southern tributaries: Tapajos (grey) and Xingu (black). Thick orange, grey and black lines indicate the corresponding catchment areas. (b) Observed streamflow time series at the three stations indicated in (a), and their summed values (Pooled).

we employ streamflow at the outlet of the Amazon river (orange circle in figure 2) to evaluate extreme floods.

2.2. Observations

The ∼100 year series mentioned above is for river stage (water level) only. The most downstream streamflow record for the main Amazon River is located at Obidos (brown circle in figure 2(a)). After Obidos, two tributaries from the south, Tapajos (grey circle) and Xingu (black circle in figure 2(a)), join the main Amazon River before the river reaches the outlet. For the period 1981–2010, streamflow data obtained using a rating curve are available for all three stations (figure 2(b)) with less than 10% missing from the catchments attributes for Brazil (CABra) series (Almagro *et al* 2021). In the CABra dataset, gauged daily streamflow from the Brazilian Water Agency are quality controlled to remove outliers, duplicate dates and values. We aggregate the daily data into monthly streamflow averages to match the simulations, then sum the streamflow values in Obidos, Xingu, and Tapajos ('Pooled', figure 2(b)). By pooling (summing) observed station records, we assume negligible streamflow losses between Obidos and Tapajos towards Xingu over monthly timescales. The catchment areas of Obidos, Tapajos, and Xingu represent 99.3% of the total catchment area within the model simulations and, hence, can be reasonably compared. We compute specific discharge (converting cumecs to millimetres per day) to normalize for the slight difference in catchment area between the observations (brown + grey + black catchment outlines) and the simulations at the outlet (orange catchment outline in figure 2(a)).

2.3. Simulations

We use the large ensemble of streamflow simulations presented in Van der Wiel *et al* (2019b). Streamflow was modelled by forcing the global hydrological model PCR-GLOBWB (Sutanudjaja *et al* 2018) with the large ensemble simulations from EC-Earth (Hazeleger *et al* 2012).

EC-Earth v2.3 is a fully coupled free-running global climate model, that combines atmospheric, oceanic, land and sea-ice model components (Hazeleger *et al* 2012), run at hourly (output at daily) time-step and 1.1° spatial resolution. A 2000 year 'present climate' ensemble was created that is representative of global mean surface temperatures (GMSTs) similar to those observed in 2011–2015 (Van der Wiel *et al* 2019b). First, 16 long transient simulations were run with historical forcing (1860–2005) and RCP8.5 (2006–2100). Then, 25 ensemble members were re-initialized with perturbed physics from model years matching observed GMST. The 400 ensemble members (16 × 25) were run for five years, resulting in a 2000 year ensemble.

EC-Earth precipitation was modified before input to PCR-GLOBWB by correcting for too many drizzle days (a recognized limitation of climate models (Dai 2006)) then by adjusting to the observed monthly total precipitation. Drizzle days were corrected using a cut-off value, whereby precipitation days below the threshold are set to 0. This value was determined for each grid cell by matching the amount of EC-Earth precipitation days to ECMWF Re-Analysis (ERA-Interim, Dee *et al* 2011). The total monthly precipitation was corrected linearly for the precipitation days after removing drizzle days by matching with ERA-Interim monthly totals. Bilinear interpolation was applied between EC-Earth gridcell (1.1°) values to regrid output to the PCR-GLOBWB resolution (0.5°).

PCR-GLOBWB is a fully distributed, macro-hydrological model that simulates the global terrestrial water cycle including natural components, with human-water interactions, such as irrigation, reservoirs, and abstractions (Sutanudjaja *et al* 2018). Historical simulations of discharge, water storage,

and water withdrawal have previously been validated against observations globally, and show a high degree of accuracy (Sutanudjaja *et al* 2018). For the streamflow large ensemble used here, PCR-GLOBWB was run on a daily time-step at 0.5° spatial resolution using standard parameterisation (Sutanudjaja *et al* 2018), with outputs reported as monthly averages. For example, the parameterisation of the land surface module (covering for example run-off generation mechanisms), is governed by soil (e.g. FAO Digital Soil Map of the World, Version 3.6), land cover (e.g. GLCC v2.0, Loveland *et al* 2010), and topographic layers (e.g. HydroSHEDS, Lehner *et al* 2008). Routing used in this study is a simplified dynamic routing based on the Manning's equation, to reduce computational demands (Sutanudjaja *et al* 2018). For more details on the streamflow simulations, we refer to (Van der Wiel *et al* 2019b).

3. Methods

In this section, the methods are described for assessing the realism of simulated 'unseen' extreme events, larger than those seen in the historical record. The ability of EC-Earth and PCR-GLOBWB to simulate Amazon floods are reviewed (Step 1 in figure 1); the statistical features of the simulations are compared with observations (Step 2 in figure 1); and the physical credibility of the largest flood simulation is evaluated (Step 3 in figure 1).

3.1. Model properties (Step 1)

This first step is to evaluate the general capability of the model to simulate the target extremes *a priori*. This may include comparing properties such as model scale, resolution, boundary conditions, process representation and model chain coupling, to the target extreme. Reviewing the credibility of a certain model structure or set-up to simulate an extreme is complicated by the complexity of climate and impact models. Whereas the development of evaluated and bias corrected standard model experiments—such as the Inter-Sectoral Impact Model Intercomparison Project (Warszawski *et al* 2014)—has much improved the uptake and uncertainty analysis of model simulations in impact analyses (e.g. Boulange *et al* 2021, Orlov *et al* 2021, Tabari *et al* 2021, Thiery *et al* 2021), model development expertise is typically separated from model analysis and decision-making. Consequently, it can be complicated for users to review the adequacy of the model structure for simulating their target extremes. A suite of questions (which may be adapted depending on the type of model) can be employed to evaluate the model properties identified above. For the Amazon floods we apply two searching questions:

(a) Is the spatial or temporal resolution of the simulations too coarse to represent key processes?

(b) Are key processes dependent upon model parameterisation as opposed to direct simulation?

These questions are intentionally phrased to test whether the '*null hypothesis*' (that the model is adequate) can be rejected rather than prove that it is true. Thus, passing these questions increases our confidence in the model, such that we progress to Step 2. These questions are not meant to, and cannot, cover the fitness-for-purpose of all possible model chains for all types of target extremes and impacts. Rather, they would need to be adjusted accordingly. We refer to IPCC AR6 chapter 10 section 3.3 for an overview of model performance across model chains and types of extreme events and their relevant processes (Doblas-Reyes *et al* 2021).

3.2. Statistical features (Step 2)

The statistical consistency of the streamflow ensemble and observations was evaluated using a fidelity test (Thompson *et al* 2017, 2019, Kelder *et al* 2020). We select the annual maximum monthly streamflow for the grid cell corresponding to the outlet of the Amazon (1.25° S, 51.75° W) and convert it into specific discharge to allow for meaningful comparison with observations. We bootstrap with replacement 10 000 timeseries of 30 years (i.e. the same length as the observations) from the 2000 year simulations. For each bootstrapped timeseries, the mean, standard deviation, skewness, and kurtosis are calculated. The resulting range of the large ensemble is compared with observations.

In addition to testing statistical consistency, we visually inspect the extreme value distributions derived from simulations and observations. We fit the univariate, stationary generalized extreme value (GEV) distribution to the observed annual maximum streamflow, using maximum likelihood estimation of the distribution parameters. We select the stationary GEV distribution because it is widely applied for flood analyses in practice (Coles 2001, Madsen *et al* 2014). Other distributions and/or nonstationary behaviour could be explored but are beyond the scope of this paper. We employ a parametric bootstrap to derive confidence intervals. In addition, we undertake a frequentist analysis of observed and simulated annual maxima using the return period as the length of the data divided by the rank of the extreme. For example, the highest value within 2000 years of simulations is estimated as a 2000 year return period, the second highest as the 1000 year return period, and so forth.

Since models are imperfect representations of reality, systematic errors may exist in model simulations. Therefore, model errors are often bias corrected before outputs are used for impact assessments (Warszawski *et al* 2014). However, bias corrections may adjust the simulated distribution of extremes.

We assess the sensitivity of the monthly specific discharge simulations to two routinely used bias correction methods: empirical quantile mapping and a scaling factor. Empirical quantile mapping is widely applied in impact studies (Zscheischler *et al* 2019) whereas scaling factors (additive for temperature and multiplicative for precipitation) are common in event attribution studies (Philip *et al* 2020). We estimate values of the empirical cumulative distribution function for regularly spaced quantiles via the 'qmap' R-package (Gudmundsson *et al* 2012). These estimates are then used to perform quantile mapping using linear interpolation and a constant correction for the extrapolation, as suggested by Boé *et al* (2007). For the constant scaling factor method, we use the ratio between the mean of the simulated and observed annual maximum monthly streamflow. We pool all members for estimating the bias correction factors, as correcting each member independently reduces the spread of the ensemble (Chen *et al* 2019).

3.3. Physical credibility (Step 3)

In Step 3, we assess the physical credibility of the processes leading to the simulation of an extreme event that has not yet occurred (figure 1). First, the processes leading to the simulation of unseen events are identified. We divide this into three sub-steps: (a) the spatial-temporal build-up of the unseen event; (b) the driving atmospheric variables and processes within the climate model; and (c) the driving processes in the impact model. Checking the credibility of these processes is not straightforward, but the processes generating the largest simulated extreme can be placed into perspective with historical events. In the case of the Amazon, one might ask whether the largest simulated monthly flood is the result of a meteorological event similar to historical events (but more intense), or whether other mechanisms were involved. If other mechanisms are identified, their theoretical plausibility can be assessed. As a final check, the model properties related to the identified processes are reviewed (feeding back in Step 1).

For illustrative purposes, the spatial and temporal characteristics of the largest simulated monthly flood are compared with the observed flood in 2009, for which data are available across all observation stations. In addition, we calculate the empirical 2-, and 20 year monthly floods, based on the 29 year pooled record. Empirical return values are estimated as the quantile corresponding to the $1 - (1/\text{return period})$, hence the two year value is the 0.5 quantile and the 20 year value is the 0.95 quantile. For the temporal build-up of the flood, we show the streamflow values in the year preceding the simulated and observed flood. We use simulations at the Amazon outlet and pooled observations at Obidos, Tapajos, and Xingu (see the Data section).

We then assess the spatial distribution of the streamflow contributing to the flood peak for each month in the year preceding the largest simulated monthly flood. For each grid cell in the Amazon basin, we calculate the percentage of the streamflow compared with the flood peak (supplementary figure 1 available online at stacks.iop.org/ERL/17/044052/mmedia). After evaluating the spatial-temporal build-up of the largest simulated flood event, we assess the credibility of the drivers in EC-Earth and in PCR-GLOBWB. We plot EC-Earth precipitation over the Amazon basin for each month in the year preceding the largest simulated monthly flood (supplementary figure 2), and we investigate the PCR-GLOBWB direct runoff and bias corrected precipitation over the Amazon in addition to the streamflow and raw precipitation.

4. Results

Step 1 of the event evaluation procedure is to review whether there are known limitations of the EC-Earth and PCR-GLOBWB resolution and process representation that may influence Amazon flood peak simulations. The daily temporal resolution of both EC-Earth and PCR-GLOBWB is sufficiently fine when compared with the averaged monthly values used in the analysis and because floods in the Amazon are part of a seasonal regime (lasting up to several months Barichivich *et al* (2018)) there is no reason to dismiss the simulations based on their temporal resolution. The large extent of the Amazon basin also means that the spatial distribution is adequately represented by the 1×1 degree climate model and 0.5×0.5 degree hydrological model. In contrast, small and steep catchments with faster rainfall-runoff responses would require higher spatial-temporal resolution (Schaller *et al* 2020).

Considering process representation, EC-Earth is a global climate model that simulates the atmosphere, ocean, land, and sea-ice components. Important modulators of Amazon floods are the El Nino Southern Oscillation (ENSO) (e.g. Marengo and Espinoza 2016) and the Walker circulation (e.g. Barichivich *et al* 2018), which are well simulated by EC-Earth (Hazeleger *et al* 2012, Sterl *et al* 2012, Pausata *et al* 2017). However, EC-Earth underestimates precipitation over the Amazon during December–February and June–August seasons (Hazeleger *et al* 2012), possibly because convection is parameterised in EC-Earth.

PCR-GLOBWB is a fully distributed global hydrological model that generates runoff as a combination of direct runoff, indirect flow (through the soil reservoir), groundwater flow, and, snowmelt. Canopy interception is included as initial loss of precipitation (Sutanudjaja *et al* 2018). The model covers all major components of the terrestrial water cycle including human-water interactions. Runoff routing is included, but backwaters are not simulated. van Schaik *et al* (2018)

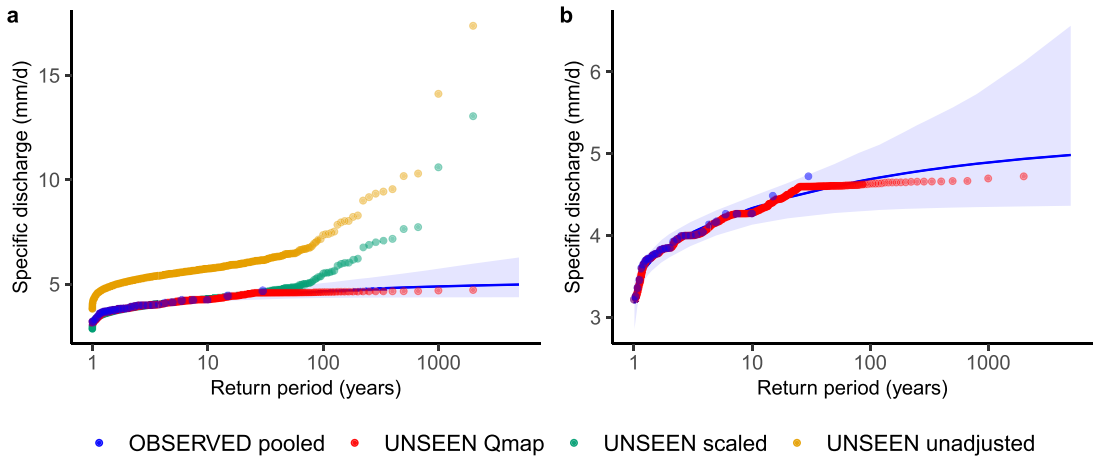

Figure 3. Evaluation and correction of UNSEEN Amazon flood simulations. (a) Extreme value distributions showing the monthly specific discharge for the historical record (blue circles) alongside the UNSEEN streamflow large ensemble, both before (orange circles) and after applying quantile mapping (Qmap, red circles) or a scaling factor (green circles). The blue line indicates the estimated extreme value distribution based on the observed record including 95% confidence intervals based on parametric bootstrapping. (b) As in (a) but highlighting Qmap and the observations to better illustrate the influence of the correction on the simulated extremes.

report that PCR-GLOBWB monthly discharge simulations forced with observed precipitation reproduces observed discharge at Obidos 'reasonably well', with a slight overestimation of the flood peaks.

As PCR-GLOBWB is a physically based, uncalibrated model, it is prone to parameter uncertainty. The parameters are based on static maps, that cannot capture any non-stationarity in catchment properties, such as changing land cover due to deforestation. PCR-GLOBWB soil parameters show the largest sensitivity for Amazon flood simulations (Sperna Weiland *et al* 2015), but high-quality precipitation data and streamflow routing are the dominant factors influencing Amazon flood peak simulations (Hoch *et al* 2017, Towner *et al* 2019). Overall, the main sources of uncertainty determined by this first step are, therefore, the underestimation of precipitation from EC-Earth, and the simplified runoff-routing scheme used in PCR-GLOBWB. There is no reason to dismiss the EC-Earth and PCR-GLOBWB simulations of unseen floods based on this first step alone, so we further validate the simulated streamflow extremes (Step 2), then identify and evaluate their drivers (Step 3).

Validation of 2000 years of present-climate Amazon monthly flood simulations is hampered by the length of the observational record (30 years in this case, 1981–2010). We therefore compare the statistical features of the simulations with observations, following Thompson *et al* (2017). The simulated annual maximum streamflow (in terms of monthly specific discharge, see 'Simulations') (UNSEEN) is overestimated when evaluated against the historical record (orange circles compared with blue circles in figure 3(a)). The bias is confirmed by the statistical consistency test, which shows that the mean of the simulated annual maximum streamflow is significantly higher than observations (orange lines compared to blue line in figure 4(a)). Furthermore, the simulations have a skewed distribution and long tail when bootstrapped to the same length of the observations (figures 4(b)–(d)), reflected by the wide range of the variability (standard deviation) and the shape (skewness and kurtosis). This means that either the simulations are wrong, or the observations are too short to well constrain the tail of the distribution.

We assess the sensitivity of the simulated distribution of Amazon monthly floods to bias corrections using quantile mapping and a scaling factor. Empirical quantile mapping corrects all moments of the distribution (red lines compared to orange lines in figure 4) and, therefore, fits the observed distribution very well (figure 3(b)). However, in the process, the correction adjusted the long tail as simulated by the climate model (orange vs. red circles in figure 3(a)). The constant scaling factor, in contrast, only corrects the mean and standard deviation of the simulated extremes (green versus orange vertical lines in figure 4) and so retains the shape of the distribution (skewness and kurtosis). Scaled simulations match observations until the 50 year period but deviate markedly beyond that (green versus blue circles in figure 3(a)). We, thus, find high sensitivity of the simulated Amazon monthly flood distribution to the bias correction method, but it cannot be statistically determined which is better—the physical credibility must be assessed.

The final step is to assess the physical credibility of a simulated unseen event (Step 3). In our example, we first evaluate the spatial and temporal characteristics of the maximum monthly flood simulation. We find that the flood peak occurred in July, and most of the discharge was generated in the month preceding the flood (figure 5(b)). This

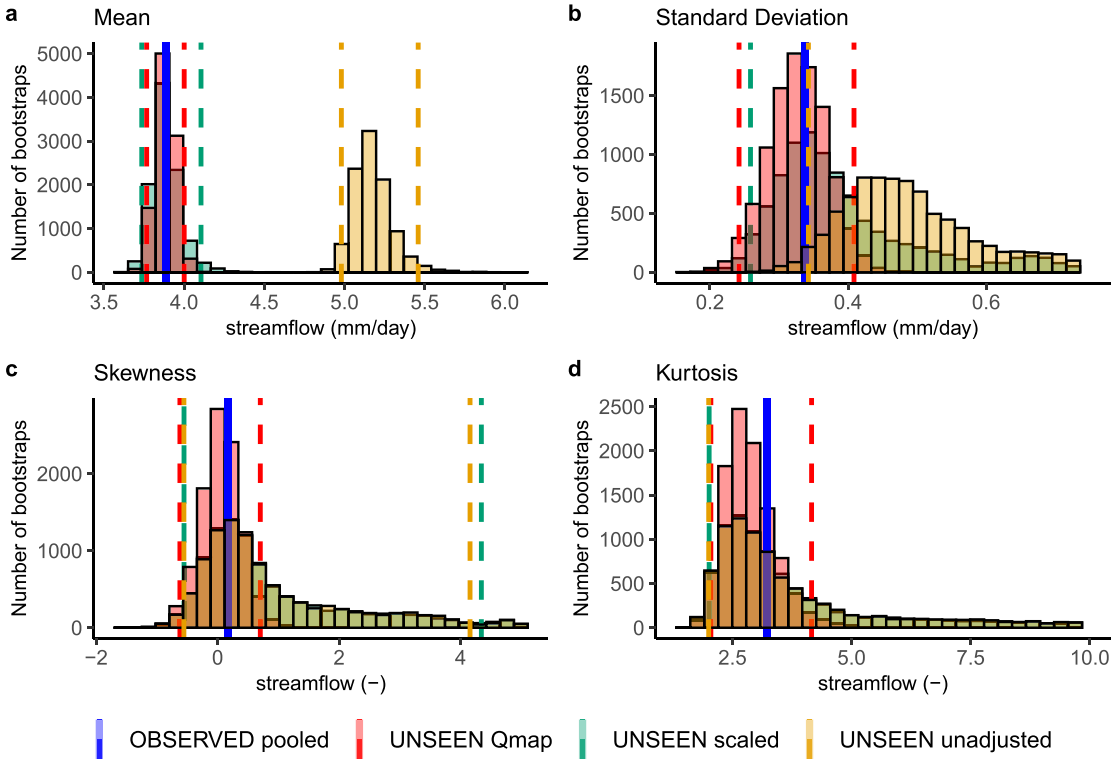

Figure 4. Testing the consistency of simulations and observations before and after bias correction. The distribution characteristics of annual maximum streamflow before (orange) and after applying quantile mapping (red) and scaling (green) are compared to observations (blue) in terms of the (a) mean (b) standard deviation (c) skewness and (d) kurtosis. Histograms show the distributions for the 10 000 simulations bootstrapped to the length of the observed record and dashed lines indicate the 95% confidence intervals. Note that the x-limits in (b) and (d) are set to 0.75 and 10 to improve the clarity of the figures (for the full range see supplementary figure 3).

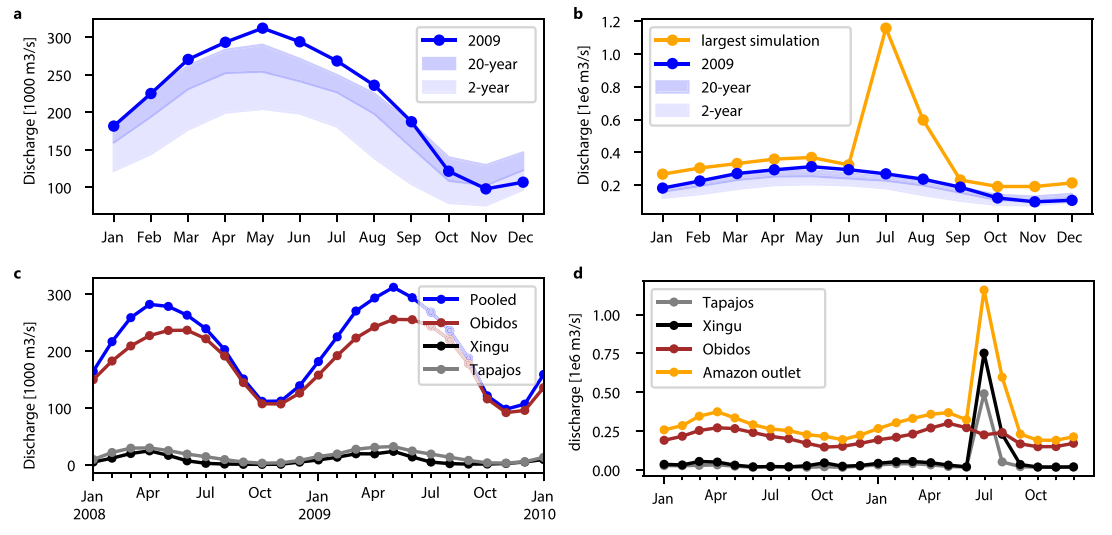

Figure 5. Spatial and temporal characteristics of observed floods and the largest simulated flood. (a), (b) Timing of observed floods for the period 1981–2010 (blue, (a), (b)) and the largest simulated flood (orange, (b)). The 2 year and 20 year floods are presented as ribbons representing 1–2 year and 2–20 year floods, to improve the clarity of the figure. (c), (d) The contribution of the southern tributaries (Xingu and Tapajos, indicated in figure 2(a)) to (c) the observed 2009 flood and (d) the largest simulated flood.

sequence is inconsistent with observed floods, which gradually build up over the season (figures 5(a) and (b)). We find that the simulated discharge originates from the southern tributaries Tapajos and Xingu (figure 5(d) and supplementary figure 1), whereas there is little contribution from these regions to observed floods (figure 2(b) and figure 5(c)). Instead, for the 2009 flood, precipitation progressed from west to east over the catchment during January–May, resulting in a temporally compounding flood peak in

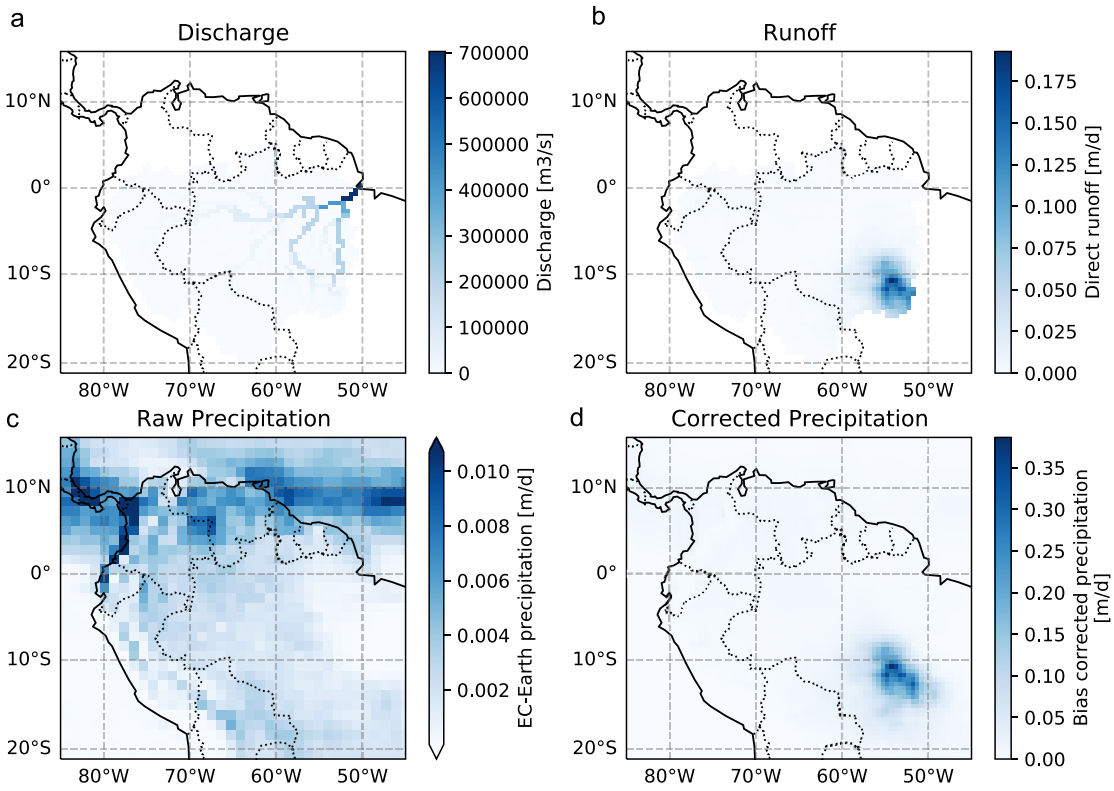

Figure 6. Identifying the driver of the largest simulated Amazon flood from (a) July average discharge, (b) direct runoff, (c) precipitation and (d) bias corrected precipitation over the Amazon preceding the most extreme flood event within the model experiments.

May (Marengo *et al* 2012, Sena *et al* 2012, Filizola *et al* 2014).

We further assess the physical drivers of the maximum simulated monthly flood to explain whether this event might be caused by an unseen, rare physical driving mechanism that has not yet been observed, or whether it might be caused by an unrealistic model bias or error (figure 6). We determine that the flood is driven by direct runoff from the south, which is linked to a local peak in the bias-corrected precipitation used to run the hydrological model. However, this peak is not found in the raw precipitation data of EC-Earth. We thus conclude that this unseen Amazon monthly flood was an artefact of a bias correction mechanism generating extreme precipitation over the Southern portion of the Amazon.

Upon further investigation of the mechanisms leading to this extreme flood, we find that a dry bias in May-September EC-Earth precipitation over the Amazon led to a high multiplication factor in the correction of monthly total precipitation (supplementary figure 4). A dry bias for the Amazon is a well-known limitation of climate models (Eyring *et al* 2019). However, the bias is especially marked in the southern tributaries of the Amazon during July (figure 7(a)). Closer inspection of a grid cell within this region (white cross in figure 7(a)) reveals how a small number of precipitation events were unrealistically inflated by the high correction ratio (figures 7(c) and (d). Indeed, the second largest simulated monthly flood also originated in the southern tributaries during summer (supplementary figure 5). Moreover, we find that the Amazon has the largest correction ratio globally (>100, figure 7(b)). Other large factors (10–100) are found in July and August over Central Asia. Conversely, the smallest corrections (1/1000) occur over the Sahara all year round, with ramifications for the realism of drought estimates there (supplementary figure 6).

5. Discussion

This work develops a procedure to evaluate simulations of unseen events, illustrated through a case study of Amazon floods. We use a large ensemble of 2000 years of simulations from the EC-Earth global climate model with offline coupling to PCR-GLOBWB hydrological model (Van der Wiel *et al* 2019b). The two largest events within 2000 years of model experiments are unexpectedly extreme when compared to observations. Conventional evaluation and correction methods (e.g. Maraun 2016, Eyring *et al* 2019) are not well-suited to simulations outside observed variability, so we follow a three-step procedure (figure 1), to evaluate the realism of these simulated events. We review the ability of EC-Earth

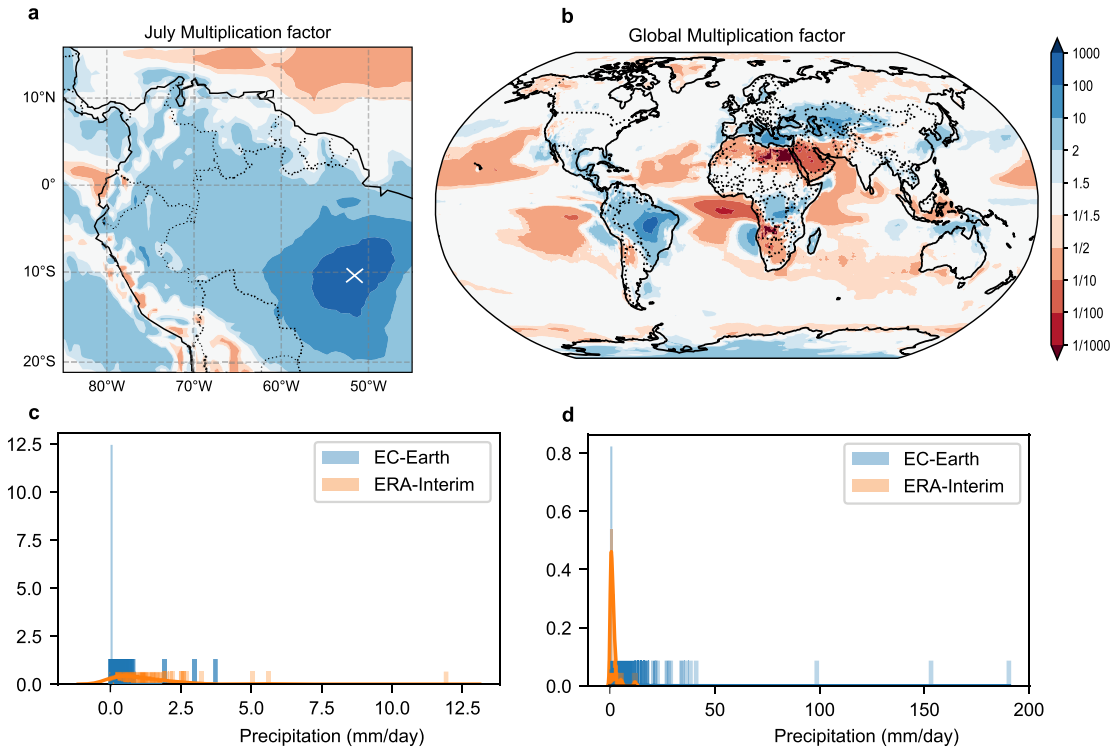

Figure 7. A global perspective on the bias correction issue. (a), (b) Precipitation multiplication factors for July over the Amazon (a) and globally (b). (c), (d) Histograms with kernel density estimation and rugplots of July precipitation for the grid cell indicated by the white cross in (a). The data consist of 2000 years of simulations (blue, EC-Earth) and 32 years of reanalysis (orange, ERA-I, 1979–2010) for raw EC-Earth simulations (c) and corrected simulations (d).

and PCR-GLOBWB to simulate Amazon flood simulations and conclude that the underestimation of precipitation in EC-Earth and simplified runoff routing scheme in PCR-GLOBWB are the dominant sources of uncertainty. However, these were insufficient reasons to dismiss the monthly flood simulations over the Amazon *a priori* (Step 1).

We compare the statistical features of the 2000 years of present-climate Amazon monthly flood simulations to 30 years of observations, following (Thompson *et al* 2017). We find that annual maximum streamflow (monthly specific discharge) simulations are inconsistent with the observations (Step 2). Most notably, simulations show a skewed distribution and long tail that is not present in the observations. This difference could be caused by infrequent compound behaviour that cannot be detected well within the comparatively short observational record. For example, large floods can be generated by spatially and temporally compounding flows from multiple sub-regions and months (Marengo *et al* 2012, Sena *et al* 2012, Filizola *et al* 2014, Zscheischler *et al* 2020). Hence, model simulations may well be realistic despite being inconsistent with observations.

We correct the monthly flood simulations for the Amazon using two commonly applied methods to study the effect of bias correction on conclusions about unseen events. We show that simulated unseen Amazon monthly floods are removed by correcting the simulations to the observations using quantile mapping, whereas scaling factors may retain such extremes (by only adjusting the mean and/or standard deviation of the distribution). Whether or not the simulated unseen extremes are realistic cannot be statistically explained; hence physical credibility should be checked.

We find that the largest simulated monthly flood is inconsistent with observations and current physical understanding, because it results from a very large precipitation bias correction factor during climatologically dry months. In this case, correctly representing spatial-temporal consistency and multi-variate dependency (Cannon 2018, Zscheischler *et al* 2019) might be more important than avoiding bias correction, which is in part justified because moderate meteorological events can cause extreme impacts (Van der Wiel *et al* 2020). Further advances in high-resolution dynamical downscaling of large ensemble simulations may one day obviate the need for such bias corrections (Huang *et al* 2020, Ødemark *et al* 2021).

The example of the Amazon reveals the utility of physical credibility checks for discerning behaviours within model worlds (Step 3). In this case, the physical credibility check is carried out manually for single events (evaluating the driver of the largest two simulated events). To assess multiple events, composite analyses can be used (e.g. Thompson *et al*

2019, Kelder *et al* 2020). Furthermore, correlation and regression methods (Wilks 2011), and causal inference methods (Runge *et al* 2019) may prove useful in systematically evaluating the realism of simulated drivers for the entire ensemble. Composite analysis of observed Amazon floods has demonstrated their connection to the ENSO (Marengo and Espinoza 2016). Nevertheless, for the most extreme floods, which may have unique driving mechanisms, single event analyses can provide insightful information in addition to the general, averaged, relationship between floods and teleconnections obtained from composite analyses (Towner *et al* 2020).

The physical credibility check can be applied to other regions and applications by assessing whether the atmospheric pattern associated with a given event is similar to the pattern that might be expected from observations, or can it be explained from theory? Answering such questions through regional evaluation of the physical drivers of the largest simulated impacts can provide insight into the credibility of the simulations. In this case, we determined that the flood originated from the southern tributaries, but that the precipitation is low over this region in the raw climate model simulations, indicating a discrepancy in the water balance. We, therefore, concluded the analysis after determining that the bias correction mechanism drives the largest simulated monthly flood. In other cases, relevant climate anomalies or hydrological state variables could be compared with anomalies during historical extreme events until the causes of the event, and its credibility, are fully understood. For example, Thompson *et al* (2019) studied unseen temperature extremes in South East China and found that variability in the Indian summer monsoon may cause temperature extremes beyond the current record.

6. Conclusion

Large-ensemble simulations are increasingly being used to explore the characteristics of plausible extreme events (van den Brink *et al* 2005, Thompson *et al* 2017, van Kempen *et al* 2021). They are also used to improve the sampling of internal variability over multi-decadal projections (Deser *et al* 2020, Lehner *et al* 2020, Maher *et al* 2020, Mankin *et al* 2020) and to attribute the causes of high-impact events (Schutgens *et al* 2017, Krishnamurthy *et al* 2018, Van der Wiel *et al* 2018, 2019a, Pascale *et al* 2020, Schlunegger *et al* 2020, Suarez-Gutierrez *et al* 2020). However, the use of large-ensemble simulations to deepen understanding of climate-related risks hinges on the realism of the simulations. It is, therefore, essential to thoroughly evaluate large ensemble simulations to avoid false confidence in statistical estimates or erroneous conclusions when model simulations may be wrong (Stainforth *et al* 2007). Conventional evaluation and correction methods are sensitive to observational uncertainties and natural variability and cannot determine whether simulations outside observed variability are correct for the right physical reasons. Here, we demonstrate a framework for, and illustrate the complexities associated with, evaluating and then correcting simulated impacts outside observed climate variability. For Amazon monthly flood simulations from EC-Earth and PCR-GLOBWB, we found large differences between simulated and observed distributions that could not be statistically explained. The physical realism must be checked, which, in this case, showed that the largest simulated monthly flood was an artefact of a bias correction mechanism. We conclude that there is high sensitivity of the simulations outside observed variability to the bias correction method, and that physical credibility checks are crucial to understanding what is driving the simulated extreme events. We, therefore, make a cautionary remark that bias correction of large ensemble simulations might unnecessarily 'tie' simulated distributions to observed distributions, but we discuss how use of such corrections may be justified to meet the needs of impact models. We, furthermore, recommend evaluating the drivers of simulations outside observed variability to explain their realism beyond what is possible from conventional approaches. Uncovering the characteristics of events in the models may reveal the most important model deficiencies limiting impact analysis which, may in turn, guide future research. Furthermore, detecting plausible drivers of extremes beyond observed impacts may improve our scientific understanding of unknown events and help provide decision makers with invaluable information to prepare for unseen impacts.

Data availability statement

The CABra streamflow observations for the Amazon River are freely available at https://doi.org/10.5281/zenodo.4655204. The annual maximum monthly streamflow simulations are available at https://doi.org/10.5281/zenodo.2536395. The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/10.5281/zenodo.4585400 (Kelder 2021).

Acknowledgments

TK was supported by Loughborough University and the NERC CENTA Doctoral Training Partnership. TK acknowledges computation facilities provided by NWO Surfsara. NW acknowledges funding from NWO 016.Veni.181.049. KW acknowledges funding from NWO ALWCL.2016.2.

Conflict of interest

The authors declare no competing interests.

Code availability

Notebooks containing the code used in this paper are openly available on the GitHub page https://timokelder.github.io/EXPLORE/Intro.html.

Authors' contributions

T K conceived and designed the study. T K drafted the paper with extensive contributions from all authors. T K analysed the data with input from all authors. N W and K W acquired the data. T K produced the figures.

ORCID iDs

T Kelder https://orcid.org/0000-0001-9802-9837
N Wanders https://orcid.org/0000-0002-7102-5454
K van der Wiel https://orcid.org/0000-0001-9365-5759
T I Marjoribanks https://orcid.org/0000-0003-0116-2952
L J Slater https://orcid.org/0000-0001-9416-488X
R l Wilby https://orcid.org/0000-0002-4662-9344
C Prudhomme https://orcid.org/0000-0003-1722-2497

References

Addor N and Fischer E M 2015 The influence of natural variability and interpolation errors on bias characterization in RCM simulations *J. Geophys. Res. Atmos.* **120** 10180–95
Alexander L V 2016 Global observed long-term changes in temperature and precipitation extremes: a review of progress and limitations in IPCC assessments and beyond *Weather Clim. Extremes* **11** 4–16
Allen M 2003 Liability for climate change *Nature* **421** 891
Almagro A, Oliveira P T S, Meira Neto A A, Roy T and Troch P 2021 CABra: a novel large-sample dataset for Brazilian catchments *Hydrol. Earth Syst. Sci.* **25** 3105–35
Barichivich J, Gloor E, Peylin P, Brienen R J W, Schöngart J, Espinoza J C and Pattnayak K C 2018 Recent intensification of Amazon flooding extremes driven by strengthened walker circulation *Sci. Adv.* **4** eaat8785
Bauer T, Ingram V, De Jong W and Arts B 2018 The socio-economic impact of extreme precipitation and flooding on forest livelihoods: evidence from the bolivian Amazon *Int. For. Rev.* **20** 314–31
Boé J, Terray L, Habets F and Martin E 2007 Statistical and dynamical downscaling of the seine basin climate for hydro-meteorological studies *Int. J. Climatol.* **27** 1643–55
Boulange J, Hanasaki N, Yamazaki D and Pokhrel Y 2021 Role of dams in reducing global flood exposure under climate change *Nat. Commun.* **12** 1–7
Brunner M I and Gilleland E 2020 Stochastic simulation of streamflow and spatial extremes: a continuous, wavelet-based approach *Hydrol. Earth Syst. Sci.* **24** 3967–82
Brunner M I and Slater L J 2022 Extreme floods in Europe: going beyond observations using reforecast ensemble pooling *Hydrol. Earth Syst. Sci.* **26** 469–82
Cannon A J 2018 Multivariate quantile mapping bias correction: an N-dimensional probability density function transform for climate model simulations of multiple variables *Clim. Dyn.* **50** 31–49
Casanueva A, Herrera S, Iturbide M, Lange S, Jury M, Dosio A, Maraun D and Gutiérrez J M 2020 Testing bias adjustment methods for regional climate change applications under observational uncertainty and resolution mismatch *Atmos. Sci. Lett.* **21** e978
Castello L, Mcgrath D G, Hess L L, Coe M T, Lefebvre P A, Petry P, Macedo M N, Renó V F and Arantes C C 2013 The vulnerability of Amazon freshwater ecosystems *Conserv. Lett.* **6** 217–29
Chen J, Brissette F P, Zhang X J, Chen H, Guo S and Zhao Y 2019 Bias correcting climate model multi-member ensembles to assess climate change impacts on hydrology *Clim. Change* **153** 361–77
Coles S 2001 *An Introduction to Statistical Modeling of Extreme Values* vol 208 (London: Springer)
Coumou D and Rahmstorf S 2012 A decade of weather extremes *Nat. Clim. Change* **2** 491–6
Dai A 2006 Precipitation characteristics in eighteen coupled climate models *J. Clim.* **19** 4605–30
Davidson E A *et al* 2012 The Amazon basin in transition *Nature* **481** 321–8
Dee D P *et al* 2011 The ERA-interim reanalysis: configuration and performance of the data assimilation system *Q. J. R. Meteorol. Soc.* **137** 553–97
Dell M, Jones B F and Olken B A 2012 Temperature shocks and economic growth: evidence from the last half century *Am. Econ. J. Macroecon.* **4** 66–95
Deser C *et al* 2020 Insights from earth system model initial-condition large ensembles and future prospects *Nat. Clim. Change* **10** 277–86
Doblas-Reyes F J *et al* 2021 Linking global to regional climate change *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge University Press) accepted
Eyring V *et al* 2016 ESMValTool (v1.0)—a community diagnostic and performance metrics tool for routine evaluation of earth system models in CMIP *Geosci. Model Dev.* **9** 1747–802
Eyring V *et al* 2019 Taking climate model evaluation to the next level *Nat. Clim. Change* **9** 102–10
Felbermayr G and Gröschl J 2014 Naturally negative: the growth effects of natural disasters *J. Dev. Econ.* **111** 92–106
Filizola N, Latrubesse E M, Fraizy P, Souza R, Guimarães V and Guyot J L 2014 Was the 2009 flood the most hazardous or the largest ever recorded in the Amazon? *Geomorphology* **215** 99–105
Gasparrini A *et al* 2015 Mortality risk attributable to high and low ambient temperature: a multicountry observational study *Lancet* **386** 369–75
Gudmundsson L, Bremnes J B, Haugen J E and Engen-Skaugen T 2012 Technical note: downscaling RCM precipitation to the station scale using statistical transformations—a comparison of methods *Hydrol. Earth Syst. Sci.* **16** 3383–90
Hallegatte S and Rozenberg J 2017 Climate change through a poverty lens *Nat. Clim. Change* **7** 250–6
Hazeleger W *et al* 2012 EC-earth V2.2: description and validation of a new seamless earth system prediction model *Clim. Dyn.* **39** 2611–29
Hoch J M, Haag A V, Van Dam A, Winsemius H C, Van Beek L P H and Bierkens M F P 2017 Assessing the impact of hydrodynamics on large-scale flood wave propagation—a case study for the Amazon basin *Hydrol. Earth Syst. Sci.* **21** 117–32
Hofmeijer I, Ford J D, Berrang-Ford L, Zavaleta C, Carcamo C, Llanos E, Carhuaz C, Edge V, Lwasa S and Namanya D 2013 Community vulnerability to the health effects of climate change among indigenous populations in the peruvian Amazon: a case study from panaillo and nuevo progreso *Mitig. Adapt. Strateg. Glob. Change* **18** 957–78
Huang X, Swain D L and Hall A D 2020 Future precipitation increase from very high resolution ensemble downscaling of extreme atmospheric river storms in California *Sci. Adv.* **6** eaba1323
Kay G, Dunstone N, Smith D, Dunbar T, Eade R and Scaife A 2020 Current likelihood and dynamics of hot summers in the UK *Environ. Res. Lett.* **15** 094099

Kelder T 2021 Dataset supporting "Interpreting extreme climate impacts from large ensemble simulations – are they unseen or unrealistic?" Zenodo (available at: https://doi.org/10.5281/zenodo.4585400)
Kelder T, Müller M, Slater L J, Marjoribanks T I, Wilby R L, Prudhomme C, Bohlinger P, Ferranti L and Nipen T 2020 Using UNSEEN trends to detect decadal changes in 100-year precipitation extremes *npj Clim. Atmos. Sci.* **3** 47
Kent C, Pope E, Dunstone N, Scaife A A, Tian Z, Clark R, Zhang L, Davie J and Lewis K 2019 Maize drought hazard in the Northeast farming region of China: unprecedented events in the current climate *J. Appl. Meteorol. Climatol.* **58** 2247–58
Kent C, Pope E, Thompson V, Lewis K, Scaife A A and Dunstone N 2017 Using climate model simulations to assess the current climate risk to maize production *Environ. Res. Lett.* **12** 054012
Klomp J and Valckx K 2014 Natural disasters and economic growth: a meta-analysis *Glob. Environ. Change* **26** 183–95
Kousky C 2014 Informing climate adaptation: a review of the economic costs of natural disasters *Energy Econ.* **46** 576–92
Krishnamurthy L, Vecchi G A, Yang X, van der Wiel K, Balaji V, Kapnick S B, Jia L, Zeng F, Paffendorf K and Underwood S 2018 Causes and probability of occurrence of extreme precipitation events like chennai 2015 *J. Clim.* **31** 3831–48
Langill J C and Abizaid C 2020 What is a bad flood? local perspectives of extreme floods in the peruvian Amazon *Ambio* **49** 1423–36
Lehner B, Verdin K and Jarvis A 2008 New global hydrography derived from spaceborne elevation data *EOS Trans. Am. Geophys. Union* **89** 93–94
Lehner F, Deser C, Maher N, Marotzke J, Fischer E M, Brunner L, Knutti R and Hawkins E 2020 Partitioning climate projection uncertainty with multiple large ensembles and CMIP5/6 *Earth Syst. Dyn.* **11** 491–508
Loveland T R, Reed B C, Ohlen D O, Brown J F, Zhu Z, Yang L and Merchant J W 2010 Development of a global land cover characteristics database and IGBP DISCover from 1 km AVHRR data *Int. J. Remote Sens.* **21** 1303–30
Madsen H, Lawrence D, Lang M, Martinkova M and Kjeldsen T R 2014 Review of trend analysis and climate change projections of extreme precipitation and floods in Europe *J. Hydrol.* **519** 3634–50
Maher N, Lehner F and Marotzke J 2020 Quantifying the role of internal variability in the temperature we expect to observe in the coming decades *Environ. Res. Lett.* **15** 054014
Mankin J S, Lehner F, Coats S and McKinnon K A 2020 The value of initial condition large ensembles to robust adaptation decision-making *Earth's Future* **8** e2012EF001610
Maraun D 2016 Bias correcting climate change simulations—a critical review *Curr. Clim. Change Rep.* **2** 211–20
Maraun D *et al* 2017 Towards process-informed bias correction of climate change simulations *Nat. Clim. Change* **7** 764–73
Marengo J A and Espinoza J C 2016 Extreme seasonal droughts and floods in Amazonia: causes, trends and impacts *Int. J. Climatol.* **36** 1033–50
Marengo J A, Tomasella J, Soares W R, Alves L M and Nobre C A 2012 Extreme climatic events in the Amazon basin *Theor. Appl. Climatol.* **107** 73–85
Ødemark K, Müller M and Tveito O E 2021 Changing lateral boundary conditions for probable maximum precipitation studies: a physically consistent approach *J. Hydrometeorol.* **22** 113–23
Orlov A, Daloz A S, Sillmann J, Thiery W, Douzal C, Lejeune Q and Schleussner C 2021 Global economic responses to heat stress impacts on worker productivity in crop production *Econ. Disasters Clim. Change* **5** 367–90
Pascale S, Kapnick S B, Delworth T L and Cooke W F 2020 Increasing risk of another cape town 'day zero' drought in the 21st century *Proc. Natl Acad. Sci. USA* **117** 29495–503
Pausata F S R, Zhang Q, Muschitiello F, Lu Z, Chafik L, Niedermeyer E M, Stager J C, Cobb K M and Liu Z 2017 Greening of the Sahara suppressed ENSO activity during the mid-holocene *Nat. Commun.* **8** 1–12
Philip S *et al* 2020 A protocol for probabilistic extreme event attribution analyses *Adv. Stat. Climatol. Meteorol. Oceanogr.* **6** 177–203
Pinho P F, Marengo J A and Smith M S 2015 Complex socio-ecological dynamics driven by extreme events in the Amazon *Reg. Environ. Change* **15** 643–55
Pinho P F, Orlove B and Lubell M 2012 Overcoming barriers to collective action in community-based fisheries management in the amazon *Hum. Organ.* **71** 99–109
Raymond C, Matthews T and Horton R M 2020 The emergence of heat and humidity too severe for human tolerance *Sci. Adv.* **6** eaaw1838
Runge J *et al* 2019 Inferring causation from time series in earth system sciences *Nat. Commun.* **10** 1–13
Samaniego L *et al* 2019 Hydrological forecasts and projections for improved decision-making in the water sector in Europe *Bull. Am. Meteorol. Soc.* **100** 2451–72
Schaller N, Sillmann J, Müller M, Haarsma R, Hazeleger W, Hegdahl T J, Kelder T, van den Oord G, Weerts A and Whan K 2020 The role of spatial and temporal model resolution in a flood event storyline approach in western norway *Weather Clim. Extremes* **29** 100259
Schlunegger S *et al* 2020 Time of emergence and large ensemble intercomparison for ocean biogeochemical trends *Glob. Biogeochem. Cycles* **34** e2019GB006453
Schöngart J and Junk W J 2007 Forecasting the flood-pulse in central Amazonia by ENSO-indices *J. Hydrol.* **335** 124–32
Schutgens N, Tsyro S, Gryspeerdt E, Goto D, Weigum N, Schulz M and Stier P 2017 On the spatio-temporal representativeness of observations *Atmos. Chem. Phys.* **17** 9761–80
Sena J A, de Deus L A B, Freitas M A V and Costa L 2012 Extreme events of droughts and floods in Amazonia: 2005 and 2009 *Water Resour. Manage.* **26** 1665–76
Sperna Weiland F C, Vrugt J A, van Beek R P H, Weerts A H and Bierkens M F P 2015 Significant uncertainty in global scale hydrological modeling from precipitation data errors *J. Hydrol.* **529** 1095–115
Stainforth D, Allen M, Tredger E and Smith L 2007 Confidence, uncertainty and decision-support relevance in climate predictions *Phil. Trans. R. Soc.* A **365** 2145–61
Sterl A, Bintanja R, Brodeau L, Gleeson E, Koenigk T, Schmith T, Semmler T, Severijns C, Wyser K and Yang S 2012 A look at the ocean in the EC-earth climate model *Clim. Dyn.* **39** 2631–57
Stevenson S, Timmermann A, Chikamoto Y, Langford S and DiNezio P 2015 Stochastically generated North American megadroughts *J. Clim.* **28** 1865–80
Stott P A *et al* 2016 Attribution of extreme weather and climate-related events *Wiley Interdiscip. Rev. Clim. Change* **7** 23–41
Suarez-Gutierrez L, Milinski S and Maher N 2021 Exploiting large ensembles for a better yet simpler climate model evaluation *Clim. Dyn.* **57** 2557–80
Suarez-Gutierrez L, Müller W A, Li C and Marotzke J 2020 Dynamical and thermodynamical drivers of variability in European summer heat extremes *Clim. Dyn.* **54** 4351–66
Sutanudjaja E H *et al* 2018 PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model *Geosci. Model Dev.* **11** 2429–53
Sutton R T 2019 Climate science needs to take risk assessment much more seriously *Bull. Am. Meteorol. Soc.* **100** 1637–42
Swain D L, Wing O E J, Bates P D, Done J M, Johnson K A and Cameron D R 2020 Increased flood exposure due to climate change and population growth in the United States *Earth's Future* **8** e2020EF001778
Tabari H, Hosseinzadehtalaei P, Thiery W and Willems P 2021 Amplified drought and flood risk under future socioeconomic and climatic change *Earth's Future* **9** e2021EF002295

Taylor K E, Stouffer R J and Meehl G A 2012 An overview of CMIP5 and the experiment design *Bull. Am. Meteorol. Soc.* **93** 485–98

Thiery W *et al* 2021 Intergenerational inequities in exposure to climate extremes *Science* **374** 158–60

Thompson V, Dunstone N J, Scaife A A, Smith D M, Hardiman S C, Ren H-L, Lu B and Belcher S E 2019 Risk and dynamics of unprecedented hot months in South East China *Clim. Dyn.* **52** 2585–96

Thompson V, Dunstone N J, Scaife A A, Smith D M, Slingo J M, Brown S and Belcher S E 2017 High risk of unprecedented UK rainfall in the current climate *Nat. Commun.* **8** 107

Towner J *et al* 2020 Attribution of Amazon floods to modes of climate variability: a review *Meteorol. Appl.* **27** e1949

Towner J, Cloke H L, Zsoter E, Flamig Z, Hoch J M, Bazo J, De Perez E C and Stephens E M 2019 Assessing the performance of global hydrological models for capturing peak river flows in the Amazon basin *Hydrol. Earth Syst. Sci.* **23** 3057–80

van den Brink H W, Können G P, Opsteegh J D, van Oldenborgh G J and Burgers G 2005 Estimating return periods of extreme events from ECMWF seasonal forecast ensembles *Int. J. Climatol.* **25** 1345–54

Van der Wiel K, Kapnick S B, van Oldenborgh G J, Whan K, Philip S, Vecchi G A, Singh R K, Arrighi J and Cullen H 2017 Rapid attribution of the August 2016 flood-inducing extreme precipitation in south Louisiana to climate change *Hydrol. Earth Syst. Sci.* **21** 897–921

Van der Wiel K, Kapnick S B, Vecchi G A, Smith J A, Milly P C D and Jia L 2018 100-Year lower mississippi floods in a global climate model: characteristics and future changes *J. Hydrometeorol.* **19** 1547–63

Van der Wiel K, Selten F M, Bintanja R, Blackport R and Screen J A 2020 Ensemble climate-impact modelling: extreme impacts from moderate meteorological conditions *Environ. Res. Lett.* **15** 034050

Van der Wiel K, Stoop L P, van Zuijlen B R H, Blackport R, van den Broek M A and Selten F M 2019a Meteorological conditions leading to extreme low variable renewable energy production and extreme high energy shortfall *Renew. Sustain. Energy Rev.* **111** 261–75

Van der Wiel K, Wanders N, Selten F M and Bierkens M F P 2019b Added value of large ensemble simulations for assessing extreme river discharge in a 2 °C warmer world *Geophys. Res. Lett.* **46** 2093–102

van Kempen G, van der Wiel K and Melsen L A 2021 The impact of hydrological model structure on the simulation of extreme runoff events *Nat. Hazards Earth Syst. Sci.* **21** 961–76

van Schaik E, Killaars L, Smith N E, Koren G, van Beek L P H, Peters W and van der Laan-luijkx I T 2018 Changes in surface hydrology, soil moisture and gross primary production in the Amazon during the 2015/2016 El Niño *Phil. Trans. R. Soc.* B **373** 20180084

Vano J A, Miller K, Dettinger M D, Cifelli R, Curtis D, Dufour A, Olsen J R and Wilson A M 2019 Hydroclimatic extremes as challenges for the water management community: lessons from oroville dam and hurricane harvey *Bull. Am. Meteorol. Soc.* **100** S9–14

Vautard R *et al* 2019 Evaluation of the HadGEM3-A simulations in view of detection and attribution of human influence on extreme events in Europe *Clim. Dyn.* **52** 1187–210

Wanders N *et al* 2019 High-resolution global water temperature modeling *Water Resour. Res.* **55** 2760–78

Warszawski L, Frieler K, Huber V, Piontek F, Serdeczny O and Schewe J 2014 The inter-sectoral impact model intercomparison project (ISI–MIP): project framework *Proc. Natl Acad. Sci.* **111** 3228–32

Weigel K *et al* 2021 Earth system model evaluation tool (ESMValTool) v2.0—diagnostics for extreme events, regional and impact evaluation, and analysis of earth system models in CMIP *Geosci. Model Dev.* **14** 3159–84

Wilby R L, Clifford N J, De Luca P, Harrigan S, Hillier J K, Hodgkins R, Johnson M F, Matthews T K R, Murphy C and Noone S J others 2017 The 'dirty dozen'of freshwater science: detecting then reconciling hydrological data biases and errors *Wiley Interdiscip. Rev. Water* **4** e1209

Wilby R L, Nicholls R J, Warren R, Wheater H S, Clarke D and Dawson R J 2011 Keeping nuclear and other coastal sites safe from climate change *Proc. Inst. Civ. Eng.-Civ. Eng.* **164** 129–36

Wilks D S 2011 *Statistical Methods in the Atmospheric Sciences* vol **100** International Geophysics Series (New York: Academic Press)

Wilks D S and Wilby R L 1999 The weather generation game: a review of stochastic weather models *Prog. Phys. Geogr. Earth Environ.* **23** 329–57

Zscheischler J *et al* 2020 A typology of compound weather and climate events *Nat. Rev. Earth Environ.* **1** 333–47

Zscheischler J, Fischer E M and Lange S 2019 The effect of univariate bias adjustment on multivariate hazard estimates *Earth Syst. Dyn.* **10** 31–43