Received: 4 November 2021 | Revised: 18 February 2022 | Accepted: 7 April 2022

DOI: 10.1002/met.2065

RESEARCH ARTICLE

Meteorological Applications
Science and Technology for Weather and Climate
Open Access
RMetS

An open workflow to gain insights about low-likelihood high-impact weather events from initialized predictions

T. Kelder[1] | T. I. Marjoribanks[2] | L. J. Slater[3] | C. Prudhomme[1,4,5] | R. L. Wilby[1] | J. Wagemann[4] | N. Dunstone[6]

[1]Geography and Environment, Loughborough University, Loughborough, UK

[2]School of Architecture, Building and Civil Engineering, Loughborough, UK

[3]School of Geography and the Environment, University of Oxford, Oxford, UK

[4]European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK

[5]UK Centre for Ecology and Hydrology, Wallingford, UK

[6]Met Office Hadley Centre, Exeter, UK

Correspondence
T. Kelder, Geography and Environment, Loughborough University, Loughborough, UK.
Email: t.kelder@lboro.ac.uk

Funding information
Loughborough University; European Union Copernicus Programme; Department for Business, Energy and Industrial Strategy, UK Government, Grant/Award Number: Met Office Hadley Centre Climate Programme

Abstract

Low-likelihood weather events can cause dramatic impacts, especially when they are unprecedented. In 2020, amongst other high-impact weather events, UK floods caused more than £300 million damage, prolonged heat over Siberia led to infrastructure failure and permafrost thawing, while wildfires ravaged California. Such rare phenomena cannot be studied well from historical records or reanalysis data. One way to improve our awareness is to exploit ensemble prediction systems, which represent large samples of simulated weather events. This 'UNSEEN' method has been successfully applied in several scientific studies, but uptake is hindered by large data and processing requirements, and by uncertainty regarding the credibility of the simulations. Here, we provide a protocol to apply and ensure credibility of UNSEEN for studying low-likelihood high-impact weather events globally, including an open workflow based on Copernicus Climate Change Services (C3S) seasonal predictions. Demonstrating the workflow using European Centre for Medium-Range Weather Forecasts (ECMWF) SEAS5, we find that the 2020 March–May Siberian heatwave was predicted by one of the ensemble members; and that the record-shattering August 2020 California-Mexico temperatures were part of a strong increasing trend. However, each of the case studies exposes challenges with respect to the credibility of UNSEEN and the sensitivity of the outcomes to user decisions. We conclude that UNSEEN can provide new insights about low-likelihood weather events when the decisions are transparent, and the challenges and sensitivities are acknowledged. Anticipating plausible low-likelihood extreme events and uncovering unforeseen hazards under a changing climate warrants further research at the science-policy interface to manage high impacts.

KEYWORDS

climate change, climate model ensemble, climate risk, Copernicus Climate Change Services, seasonal predictions, Weather extremes

1 | INTRODUCTION

Understanding the likelihood, trends, and driving processes of extreme hydro-meteorological events is crucial for decision-making (Salas et al., 2018; Slater et al., 2021). However, it is challenging to compute robust statistics for low-likelihood weather events from short historical records, especially in data scarce regions. Instrumental records are typically only a few decades long and are not available everywhere (e.g., Alexander, 2016). Reanalysis products are increasingly employed to estimate extremes, as they blend observational datasets with model simulations into spatially and temporally coherent outputs, that is, 'maps without gaps' (e.g., European Centre for Medium-Range Weather Forecasts [ECMWF], 2018). For example, the ECMWF ERA5 reanalysis (Hersbach et al., 2020) has been used to estimate rainfall intensity–duration–frequency (IDF) curves globally (Courty et al., 2019), trends in extremes (Faranda, 2020; Geirinhas et al., 2021; Kim et al., 2021), driving processes behind extreme events (Grazzini et al., 2020), and extreme weather indices (Kennedy-Asser et al., 2021; Wehner et al., 2020). Although reanalyses overcome spatial data scarcity, they can exhibit model deficiencies or inhomogeneities (Parker, 2016), and their typical length (~70 years) may still be a limiting factor when studying extreme events.

Many approaches have been developed to reduce sampling uncertainties, ranging from traditional statistical weather generators (Brunner & Gilleland, 2020; Wilks & Wilby, 1999; Yiou, 2014), extreme value approaches (Coles, 2001; Katz, 2013), and dynamical systems theory (De Luca et al., 2020; Faranda et al., 2017); through pooling of observations (e.g., Berghuijs et al., 2017; Robinson et al., 2021), the use of long archives (Hawkins et al., 2019; Murphy et al., 2020), and paleoclimatic records (Yan et al., 2020); to probing ensemble members from weather and climate models (Box 1). Ensembles of opportunity (e.g., King et al., 2017; Lewis et al., 2017), single-model initial-condition large ensembles (SMILEs) (e.g., Suarez-Gutierrez et al., 2020a, 2020b), ensemble reinitialization methods (e.g., Gessner et al., 2021), and targeted large ensemble experiments (e.g., Guillod et al., 2017; Hall et al., 2019; Mitchell et al., 2017), have all been used for the study of low-likelihood high-impact hydro-climatic events.

From the many datasets and methods available, here, we build on the UNprecedented Simulated Extremes using ENsembles (UNSEEN) approach (Thompson et al., 2017). UNSEEN uses initialized ensemble predictions to assess present climate risks from extreme weather events. Initialized predictions provide an opportunity to assess plausible high-impact weather events as they contain a larger sample size than observations or reanalysis, provide physically plausible limits whereas statistical approaches might not, and are perceived more realistic than global climate models. Weather prediction systems are generally of a higher global resolution than global climate models, benefit from continuous evaluation at weather services, and are initialized from

BOX 1 Ensemble pooling for the study of low-likelihood high-impact hydro-climatic events

By treating model ensemble members as different, but equally plausible versions of the past, then pooling them, the sample size of weather events can be increased to explore the characteristics of rare extreme events (Allen, 2003; van den Brink et al., 2005). So far, this approach has helped estimate the likelihood of heavy precipitation (Jain et al., 2020; Kelder et al., 2020; Kent et al., 2022; Thompson et al., 2017), floods (e.g., Brunner & Slater, 2022; van den Brink et al., 2004), droughts (Kent et al., 2017; Kent et al., 2019; Pascale et al., 2020), wind losses (Osinski et al., 2016; Walz & Leckebusch, 2019), heatwaves (Cowan et al., 2020; Kay et al., 2020; Thompson et al., 2019), and fire weather (Squire et al., 2021). Furthermore, ensemble members have been used to evaluate compound hazards (Bevacqua et al., 2021; Hillier & Dixon, 2020) and to detect trends in rare extreme events over past decades (Diffenbaugh et al., 2017; Kay et al., 2020; Kelder et al., 2020; Kirchmeier-Young & Zhang, 2020) and in future projections (e.g., King et al., 2017; Lehner et al., 2017; Suarez-Gutierrez et al., 2020a, 2020b; Swain et al., 2020). Ensemble members from weather and climate models have been used across timescales, ranging from weeks (Breivik et al., 2013, 2014; Meucci et al., 2018; Osinski et al., 2016), through months (Hillier & Dixon, 2020; Jain et al., 2020; Kelder et al., 2020; van den Brink et al., 2004, 2005; Walz & Leckebusch, 2019), years (Cowan et al., 2020; Dunstone et al., 2016; Guillod et al., 2017; Kay et al., 2020; Kent et al., 2017, 2019; Thompson et al., 2017, 2019; Van der Wiel et al., 2019, 2020; van Kempen et al., 2021), and decades (Mitchell et al., 2017; Poschlod et al., 2021) to centuries (Bhatia & Ganguly, 2019; King et al., 2017; Lehner et al., 2017; Stevenson et al., 2015; Swain et al., 2020; Van der Wiel et al., 2018).

observations and reanalysis, therefore remain closer to the observed climate than uninitialized projections (Meehl et al., 2021).

Whereas event attribution methods raise awareness about the anthropogenic influence of high-impact weather events that *have* occurred (e.g., Allen, 2003; Philip et al., 2020; Stott et al., 2016) and large ensembles of climate model simulations help project future high-impact events (e.g., Deser et al., 2020; Mankin et al., 2020), UNSEEN may raise awareness about low-likelihood high-impact weather events that *could* occur in the present climate. UNSEEN, therefore, has considerable potential to improve engineering design standards (e.g., Jain et al., 2020; Kent et al., 2022; Thompson et al., 2017), detecting and explaining trends in rare extremes over recent decades (e.g., Kay et al., 2020; Kelder et al., 2020), and developing event-based storylines of plausible—yet unseen—weather extremes (Matthews et al., 2016; Sillmann et al., 2021). However, uptake is hindered by large data and processing requirements, and there are challenges associated with providing credible, *multiple* 'maps without gaps' from initialized predictions (Box 2).

This paper presents a protocol and open workflow to apply and ensure credibility of UNSEEN for studying low-likelihood high-impact weather events globally (Figure 1). The procedure begins by selecting the type of hydro-meteorological event of interest with requisite spatial and temporal scales (step 1). The type of event being studied informs the selection of the most appropriate prediction system (step 2). We then discuss how data can be retrieved (step 3) and pre-processed (step 4) before being evaluated (step 5). For events that are not deemed credible, practical solutions are discussed (step 6). The final step is to apply statistics and gain insights about plausible

BOX 2 Three challenges associated with generating an UNSEEN ensemble

The credibility of initialized ensemble predictions to represent large samples of weather events hinges on three common challenges faced by all prediction systems: the independence of the ensemble members; the stability of the model; and the fidelity of the simulations (Kelder et al., 2020; Thompson et al., 2017, 2019).

Independence. Ensemble member independence (i.e., the uniqueness of each model ensemble member) is closely linked to the spread and predictability of forecasts. When a forecast is initialized, the ensemble member independence is low because the ensemble members only differ slightly in their initial conditions. The spread of the individual ensemble members increases over the forecasting horizon because they develop their own 'virtual world' induced by stochastic processes in the atmosphere. The importance of ensemble member independence can be best explained through an example of extreme value analysis. Dam safety standards require preparedness for very unlikely scenarios, such as the 10,000-year inflow return value. Large ensemble hindcasts might be used to generate an UNSEEN ensemble that can capture such events. However, if the UNSEEN ensemble members are correlated, one might think that 10,000 years were simulated adequately, whereas the effective ensemble size is in practice much smaller.

Stability. Ensemble members may drift away from their initial climatology (near to an observed state) towards a steady virtual climatology (Covey et al., 2006; Hermanson et al., 2018; Sen et al., 2009, 2013). Such drifts are not caused by external forcing or internal low-frequency variability but by numerical errors (e.g., Liepert & Previdi, 2012; Lucarini & Ragone, 2011), model imbalances and/or discontinuities (e.g., Rahmstorf, 1995). Drift is mostly present in physical ocean variables, but can be evident in atmospheric properties (Sen et al., 2013). Hence, model instability (i.e., the presence of drift) may deteriorate the realism of the hindcast ensemble.

Fidelity. Model simulations are virtual representations of reality, and 'fidelity' refers to their ability to realistically simulate the target event(s). Hence, for robust analyses using climate model simulations, their 'virtual world' must realistically describe 'reality', that is, the extreme event being studied. Systematic errors such as in cloud microphysics, tropical cyclones, convective precipitation, teleconnections, and synoptic regimes in numerical prediction systems may bias the simulation of extreme events (e.g., Zadra et al., 2018). Processes that occur on scales smaller than the model grid cannot be resolved but must be parameterized, leading to lower fidelity (Sillmann et al., 2017). Furthermore, mechanisms such as self-intensification of droughts via land-atmosphere feedbacks are currently not well-represented by climate models (Miralles et al., 2019). Therefore, an evaluation of model fidelity is crucial.

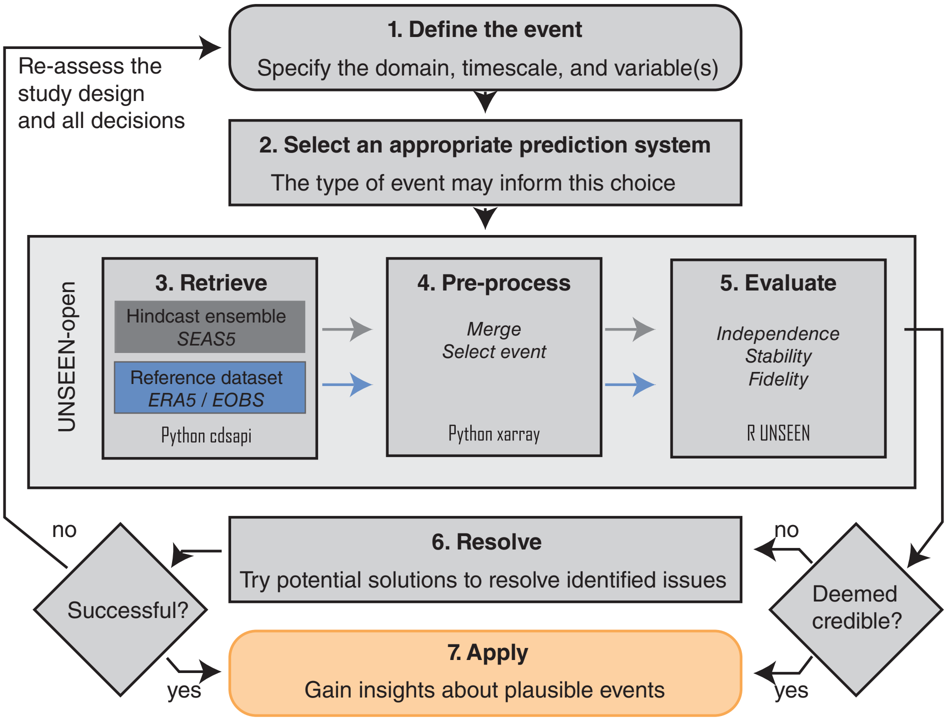
Re-assess the study design and all decisions
1. Define the event
Specify the domain, timescale, and variable(s)
2. Select an appropriate prediction system
The type of event may inform this choice
UNSEEN-open
3. Retrieve
Hindcast ensemble SEAS5
Reference dataset ERA5 / EOBS
Python cdsapi
4. Pre-process
Merge
Select event
Python xarray
5. Evaluate
Independence
Stability
Fidelity
R UNSEEN
no
6. Resolve
Try potential solutions to resolve identified issues
no
Successful?
Deemed credible?
yes
7. Apply
Gain insights about plausible events
yes

FIGURE 1 Protocol to ensure credibility and apply UNSEEN for gaining insights about low-likelihood high-impact events. Grey boxes indicate the six steps of the protocol ensuring credibility of large samples of weather events obtained from initialized predictions for the user-defined domain, timescale, and variable(s). The orange box indicates the final step to gain insights about plausible weather events, after a successful outcome of a credible large ensemble. The larger light grey box surrounding steps 3–5 shows the UNSEEN-open workflow, documented at https://unseen-open.readthedocs.io. For these steps, the programming language and package are indicated.

weather events, if the large ensemble was deemed credible or identified issues could be resolved (step 7). A technical workflow (UNSEEN-open, documented at https://unseen-open.readthedocs.io) was developed for steps 3–5 during the *ECMWF Summer of Weather Code* 2020 (https://esowc.ecmwf.int/). This workflow facilitates the process of retrieving, pre-processing, and evaluating the latest ECMWF seasonal prediction system 5 (SEAS5, Johnson et al., 2019) but could be adapted for other modelling systems. Through the protocol and workflow, this paper aims to transparently detail the data and code that are applied, decisions that were made, challenges that were faced, and sensitivities that may influence the outcome of the UNSEEN approach.

Following sections step through the protocol, illustrated by three worked examples of extreme events that occurred in 2020. During February, the United Kingdom endured floods that caused more than £300 million in damage and destroyed 3400 houses (Copernicus EMS, 2020). Later that year, prolonged heat over Siberia caused wildfires, invasion of pests, and infrastructure failure, as well as global impacts through the release of greenhouse gasses from thawing permafrost (Ciavarella et al., 2021; Overland & Wang, 2021). Meanwhile, wildfires in California contributed to the (then) worst fire season on record (Pickrell & Pennisi, 2020), amplifying hardships faced by communities during the COVID-19 pandemic (Moore et al., 2020). Section 2 describes the seven steps of the protocol. Section 3 provides an overview for each of the three case studies. In Section 4, we discuss the challenges and sensitivities to be mindful of when applying UNSEEN. Section 5 concludes the paper with final remarks.

2 | THE PROTOCOL

2.1 | Step 1: Define the event

To apply the UNSEEN method, a hydro-meteorological event is first defined. The event definition depends on the scope of the analysis in terms of the target domain, timescale, and (meteorological) variable of interest. Any domain, timescale, and variable(s) can be selected, for example, to estimate design values or to quantify the likelihood of unprecedented events. The event definition should reflect the impact being studied. For example, larger spatial and temporal scales are used to study winter precipitation (Thompson et al., 2017) than to study summer precipitation over the United Kingdom (Kent et al., 2022), reflecting widespread flooding from winter storms as opposed to more localized convective storms in summer. In the examples below, events were defined to best represent the footprint of historical low-likelihood high-impact weather incidents because we want to assess whether these phenomena could have been 'seen' before they occurred. For some cases, the definition is straightforward, such as for studying UK-average extreme precipitation in February (Section 3.3). In other cases, such as for the Siberian heatwave (Figure 2a,c and Section 3.1)

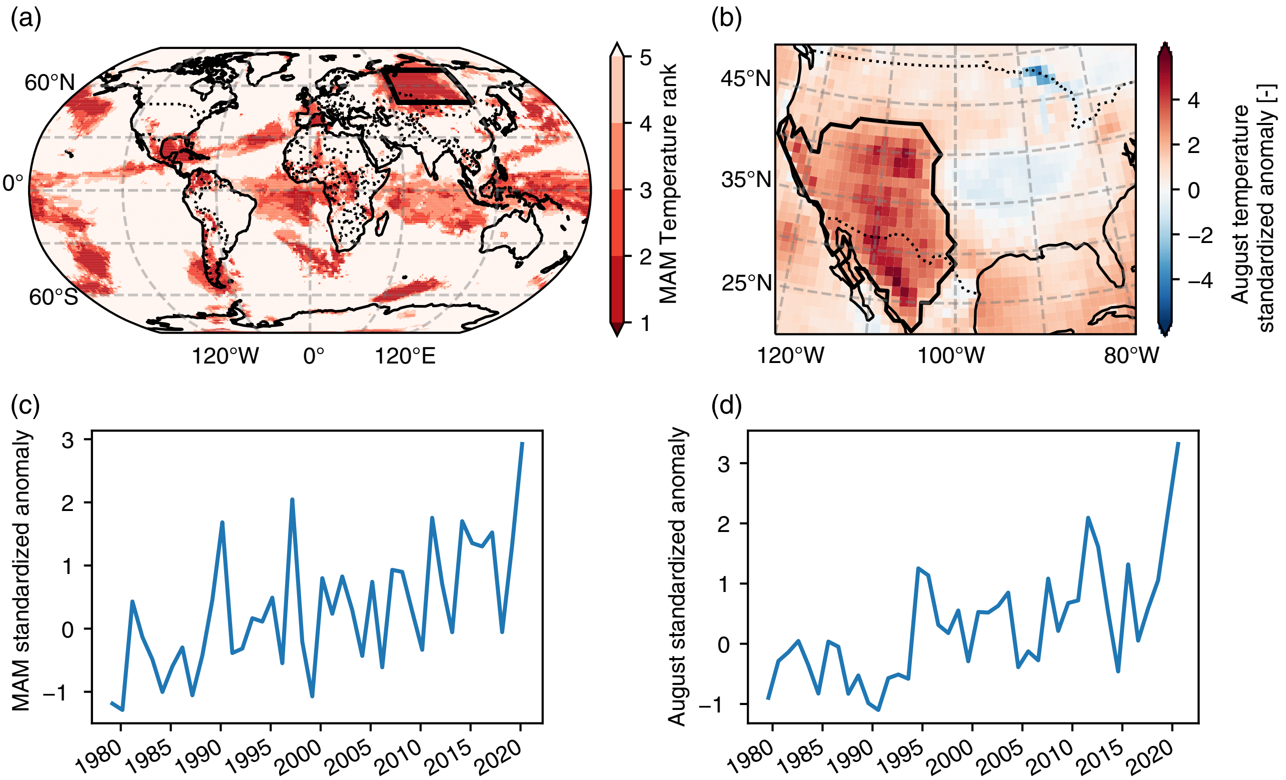

FIGURE 2 The domains and temperature anomalies for extreme heat over Siberia (a,c) and California-Mexico (b,d). (a) March–May 2020 temperature rank within the 1979–2020 ERA5 record. Rank 1 means that temperature records were broken in 2020. (b) August 2020 standardized temperature anomaly with respect to ERA5 1979–2010 climatology. Standardized anomalies are calculated by subtracting the mean and dividing by the standard deviation. Thick black lines indicate the selected domains, which for (a) is the region where March–May 2020 temperature records were broken over Siberia, and for (b) is where August temperature anomalies exceeded twice the climatological standard deviation. (c,d) The standardized temperature anomalies averaged over the domains indicated in (a,b).

and temperature anomalies during peak California wildfire activity (Figure 2b,d and Section 3.2), the domain and timescale may be informed by an assessment of the event anomaly[1] or the region where historical records were broken.[2] Detailed protocols for defining the extent, timescale, and meteorological variable representative of target events can be found in Philip et al. (2020) and van Oldenborgh et al. (2021).

2.2 | Step 2: Select an appropriate prediction system

Increasing computational resources and improved physical understanding of the Earth System have led to advances in seamless prediction systems over recent decades (Alley et al., 2019; Bauer et al., 2015; Hoskins, 2013; Palmer, 2019). The resolution and prediction timescale of ensemble prediction systems are important considerations to inform the choice of model (Figure 3). Predictions ranging from weeks to years provide high-resolution but independent events well suited for regional-scale multi-day to monthly events, such as heatwaves (Cowan et al., 2020; Kay et al., 2020; Thompson et al., 2019), cold spells, wind storms (Walz & Leckebusch, 2019), and extreme precipitation (Jain et al., 2020; Kelder et al., 2020; Thompson et al., 2017) (Figure 3, Table 1). For sub-daily extremes—such as ocean wind and wave extremes, convective storms, or wind gusts—high-resolution simulations are required to resolve sub-grid processes (Sillmann et al., 2017). In general, global medium-range simulations (10–15 days) are likely to be most appropriate for studying local, short-duration events (e.g., Breivik et al., 2014, 2013; Meucci et al., 2018; Osinski et al., 2016, Table 1), or additional downscaling might be needed (e.g., Guillod et al., 2018; Poschlod et al., 2021). For events with long persistence such as droughts, the ensemble members from medium-range predictions are unlikely to be unique (low

[1] See https://unseen-open.readthedocs.io/en/latest/Notebooks/California_august_temperature_anomaly.html.
[2] See https://unseen-open.readthedocs.io/en/latest/Notebooks/Global_monthly_temperature_records_ERA5.html.

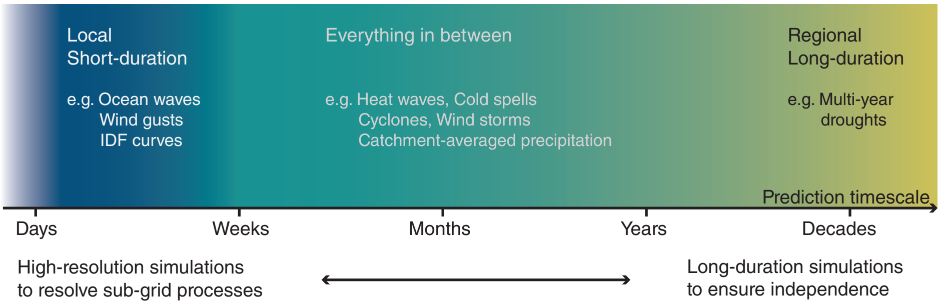

FIGURE 3 The appropriateness of hindcast ensembles for diverse types of events. This schematic shows which prediction systems are most likely to be appropriate for diverse types of extreme events. The horizontal axis represents seamless prediction timescales, where the arrows beneath indicate the different weather prediction systems covering the respective prediction timescales. The corresponding types of extreme event range from local, short-duration events requiring high-resolution simulations, to regional, persistent events involving long-duration simulations. The gradual shading indicates that multiple prediction systems might be equally appropriate for some type of events. Fading on the left-hand side is a reminder that the first few days of forecasts cannot be used because of low independence between ensemble members due to similar initial conditions. The schematic is based on Table 1.

TABLE 1 Hydro-meteorological extremes (variable) with spatial resolution and timescale that have been studied by pooling ensembles from medium-range, seasonal, and decadal prediction systems

Prediction timescale	Variable	Spatial resolution	Timescale	References
Medium-range (10–15 days)	Ocean wind speed and wave height; windstorms	0.1° × 0.1°; 0.25° × 0.25°	6 h	Breivik et al. (2014), Breivik et al., (2013), Osinski et al. (2016), Meucci et al. (2018), and
Extended range (22–46 days)	Floods	5 km× 5 km	Day	Brunner and Slater (2022)
Seasonal (6 months)	Rainfall; wind losses; river discharge	0.4° × 0.4°; 1° × 1°	6 h; 3 days; season	van den Brink et al. (2005), van den Brink et al. (2004), Walz and Leckebusch (2019), Hillier and Dixon (2020), Jain et al. (2020), and Kelder et al. (2020)
Decadal (1–10 years)	Rainfall; temperature; water shortage; drought	0.5° × 0.5°; 1.875° × 1.25° GCM, 0.22° × 0.22° RCM	Day; month; season	Thompson et al. (2019), (2017), Kent et al. (2019), (2017), Kay et al. (2020), (2018), Hall et al. (2020), (2019)

Note: We present the spatial resolution of the most recent prediction systems for consistency, but some of the cited studies may have used earlier systems with lower resolutions. GCM, Global Climate Model; RCM, Regional Climate Model.

independence, Box 2). Hence, decadal predictions (1–10 years) are recommended for events with long memory (e.g., Hall et al., 2019, 2020; Kay et al., 2018).

2.3 | Step 3: Retrieve the ensemble hindcast and reference dataset

The UNSEEN-open technical workflow was developed for steps 3–5 of the protocol with a focus on the SEAS5 prediction system, but with other systems from Copernicus Climate Change Services (C3S) also in mind (see Figure 1 and supporting technical documentation: https://unseen-open.readthedocs.io). The protocol is applicable to any prediction system, while the code and guidance for UNSEEN-open is developed to work with the Copernicus Data Store (CDS, https://cds.climate.copernicus.eu/). For the case studies presented here, we retrieve all relevant SEAS5 forecasts, ERA5 reanalysis, and EOBS observational data from CDS via a Python API (https://pypi.org/project/cdsapi/). Jupyter notebooks showing how the data are retrieved are available at https://unseen-open.readthedocs.io/en/latest/Notebooks/1.Download/1.Retrieve.html. SEAS5 data dimensions and

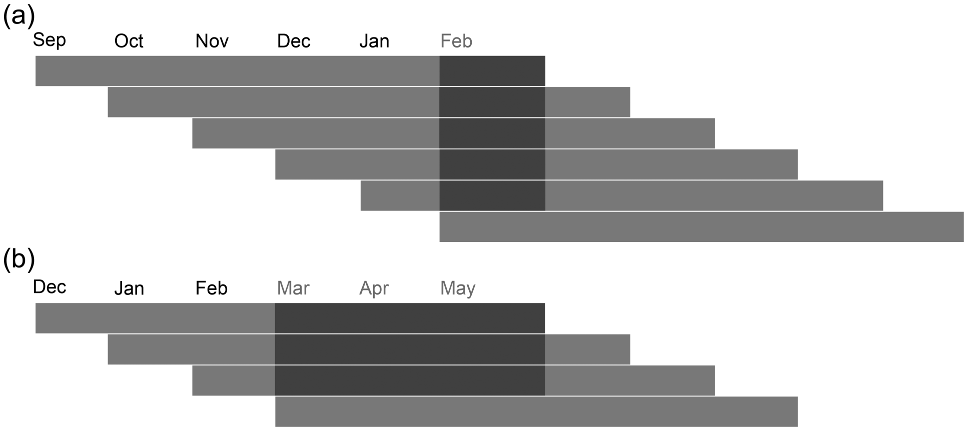

FIGURE 4 Schematic showing how forecasts initialized in different months can be pooled to extend the sample size for the same target event. The grey horizontal bars represent the seasonal forecasts, which are initialized every (leftmost) month and run for 6 months after their initialization. Dark shading indicates the relevant section of the target period (lead times greater than 1 month) used for the February UK precipitation (a) and March–May Siberian heat (b) case studies.

retrieval time are optimized by (1) retrieving pre-computed monthly statistics (minimum, maximum, or average) instead of retrieving all forecasts in full; (2) selecting only the target domain and months, then converting those into the relevant initialization months and lead times required for the request and; (3) optimizing retrieval functions to the structure of the ECMWF MARS archive (see the ECMWF documentation).

SEAS5 ensemble members and lead times are pooled to create the UNSEEN ensemble (e.g., Kelder et al., 2020). For example, UK February precipitation is forecasted from six initialization months (i.e., the preceding September to February, Figure 4a). For longer duration 'target events', such as March–April–May average temperature over Siberia, there are fewer forecasts that can be pooled together (from four initialization months, that is, the preceding December to March, Figure 4b). We discard the first month of the forecast because ensemble members are still likely to be overly constrained to initial conditions (Kelder et al., 2020). In the end, we are left with five initialization months for monthly blocks (such as the United Kingdom and California examples) and three initialization months for seasonal blocks (such as for the Siberia example). Pooling across the 25 ensemble members yields a potential increase to 125 (monthly blocks) and 75 (seasonal blocks) compared with a single observed period.

SEAS5 is, at present, the latest seasonal prediction system of ECMWF, launched in November 2017. SEAS5 is run on a 36 km horizontal resolution and is upscaled to a 1° grid to create a homogenous dataset with the same resolution for all C3S seasonal prediction systems. SEAS5 contains 51 ensemble members (25 members were available through C3S at the time of analysis). The historical seasonal predictions that are used to generate the UNSEEN ensemble consist of two datasets: archived operational forecasts since 2017 (years 2017–2020 are used) and hindcasts that were originally run to evaluate and calibrate the operational forecasts (years 1981–2016). Inhomogeneity between the hindcasts and forecasts is not expected, but can occur because of the differences in initialization: SEAS5 hindcasts are initialized from ERA-Interim (Dee et al., 2011) and OCEAN5 (Zuo et al., 2018), but the operational forecasts use ECMWF operational analyses instead of ERA-Interim. For further details on SEAS5, see the ECMWF page (https://confluence.ecmwf.int/display/CKB/C3S+Seasonal+Forecasts) and Johnson et al. (2019).

2.4 | Step 4: Pre-process the data

In the pre-processing step, the retrieved files are merged into one multi-dimensional dataset (Hoyer & Hamman, 2017). This dataset can be stored as a NetCDF file containing the dimensions latitude, longitude, ensemble members, time (years), and lead time (initialization months). Then, a domain and timescale representative of the event being studied is selected. In the workflow, the resulting data array (with *dimensions* ensemble members, time, and lead time) is converted to a data frame (with *variables* ensemble members, time, and lead time) and stored as a csv file to match ggplot functionalities in R. This step is provided in python and is run on a local machine.

2.5 | Step 5: Evaluate the independence, stability, and fidelity

In the evaluation step, ensemble member independence, model stability, and model fidelity are tested (Box 2).

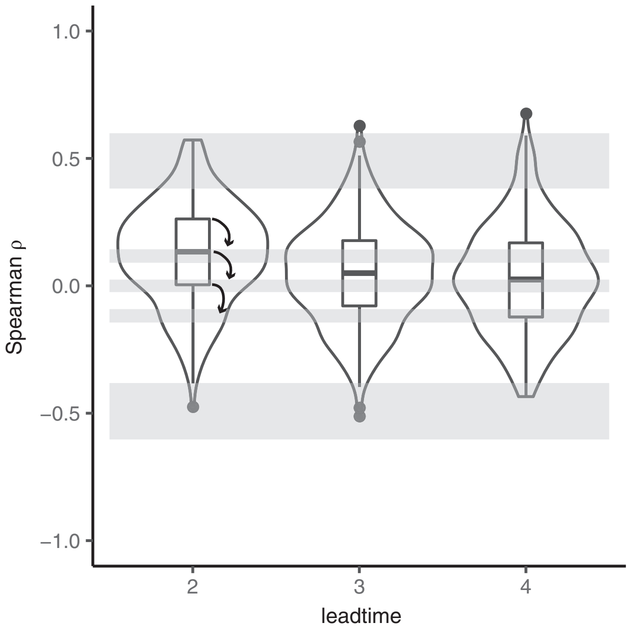

FIGURE 5 Example of an ensemble with inter-member dependence. Boxplots and violin plots show the correlation between all pairs of ensemble members for each lead time within the hindcast ensemble of March–May Siberian temperatures. Boxplots show the median, inter-quartile range, 1.5× interquartile range, and outliers of correlation values (small squares). Grey horizontal shading denotes the correlation (median, interquartile range, and 1.5× interquartile range) that is expected by chance. Arrows indicate for lead time 2, where the median and inter-quartile range are higher than would be expected by chance, revealing that the ensemble members are not fully independent.

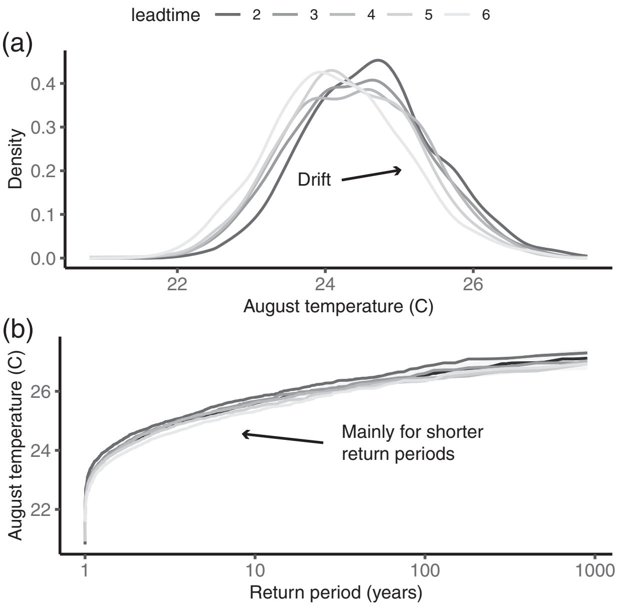

FIGURE 6 Testing the model stability for August California temperatures within the SEAS5 hindcast ensemble. The (a) probability density and (b) extreme value distributions are plotted for each lead time. Grey shading in (b) illustrates 95% confidence bounds. Arrows highlight the presence of model drift, most pronounced in lead time 6 for shorter return periods.

Thompson et al. (2017) developed the model fidelity test and Thompson et al. (2019) discussed the general applicability also in terms of the ensemble member independence and model stability. Kelder et al. (2020) then developed methods for the evaluation of the independence and stability for a case study of extreme precipitation events over Norway and Svalbard. Here, we build upon and extend these evaluation tests so they can be tailored to the selected event definition. We provide functions for testing the three criteria in the 'UNSEEN' R-package (https://github.com/timokelder/UNSEEN). We switch from python to R since we believe R has a better functionality in extreme value statistics. This section describes the evaluation tests.

Ensemble member *independence* is tested using a modification of the 'potential predictability' test—the ability of the forecast to predict itself (Kelder et al., 2020; Lavers et al., 2014; Wilks, 2011). The correlation between each ensemble member over the hindcast period is calculated, resulting in 300 distinct pairs per lead time (Kelder et al., 2020). A trend over the hindcast period can result in artificial detection of dependence. The independence test, therefore, includes detrending of values by first-differencing, whereby a new series is created from the differences between each successive value in a time series. Then, the non-parametric Spearman rank correlation between ensemble members is compared with the correlation arising by chance from uncorrelated members. This may be represented as a boxplot of the correlations between all pairs of ensemble members, with background values for each of the boxplot statistics given by those expected between all pairs of uncorrelated members.

For example, for the Siberian heat case study, the independence test shows that there is stronger correlation between ensemble members than would be expected by chance (Figure 5). The dependence between ensemble members is most pronounced for the shortest lead time used (recalling from Section 2.3 that the first month of the forecasts are removed to avoid dependence). The correlation is not caused by a trend because the time series have been detrended.

Model *stability* is tested by comparing distributions between the different lead times (Kelder et al., 2020), which is performed on the original, raw data. For example, for the California wildfire danger case study, we find that August temperatures tend to drift over forecast lead time (Figure 6). First, the probability density function is plotted for each lead time (Figure 6a). This shows that lead time 6 seems to be colder for the tail of the distribution, which contains the extreme values of interest. Then, an empirical extreme value distribution is plotted (Figure 6b), which focuses more on the tail of the distribution. The extreme value distributions show that the drift is less pronounced for rare events. For more details about the model stability test, refer to Kelder et al. (2020).

Model *fidelity* is tested by evaluating the consistency between the hindcast ensemble and a reference dataset. For illustrative purposes, the hindcast ensemble for February rainfall is bootstrapped into 10,000 series of equal length to the reference dataset, with all lead times pooled together (Kelder et al., 2020; Thompson et al., 2017). The mean, standard deviation, skewness, and kurtosis are calculated for each of the series. Histograms of these distribution characteristics are plotted, including their 95% confidence interval. The range of the distribution characteristics within the hindcast ensemble can then be compared with the reference dataset (Figure 7).

2.6 | Step 6: Resolve detected issues

If the above three tests are passed, the ensemble is considered credible for applications (Figure 1). However, if one or more tests fail, identified issues need to be resolved prior to further use. This section discusses potential solutions to resolve the issues, which are summarized in Table 2.

When the independence test or stability tests are failed, the simplest solution is to remove the problematic lead times (Solution I1 in Table 2). If ensemble member *dependence* cannot be corrected by removing problematic lead times—for example, when dependence persists across all lead times—it is possible to assess whether forecasts are over-dispersive or under-dispersive (Solution I2 in Table 2) by calculating the signal-to-noise ratio and/or the relationship between ensemble mean root-mean-square error (RMSE) and ensemble spread (e.g., Weisheimer et al., 2019). Another desirable (but not always practical) approach is to assess the spread in large-scale physical drivers and surface states relevant to the hydro-climatic extreme being studied (Solution I3), such as sea-surface temperatures, sea-ice conditions, soil moisture, or atmospheric patterns. The spread shows the extent to which the ensemble is tied to slowly varying properties within the prediction systems. A bounded ensemble can still provide valuable information. In fact, many weather generators are constructed to be constrained and bounded to typical

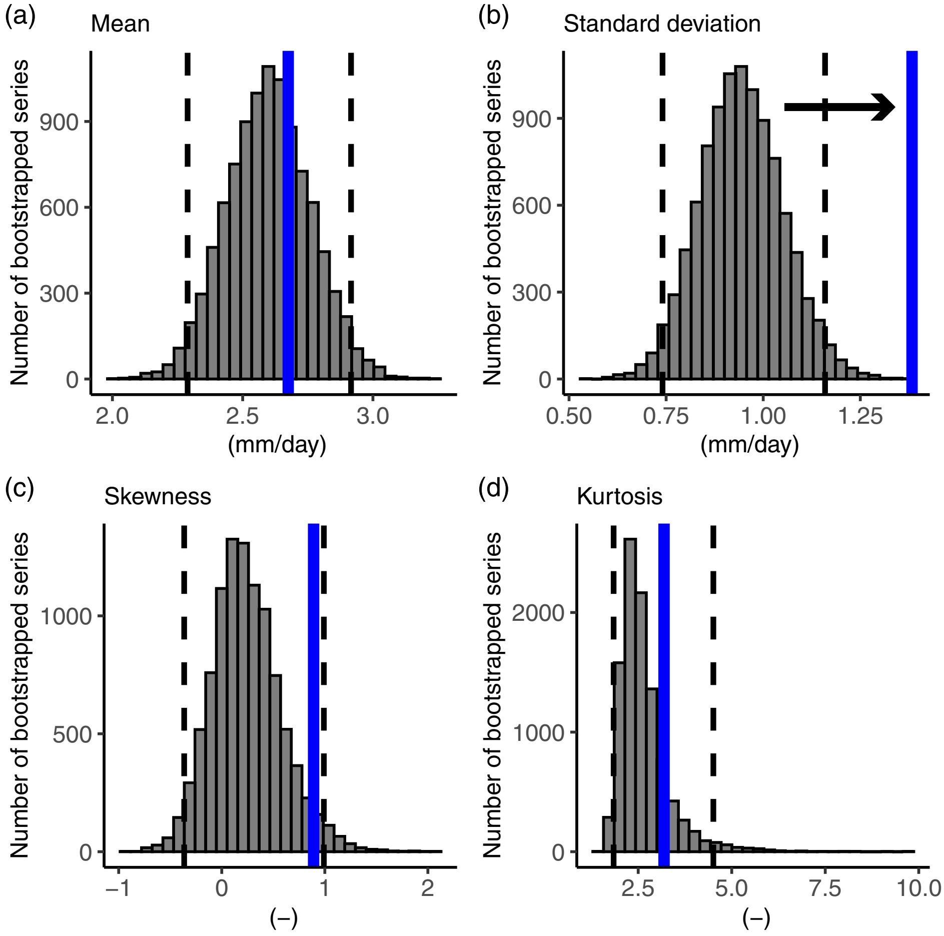

FIGURE 7 Testing the fidelity of UK February precipitation simulations within the SEAS5 hindcast. Distribution characteristics of SEAS5 are compared with EOBS. Histograms show the distribution of the (a) mean (b) standard deviation (c) skewness and (d) kurtosis for SEAS5, including 95% confidence intervals (dashed lines). EOBS statistics are derived for the period 1981–2016 (blue lines). The arrow denotes where the moment from EOBS lies outside the confidence interval of the SEAS5 ensemble—in this case, for the standard deviation.

TABLE 2 Potential solutions when issues with ensemble member independence, stability, or fidelity are detected

Independence	Stability	Fidelity
Solution I1: Remove problematic lead times	Solution S1: Remove problematic lead times	Solution F1: Additive/ multiplicative adjustment
Solution I2: Assess whether forecasts are over-dispersed or under-dispersed	Solution S2: Scale individual lead times	Solution F2: Apply other evaluation tests
Solution I3: Assess the spread in large-scale physical drivers		Solution F3: Evaluate drivers and feedback processes
Solution I4: Calculate the effective sample size		Solution F4: Advanced bias adjustments

weather types. Therefore, predictability is only an issue when it originates from the initial conditions. Initial-condition predictability implies that the ensemble members are not unique, whereas predictability from boundary conditions means that the ensemble members are unique but conditioned. Note that for events with short memory and low persistence, no initial-condition predictability is expected beyond 2 weeks (Lorenz, 1963). Finally, the option to calculate the *effective* sample size (Solution I4) is recommended when dependence remains an issue (Breivik et al., 2013). The effective sample size represents the size of the dependent sample that an independent sample would have. For example, an ensemble consisting of 1000 years of weather events containing some dependence may *effectively* represent only 500 unique, independent years. For an ensemble with sample size (N) that expresses dependence (correlation between ensemble members, r), the effective sample size (N^*) can be calculated following Breivik et al. (2013):

$$N^* = \frac{N}{1 + (N-1) * r}.$$

If model *stability* is an issue and cannot be corrected by removing problematic lead times (Solution S1 in Table 2), an option could be to scale each lead time to the distribution of the shortest lead time (Solution S2).

When model *fidelity* is an issue, additive (for temperature) or multiplicative (for precipitation) adjustment may be applied (Jain et al., 2020; Kelder et al., 2020; Thompson et al., 2019) (Solution F1). If issues with model fidelity remain, it is recommended to apply other evaluation tests (Solution F2) plus assess large-scale drivers and land surface feedbacks related to the extreme event (Solution F3).

In this workflow, the fidelity test was used for its focus on rare extremes. The sensitivity of the model fidelity results to the method of assessment can be tested (Solution F2). A wide range of methods and tools to identify biases in the simulation of extreme events exist (Eyring et al., 2019) that can be applied as tests for UNSEEN applications. For example, the 'ESMValTool' has been developed for climate model evaluation (Eyring et al., 2016) including extreme events (Weigel et al., 2021). Furthermore, metrics common to the evaluation of numerical weather prediction systems, such as the forecast reliability and rank histograms, can be used for prediction systems across timescales (Bellprat et al., 2019; Palmer & Weisheimer, 2018; Suarez-Gutierrez et al., 2021; Weisheimer & Palmer, 2014). In addition to the statistical evaluation methods presented so far, it may be desirable to evaluate the large-scale drivers and feedback processes of the extreme events (Solution F3) and how they are represented in the model (e.g., van der Wiel et al., 2017; Vautard et al., 2019). For example, Kay et al. (2020) and Thompson et al. (2019) assessed the large-scale drivers of simulated unseen temperature events. When inconsistencies in the variability (standard deviation) or shape (skewness and kurtosis) remain, more advanced correction methods can be applied with caution, such as the Inter-Sectoral Impact Model Intercomparison Project (ISI–MIP, Warszawski et al., 2014) bias adjustment approach (Hempel et al., 2013; Lange, 2019), which is commonly used to study climate impacts (e.g., Mitchell et al., 2017). For more guidance on bias adjustment methods, see for example Cannon et al. (2020) and Maraun and Widmann (2018).

2.7 | Step 7: Apply statistics

When the previous steps resulted in a credible large ensemble for analysis, extreme value theory can be applied to gain insight into low-likelihood events. This section discusses the steps that we take to analyse the event statistics.

One way to determine whether UNSEEN could have detected historical low-likelihood weather events is to simply assess whether one (or multiple) of the ensemble members exceeds the magnitude of the historical event. The likelihood of the historical event can then be estimated as the percentage of the ensemble members that exceeded the threshold (e.g., Thompson et al., 2017). It is also possible to compare the UNSEEN ensemble member with the highest magnitude to the magnitude of the historical event to demonstrate how severe an event could get based on a worst-case scenario from UNSEEN (e.g., Walz & Leckebusch, 2019). Furthermore, a threshold can be selected based on system vulnerabilities. For example, for the Siberian heatwave, we count the number of thawing events (MAM mean temperature >0°C) within the UNSEEN ensemble and the reference data.

Another, more advanced way is to apply extreme value theory (Coles, 2001). We demonstrate this method for the record-shattering California-Mexico August temperatures. We use the UNSEEN-trends approach (Kelder et al., 2020) to estimate changes in the event by fitting a non-stationary generalized extreme Value (GEV) distribution (Katz, 2013) to the pooled UNSEEN data and to the reference data, excluding the 2020 event. As in Kelder et al. (2020), we allow the location and scale parameters of the GEV distribution to vary linearly with time, whereas the shape parameter is assumed constant over time. In addition, we fit a stationary GEV distribution and test which distribution (stationary or non-stationary) better fits the data using a likelihood-ratio test. The parameters of the GEV distributions are estimated using maximum likelihood estimation (MLE) and statistical uncertainty is estimated as 95% confidence intervals based on the normal approximation, employing the extRemes package in R (Gilleland & Katz, 2016). The resulting distribution can then be used to determine the likelihood of the historical events, and whether, and by how much, the frequency and magnitude of such events have changed over recent decades.

3 | CASE STUDIES

We now present three case studies where we apply the UNSEEN protocol to the 2020 Siberian heatwave, temperature anomalies during peak California wildfire activity, and UK extreme precipitation events. We describe the steps taken to generate and evaluate the UNSEEN ensemble for each of the case studies. When issues are identified, the options to resolve them are discussed and appropriate solutions applied.

3.1 | Siberian heatwave

The detailed technical steps involved in producing this example may be followed at https://unseen-open.readthedocs.io/en/latest/Notebooks/examples/Siberian_Heatwave.html.

For the Siberian heat case study, our choice of domain and duration was informed by the location and season in which monthly temperature records were broken (Section 2 and Figure 2a). We selected the area bounded by 65–120°E, 50–70°N for the March–May (MAM) season. Seasonal predictions (SEAS5) were selected as the hindcast ensemble and reanalysis (ERA5) was chosen for reference data. All forecasts simulating March–May monthly temperatures were retrieved and pre-processed (averaged and merged) to represent the event definition. Time series show that the 2020 event was the highest within the ERA5 record and exceeded the simulations within the UNSEEN ensemble (blue cross compared with grey boxplot in Figure 8a). One interpretation is that the 2020 event was rarer than the 75 ensemble members within UNSEEN; another is that UNSEEN does not represent the true likelihood of such an extreme event. Therefore, an evaluation of the applicability of UNSEEN for this event definition is crucial.

We find some ensemble member dependence for lead time 2 (Figure 8b) but no drift over any lead times (Figure 8c–d). The fidelity test using all lead times shows that there is a cool bias in SEAS5 MAM temperatures compared with ERA5 (Figure 8e). The standard deviation, skewness, and kurtosis are within the 95% confidence intervals, but the standard deviation is at the lower end (Figure 8f–h). Note that the difference between SEAS5 and ERA5 could also be due to temperature overestimation by ERA5 for this particular season and domain. However, Ciavarella et al. (2021) report little difference between ERA5 and GISTEMP 250-km anomalies (Hansen et al., 2010) for the Siberian heat event.

The UNSEEN March–May average Siberian temperature ensemble based on SEAS5 expresses low ensemble member dependence for lead time 2, a cool mean bias, as well as low variability compared with ERA5. Lead time 2 could be removed from the ensemble to avoid dependence (Solution I1 in Table 2) but we choose to keep the ensemble members to retain a large sample size, because the low median correlation values of ~0.2 fall within accepted correlation values of <0.25 previously found for floods (Brunner & Slater, 2022). As a result of this decision, the effective sample size may be slightly lower than the 75 members being used because of the low dependence between ensemble members (see ‘independence’ in Box 2). The dependence and the sensitivity of the dependence result to other tests could be further assessed (Solutions I2–4 in Table 2) but is not deemed necessary in this case. We apply a mean bias adjustment (Solution F1) to solve the issue with the cold bias of the UNSEEN ensemble (Figure S1). The UNSEEN ensemble remains conservative with a low SEAS5 standard deviation (although not statistically significant from ERA5 at the 95% confidence level). Further evaluation tests can be applied (Solution F2), and large-scale drivers and land surface feedbacks that might be unrealistic can be

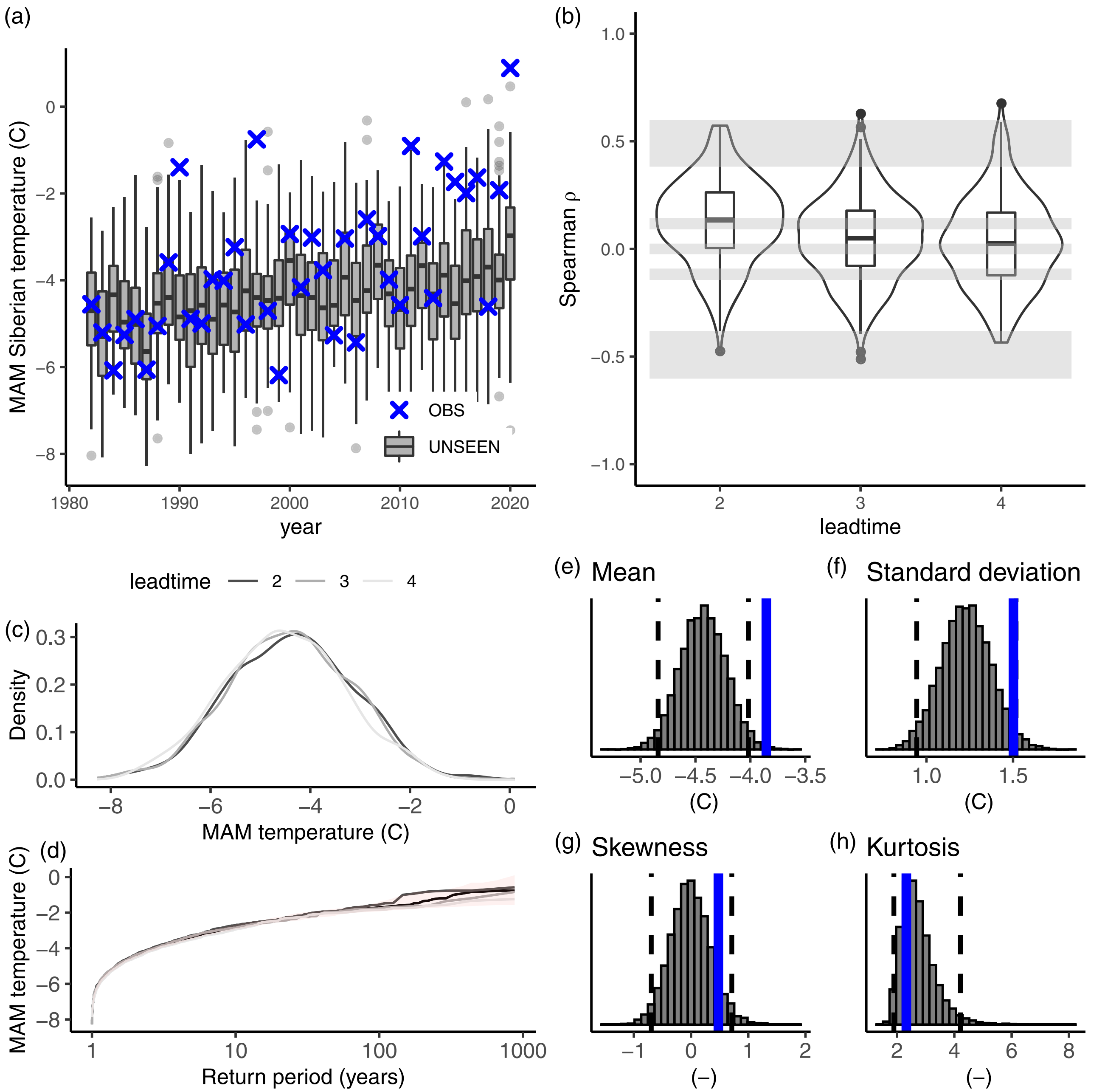

FIGURE 8 The (a) time series (b) independence (c–d) stability and (e–h) fidelity of the ensemble for March–May 2020 Siberian temperatures. (a) Blue crosses denote events in ERA5 (OBS) and grey boxplots represent the 75 events per year in the raw unadjusted SEAS5 hindcast (UNSEEN). Boxplots show the median, inter-quartile range, 1.5× interquartile range, and outliers (data outside the 1.5× interquartile range). (b) As in Figure 5, box and violin plots show the correlation between ensemble members for each of the lead times within the hindcast ensemble. (c,d) As in Figure 6 but for MAM Siberian temperature, showing the probability density (c) and extreme value distribution (d) for each lead time. Grey shading shows 95% confidence bounds. (e–h) As in Figure 7 but for MAM Siberian temperature and using ERA5 statistics derived for the period 1981–2016. Histograms show the distribution characteristics for SEAS5, dashed lines indicate 95% intervals, and blue lines represent ERA5 statistics.

assessed (Solution F3) but are beyond the scope of this paper. An evaluation of feedbacks involving soil moisture or snow cover that contributed to the 2020 anomaly (Ciavarella et al., 2021; Overland & Wang, 2021) merits further research.

The resulting bias-adjusted UNSEEN ensemble is inherently conservative because of the low variability, yet captures the 2020 event, along with five thawing events (MAM mean temperature >0°C), with a near possibility as early as the 1990s (Figure 9). In comparison, there had

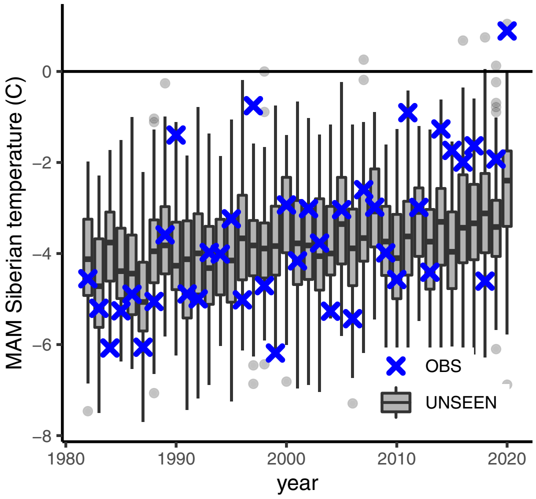

FIGURE 9 Thawing events over Siberia within the bias-adjusted UNSEEN ensemble compared with ERA5. As in Figure 8a, but after applying a mean bias adjustment to the UNSEEN ensemble. The horizontal line shows the threshold for thawing events (temperature >0°C).

been no observed thawing events within the reanalysis before 2020.

3.2 | Temperature anomalies during peak California wildfire activity

The technical steps to reproduce this example are available at https://unseen-open.readthedocs.io/en/latest/Notebooks/examples/California_Fires.html.

Wildfire activity cannot be studied from meteorological variables alone, because wildfire activity depends not only on weather conditions but also on fuel stock, ignition agents, and management (Flannigan et al., 2013). For example, weather conditions may be very dry, but without fuel or ignition source(s), wildfire activity is unlikely. However, weather-driven fire danger conditions can be studied from meteorological variables (e.g., Vitolo et al., 2020). For example, trends in temperature and precipitation are associated with rising likelihood of wildfire conditions across California (Goss et al., 2020). In 2020, the wildfire season peaked in August, coinciding with record high-temperature anomalies (Figure 2b,d). Here, we demonstrate the applicability and potential of SEAS5 in estimating the likelihood and trend of such a temperature anomaly. We selected a contiguous, land-only region where August temperature anomalies were more than twice the climatological (1979–2010) standard deviation based on ERA5 over the domain 100–125°W, 20–45°N (Figure 2b).

The ERA5 time series shows a strong increase in August temperatures over 1981–2020 for this domain, which is also present in SEAS5 (Figure 10a). We find low ensemble member dependence in the UNSEEN ensemble for all lead times (Figure 10b). We also find that the model is not stable, especially for lead time 6 the model has a cold bias (Figures 6 and 10c,d). Lastly, we find that SEAS5 overestimates mean August temperatures when compared with ERA5, but that the standard deviation, skewness, and kurtosis are not significantly different at a 95% confidence level (Figure 10e–h).

Following these tests, we remove lead time 6 from the ensemble to solve the instability for lead time 6 (Solution S1 in Table 2) and apply a mean bias adjustment to solve the warm bias (Solution F1), leaving 100 members in the pooled data. Removing problematic lead times is not an option to solve the dependence (Solution I1 in Table 2) because the issue persists across all lead times. Further assessment of the independence between ensemble members was not deemed necessary in this case because of the low median correlation values (Figure 10b).

We then use extreme value statistics and find that the trend in 2-year temperature extremes, which can be detected well within short observational records, is similar between UNSEEN and reanalysis (Figure 11a). Both reanalysis and UNSEEN suggest a strong increase in the magnitude of 100-year temperature extremes (Figure 11b), but the statistical uncertainty is much larger within the 40-year reanalysis record (blue envelope in Figure 11b) than within large sample size of UNSEEN (100 × 40 years, orange envelope in Figure 11b). When we compare the GEV distributions with the 'year' covariate for 1981 as opposed to 2020, we find that the distribution of temperature for 1981 does not reach the magnitude of the 2020 event, whereas the distribution for the year 2020 does capture the event for both reanalysis and UNSEEN (Figure 11c). This result suggests that the temperature anomaly observed in 2020 could not have occurred a few decades ago and that it was still unlikely to occur in the present climate (i.e., the distribution for the year 2020), with a return period of more than 100 years, that is, <1% chance of occurrence. This trend is consistent with record-breaking or 'record-shattering' temperatures being expected to occur more frequently in a rapidly warming climate (Coumou et al., 2013; Fischer et al., 2021; Power & Delage, 2019).

3.3 | UK extreme precipitation

The technical steps to reproduce this example are available at https://unseen-open.readthedocs.io/en/latest/Notebooks/examples/UK_Precipitation.html.

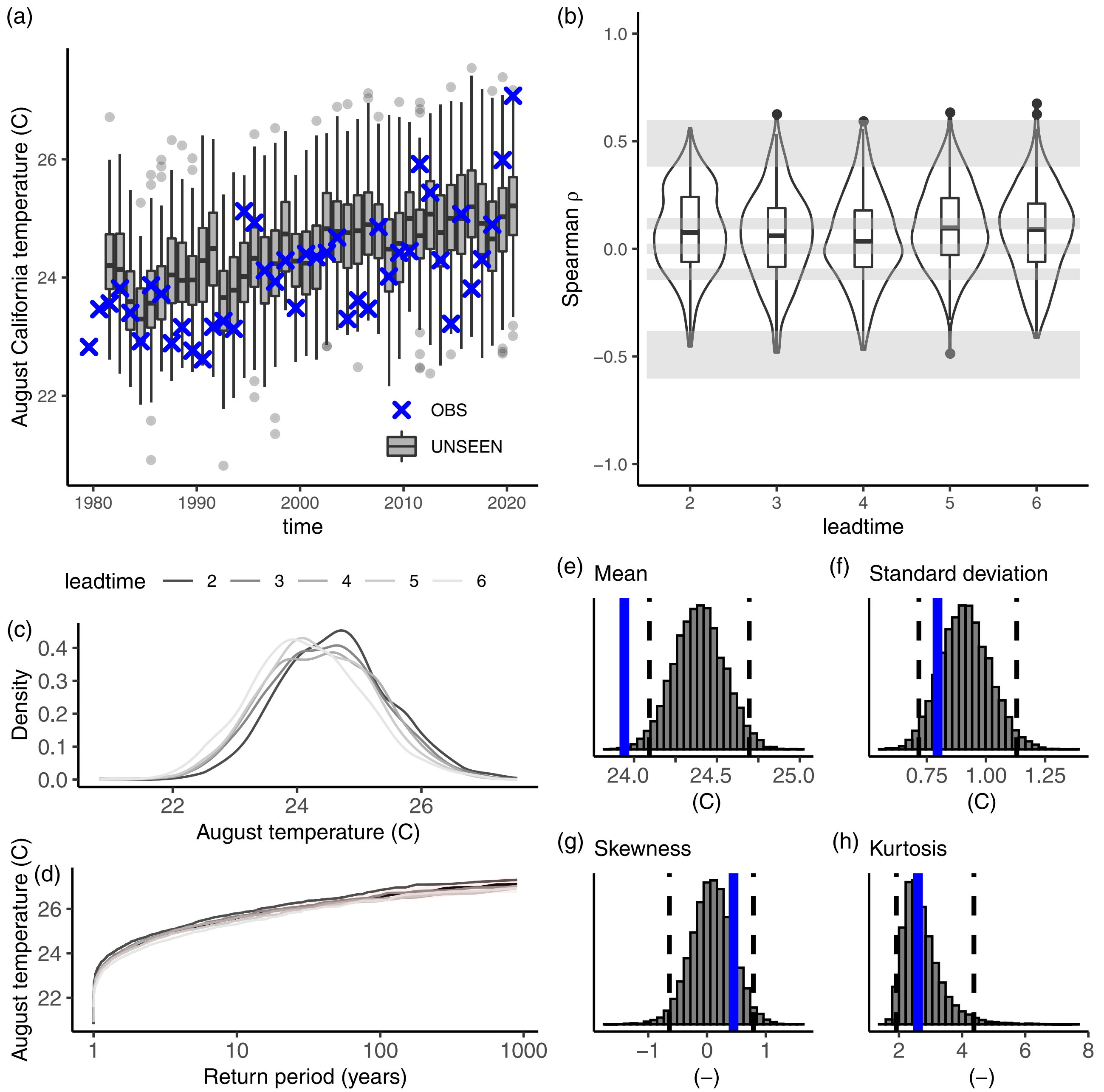

FIGURE 10 As in Figure 8 but for (a) time series (b) independence (c–d) stability and (e–h) fidelity of the ensemble for August temperature anomalies during peak California wildfire activity. (a) Blue crosses denote events in ERA5 (OBS) and grey boxplots represent the 125 events per year in the raw unadjusted SEAS5 hindcast (UNSEEN). (b) Box and violin plots show the correlation between ensemble members for each of the lead times for California-Mexico August temperatures within the hindcast ensemble. (c,d) The probability density (c) and extreme value distribution (d) of California-Mexico August temperatures for each lead time. (e–h) California-Mexico August temperature distribution characteristics for SEAS5 (histograms, including dashed lines indicating 95% intervals) and for ERA5 derived for the period 1981–2016 (blue lines).

Three storms hit the United Kingdom in February 2020, breaking the UK-average monthly precipitation record according to the Met Office (2020). Hence, we select country-averaged February precipitation for the UK case study. In this case, we employ the EOBS version 20.0e observational dataset as reference (Cornes et al., 2018) because precipitation observations (UK Met Office, 2020) suggest the reanalysis values may have underestimated the event. We upscale this dataset to the resolution of SEAS5 using bilinear interpolation and take the same UK spatial average as for SEAS5.

The UK February precipitation time series shows that the 2020 event was not the highest on record within the EOBS dataset (Figure 12a), while it was the highest

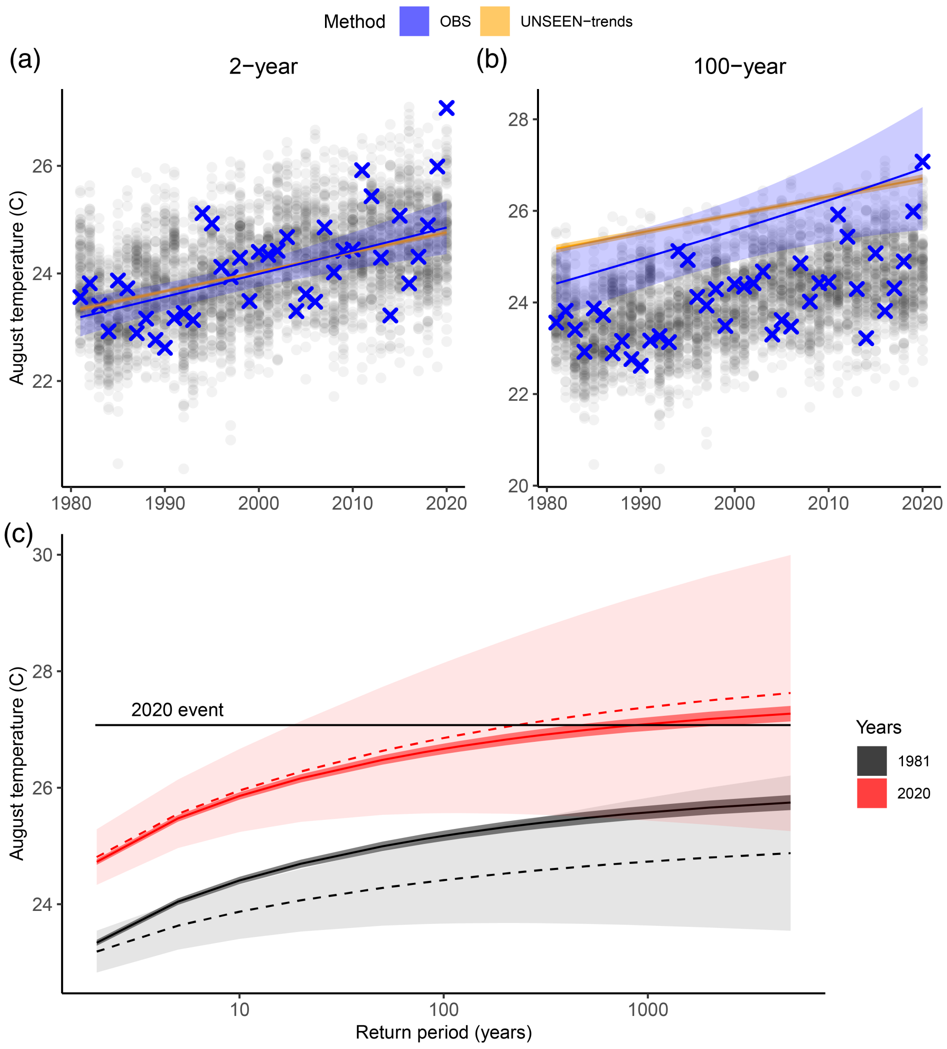

FIGURE 11 Trends in extreme temperatures estimated by UNSEEN and reanalysis. (a,b) The temporal change in 2-year (a) and 100-year (b) August temperature extremes. Blue crosses indicate events in ERA5 (OBS). Grey circles indicate SEAS5 ensemble members (UNSEEN). (c) The GEV distribution for the covariates 1981 and 2020. Distributions based on UNSEEN are indicated by solid lines with uncertainty estimates in darker shading. The distributions based on ERA5 data are indicated by dashed lines and the uncertainty range by lighter shading. The magnitude of the 2020 event is indicated with a black horizontal line. In all plots and for both OBS and UNSEEN, statistical uncertainty is estimated as 95% confidence intervals based on the normal approximation.

within the HadUK-Grid dataset (Davies et al., 2021; Hollis et al., 2019). The discrepancy likely arises from the number of observation stations being incorporated, with the local HadUK-Grid dataset containing more rain gauges. Note that later versions of EOBS may have incorporated more observation stations for the year 2020, but these versions were not available at the time of analysis.

We find that SEAS5 UK February precipitation ensemble members are independent (Figure 12b) and stable (Figure 12c,d). However, there is too little variability within SEAS5 when compared with EOBS (Figure 12e–h), raising concerns about model fidelity. Independent UNSEEN analysis of February 2020 UK precipitation using the Met Office decadal prediction system and observations also found a lack of fidelity, with observed variability outside the range of that simulated (Kay 2021, personal communication). A mean bias adjustment (Solution F1 in Table 2) does not help in this case, because it will not sufficiently adjust the standard deviation. The result will likely not be sensitive to the evaluation test (Solution F2), such as a rank histogram or reliability diagram, given that the lack of variability is also evident in the time series (Figure 12a). Further evaluating the drivers (Solution F3) and comparing the results to other datasets (Section 4.2) would be recommended, as the realistic simulation of large-scale winter precipitation variability over the United Kingdom may be hampered by the SEAS5's resolution. For example, Thompson et al. (2017) also found that DePreSys3 does not simulate the orographic enhancement over the Scottish highlands. Flat regions are better simulated, such as southern England.

We do not take this case study further, as the generated ensemble of UK-average February precipitation did not pass the fidelity test and could not be resolved. Note, however, that UNSEEN can successfully be applied to monthly winter precipitation over Southeast England (Thompson et al., 2017). Furthermore, for a detailed analysis of the dynamics of the wet Winter 2019/2020, including the attribution of the record-breaking February 2020 precipitation to climate change, see Davies et al. (2021) and Hardiman et al. (2020).

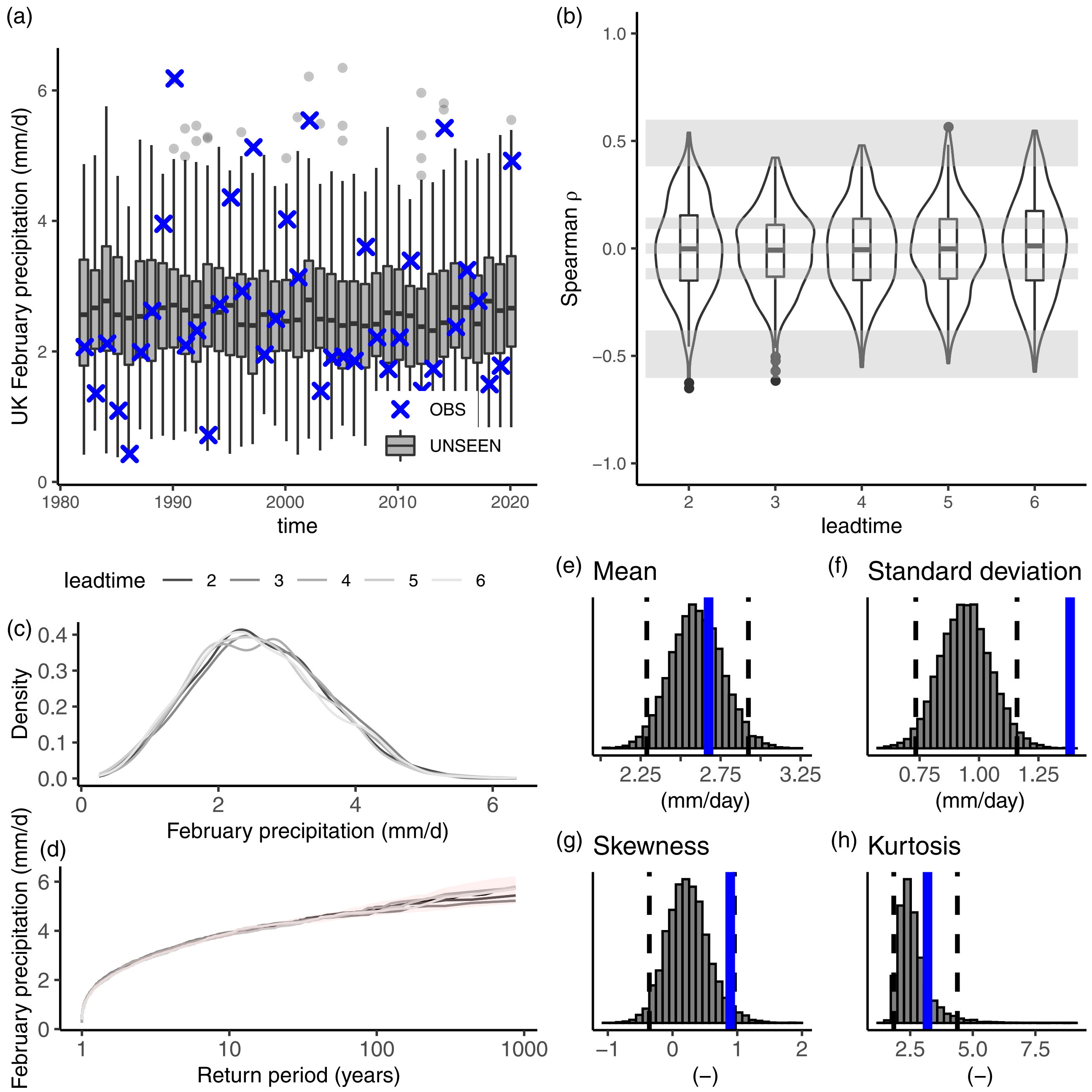

FIGURE 12 As in Figure 8 but for the (a) time series (b) independence (c–d) stability and (e–h) fidelity of the ensemble for February UK precipitation. (a) Blue crosses denote events in EOBS (OBS) and grey boxplots represent the 125 events per year in the raw unadjusted SEAS5 hindcast (UNSEEN). (b) Box and violin plots show the correlation between ensemble members for each of the lead times within the hindcast ensemble for UK February precipitation. (c,d) The probability density (c) and extreme value distributions (d) of UK February precipitation for each lead time. (e–h) The distribution characteristics of UK February precipitation for SEAS5 (histograms, including dashed lines indicating 95% intervals) and for EOBS derived for the period 1981–2016 (blue lines).

4 | DISCUSSION

This paper sets out a protocol for generating a credible, large ensemble for event definitions specified by a user (step 1). The protocol guides the user through the selection of an appropriate prediction system (step 2), the retrieval (step 3), pre-processing (step 4), and evaluation (step 5) of the data, and how to resolve detected issues (step 6). Finally, the protocol describes the use of statistics to gain insight into low-likelihood events (step 7). A technical UNSEEN-open workflow for steps 3–5 is presented using ECMWF seasonal prediction system SEAS5 (Johnson et al., 2019). In this section, we reflect on the challenges and sensitivities to be mindful of when applying UNSEEN.

4.1 | Sensitivity to the choice of event

Here, the event definition is informed by the meteorological rarity of observed low-likelihood high-impact events. This approach does not necessarily best match target impacts. For example, local fire weather indices over California would be more informative than the temperature anomaly over a larger domain that we assessed, also noting that high impacts do not necessarily need to result from rare meteorological events (Van der Wiel et al., 2020). Furthermore, the outcome of UNSEEN is sensitive to the selection of the spatial–temporal scale of the event, as is well documented for event attribution studies (e.g., Angélil et al., 2014; Leach et al., 2020; Uhe et al., 2016). Finally, an event definition based on observed events is not available for unprecedented events. The outcomes of our case studies are therefore (1) not optimized to study impacts, (2) only hold for the specified spatial and temporal scales, and (3) can only be applied to events in hindsight. We believe these assumptions are justified to demonstrate how UNSEEN can be used to gain insight into low-likelihood events, for example, whether past events could have been detected with UNSEEN. When aiming to inform decision-making based on UNSEEN, sensitivities must be assessed and the event definition should best match target impacts, for example, through local expert knowledge of the domain and timescale that are most vulnerable to the target weather event. Practical factors may weigh in as well, such as our workflow considerations to use the average of, or daily max/min values within, monthly or seasonal blocks (Section 4.3). The protocol can help to understand the applicability of UNSEEN for the chosen event definition and can be used to test the sensitivity of the outcomes to various spatial–temporal scales.

4.2 | Model adequacy and availability

Ideally, the adequacy of the models should inform the selection of an appropriate prediction system (step 2 and Figure 3). However, the availability of model simulations may be another important consideration. The UNSEEN-open workflow was developed to generate and evaluate an UNSEEN ensemble from open access Copernicus SEAS5 simulations. SEAS5 has been used in other UNSEEN studies because it provides a stable, homogeneous, global, high-resolution, large ensemble of weather variables (Hillier & Dixon, 2020; Kelder et al., 2020).

The outcomes from UNSEEN based on SEAS5 initialized ensemble predictions are not completely independent from outcomes based on reanalysis data. For example, the California-Mexico August temperature trends based on SEAS5 are conditioned on ERA-Interim (1981–2016) and ECMWF operational analysis (ERA5 for 2017–2020, see ‘Step 3: Retrieve’). Trends in low-likelihood weather events such as the August 2020 California-Mexico temperatures are hard to constrain from reanalysis data alone, and the large sample size from UNSEEN (SEAS5) can help. The prediction timescale used here is at least 1 month, so the ensemble members are less constrained to observation stations than reanalysis. UNSEEN is, therefore, less reliable than reanalysis but represents many weather realizations that may face less of an issue with assimilation inhomogeneity over time. The evaluation of SEAS5 to ERA5 can confirm if the large sample size from UNSEEN (SEAS5) is as reliable as ERA5, but both may equally face model errors.

4.3 | UNSEEN-open workflow considerations

In this workflow (steps 3–5 in Figure 1), SEAS5 monthly statistics are retrieved locally from the Copernicus Climate Data Store (Buontempo et al., 2020; Thepaut et al., 2018), which are openly available, freely accessible, and can be retrieved without an intermediary. The case studies in this paper include monthly average statistics from CDS, but the workflow is sufficiently flexible to draw on monthly minimum or maximum data. For compound or multi-day events, daily data can be retrieved and processed to obtain the required metric.

There are two points of attention for users to consider when using SEAS5: (1) the ensemble size depends on the selected block length and (2) the ensemble represents the conditions of the most recent decades only. Forecasts run for 6 months and, therefore, an ensemble size of 125 members can be created for monthly blocks, 75 members for seasonal blocks, and events longer than 5 consecutive months are not possible without stitching forecasts (Section 2, step 3). When longer time periods are required to evaluate internal climate variability, century-long seasonal hindcasts with a similar set-up to SEAS5 but at lower resolution, such as the Coupled Seasonal Forecasts of the 20th Century (CSF-20C, Weisheimer et al., 2021), or the Atmospheric Seasonal Forecasts of the 20th Century (ASF-20C, Weisheimer et al., 2017) may be useful. The workflow is adjustable for other prediction systems, including medium/extended range, seasonal and decadal (Table 1), but, here, retrieval was optimized for Copernicus SEAS5.

At present, hindcast datasets are available for download and need to be pre-processed, which can be a time-consuming process. An open workflow as presented in this paper would benefit from having large volumes of data such as the SEAS5 hindcast accessible on-demand via a cloud-based service (Pappenberger et al., 2021; Wagemann et al., 2018). An example of a cloud service for the meteorological and climate community, which in the future may be incorporated in the UNSEEN-open workflow to obviate retrieval of data, is the European Weather Cloud (Pappenberger & Palkovic, 2020).

4.4 | Sensitivity to evaluation metrics and solutions

The applicability of UNSEEN is determined by multiple, interrelated factors. Ensemble member independence, model stability, and model fidelity depend on the type of event being studied (variable of interest, spatial and temporal extent, and geographical location), as well as on the prediction system applied. The prediction timescale furthermore influences the independence and stability, as longer simulations are more independent but have a higher chance of drifting away from climatology. Different systems, and the way they have been downscaled, initialized, and coupled, may yield different biases. Therefore, it is recommended that our protocol is used to explore the applicability for the selected event definition and prediction system(s).

All three case studies of extreme weather events in 2020 express challenges with respect to the credibility of UNSEEN (independence, stability, and fidelity, see Box 2). We note that a wide range of evaluation metrics exist, especially to evaluate the model fidelity (described under Solution F2 and F3), and that the sensitivity of the identified challenges to evaluation metrics could be further assessed. For two out of three case studies, we find that solutions could be applied to deem the UNSEEN ensemble credible for further analysis. We note that the outcome of UNSEEN is sensitive to the judgement of appropriate solutions. For example, here, we identified weakly dependent ensemble members for the Siberian heatwave and for the California-Mexico temperature anomalies. There is a trade-off between discarding useful information compared with keeping dependent members. Here, we chose to keep all members as the ensemble member dependence was low. If we would have removed certain lead times, the results would be different.

The outcome of UNSEEN is furthermore sensitive to the type of bias adjustment. The UNSEEN ensemble may sample plausible extreme events that never occurred, and bias adjustment techniques may constrain the ensemble to observed extremes—thereby removing information about unseen events. Furthermore, observations are not the ‘truth’ under internal variability, resolution mismatch, and other sources of error (Casanueva et al., 2020; Wilby et al., 2017). Attention is therefore needed to evaluate which statistical properties of the extremes are being constrained to observations.

4.5 | Sensitivity to the analysis and framing

The outcome of UNSEEN is sensitive to the statistical analysis being applied. In particular, the estimation method, time window, and initialization method of UNSEEN (SEAS5) are factors that may influence estimated likelihoods and detected trends. Here, we allow the location and scale parameters to vary linearly with time (step 7). Other regression methods, other covariates than time, other reference data, and other prediction systems allowing longer time periods could be explored, but such analyses are beyond the scope of this paper.

Furthermore, correct framing of the result of an UNSEEN study is crucial. For example, for the Siberian heatwave, we did not want to apply extreme value theory or attach likelihood estimates to the event, because the UNSEEN ensemble was conservative with a low standard deviation. Nonetheless, we are confident in saying that the 2020 March–May Siberian heatwave was predicted by one of the ensemble members prior to the event happening, along with other simulations of thawing events that had not yet been observed. The sensitivity of likelihood statistics to portfolio risks should also be considered: whereas the chance of a single high-impact weather event to occur might be very low, the chance of any type of high-impact weather event to occur anywhere in the world is substantially larger. For example, Thompson et al. (2017) showed a 7% chance for unprecedented winter monthly precipitation to occur in a given year for southern England, but 34% when also including the chance of unseen events over the Midlands, East Anglia, or northeast England.

4.6 | Scope for multi-model multi-method approaches

Most studies evaluating unprecedented extreme events have used single models to assess their magnitude and frequency, but such analyses are sensitive to model structures (e.g., van Kempen et al., 2021; Wilcke et al., 2020). Multi-model approaches have therefore been used in weather predictions, climate projections, and event attribution studies (Palmer et al., 2005; Philip et al., 2020; Tebaldi & Knutti, 2007). Jain et al. (2020) were the first to apply a multi-model ensemble in an UNSEEN approach using the Climate-System Historical Forecast Project (Tompkins et al., 2017) to study extreme summer rainfall over India. Future work may investigate the extension of the UNSEEN-open workflow to include all seasonal prediction systems available in the CDS.

Furthermore, UNSEEN is one tool within many available tools to study plausible low-likelihood high-impact weather events. There is scope to assess the mutual benefit of various approaches, as is common for event attribution (Philip et al., 2020; van Oldenborgh et al., 2021), including ensembles of opportunity (e.g., King et al., 2017; Lewis et al., 2017), single-model initial-condition large ensembles (e.g., Suarez-Gutierrez et al., 2020a, 2020b), ensemble reinitialization methods (e.g., Gessner et al., 2021), targeted large ensemble experiments (e.g., Guillod et al., 2017; Mitchell et al., 2017), pooling of observations (e.g., Berghuijs et al., 2017; Robinson et al., 2021), long archives (Hawkins et al., 2019; Murphy et al., 2020), paleoclimatic records (Yan et al., 2020), and statistical weather generators (Brunner & Gilleland, 2020; Wilks & Wilby, 1999; Yiou, 2014). Seasonal and decadal prediction systems may, furthermore, contribute additional lines of evidence to event attribution statement if their trend estimates can be extrapolated to represent pre-industrial climates.

5 | CONCLUSION

Hindcast ensembles from weather predictions have considerable potential for advancing understanding of low-likelihood high-impact weather events. Estimates of rare extreme events or compound extremes can be improved through the large number of weather events that can be generated from these ensembles (UNSEEN, van den Brink et al., 2004, Thompson et al., 2017). To improve uptake and ensure rigour of these methods, we provide a protocol and open workflow to apply UNSEEN for gaining insights about low-likelihood high-impact events.

Demonstrating our protocol and open workflow using SEAS5, we show that a stress-test of March–May thawing events over Siberia would have shown their plausibility within the UNSEEN ensemble before the event happened, for which the far-reaching impacts on permafrost peatlands were already widely known (e.g., Swindles et al., 2015). Assessing UK February monthly precipitation revealed an issue with the variability of SEAS5 for this event definition, illustrating how the protocol may help understand the limitations of UNSEEN and diagnose the lack of simulated precipitation variability in the underlying forecasting system. In the case of August 2020 temperatures during peak California wildfire activity, anomalies exceeded previous records by a considerable margin (Figure 2c,d). Such anomalous events can have large socio-economic consequences, especially when climate risk perception is driven by past experiences (Aerts et al., 2018; Weber, 2006). The UNSEEN approach reveals a strong trend in temperature extremes over the last 40 years, which has increased the likelihood of events like the August 2020 temperature anomalies in the present climate. UNSEEN shows the plausibility of such a record-shattering event in the present climate, but not in the past climate. This case study shows how UNSEEN may help to understand what kind of unseen weather events could now occur in the present climate, and thus in the near future.

Based on these case studies, we conclude that UNSEEN can provide new insights into low-likelihood high-impact events, but that there are several challenges and sensitivities of which to be mindful. It is, therefore, key to be transparent about all decisions that are made throughout the analysis, given the many sensitivities that can arise from these decisions. Our protocol and open workflow assist users to identify challenges and sensitivities, and can help gain credible insights for target high-impact weather events. The results warrant further research and application of UNSEEN at the science-policy interface, to improve our preparedness to low-likelihood high-impact weather events.

ACKNOWLEDGEMENTS

T. Kelder acknowledges support from Loughborough University and the NERC CENTA Doctoral Training Partnership, acknowledges computation facilities provided by the University of Oxford, and acknowledges support from the European Union Copernicus Programme within the *ECMWF Summer of Weather Code* 2020. N. Dunstone acknowledges support from the Met Office Hadley Centre Climate Programme funded by BEIS. We acknowledge the EU-FP6 project UERRA (https://www.uerra.eu), data providers in the ECA&D project (https://www.ecad.eu), and the Copernicus Climate Change Service for the SEAS5, ERA5, and EOBS datasets.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

AUTHOR CONTRIBUTIONS

T. Kelder: Conceptualization (lead); data curation (lead); formal analysis (lead); investigation (lead); methodology (lead); software (lead); visualization (lead); writing – original draft (lead); writing – review and editing (equal). **T. I. Marjoribanks:** Conceptualization (supporting); formal analysis

(supporting); methodology (supporting); supervision (equal); writing – review and editing (equal). **L. J. Slater:** Conceptualization (supporting); formal analysis (supporting); methodology (supporting); supervision (equal); writing – review and editing (equal). **C. Prudhomme:** Conceptualization (supporting); formal analysis (supporting); methodology (supporting); supervision (equal); writing – review and editing (equal). **R. L. Wilby:** Conceptualization (equal); formal analysis (supporting); methodology (supporting); supervision (equal); writing – review and editing (equal). **J. Wagemann:** Software (supporting); supervision (equal); writing – review and editing (supporting). **N. Dunstone:** Formal analysis (supporting); writing – review and editing (supporting).

DATA AVAILABILITY STATEMENT

UNSEEN-open at present uses Copernicus SEAS5, ERA5, and EOBS datasets, which are openly available through the Copernicus Data Store (CDS, https://cds.climate.copernicus.eu/). The pre-processed data for the example case studies are available on the GitHub repository. The European Commission and ECMWF are not responsible for any use of UNSEEN-open. Documentation: https://unseen-open.readthedocs.io. Repository: https://github.com/esowc/UNSEEN-open.

ORCID

T. Kelder https://orcid.org/0000-0001-9802-9837
T. I. Marjoribanks https://orcid.org/0000-0003-0116-2952
L. J. Slater https://orcid.org/0000-0001-9416-488X
C. Prudhomme https://orcid.org/0000-0003-1722-2497
J. Wagemann https://orcid.org/0000-0002-3075-2559
N. Dunstone https://orcid.org/0000-0001-6859-6814

REFERENCES

Aerts, J.C.J.H., Botzen, W.J., Clarke, K.C., Cutter, S.L., Hall, J.W., Merz, B. et al. (2018) Integrating human behaviour dynamics into flood disaster risk assessment. *Nature Climate Change*, 8, 193–199 Available from: www.nature.com/natureclimatechange

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 Climate Extremes*, 11, 4–16.

Allen, M. (2003) Liability for climate change. *Nature*, 421, 891–892.

Alley, R.B., Emanuel, K.A. & Zhang, F. (2019) Advances in weather prediction. *Science*, 363, 342–344. https://doi.org/10.1126/science.aav7274

Angélil, O., Stone, D.A., Tadross, M., Tummon, F., Wehner, M. & Knutti, R. (2014) Attribution of extreme weather to anthropogenic greenhouse gas emissions: sensitivity to spatial and temporal scales. *Geophysical Research Letters*, 41, 2150–2155. https://doi.org/10.1002/2014GL059234

Bauer, P., Thorpe, A. & Brunet, G. (2015) The quiet revolution of numerical weather prediction. *Nature*, 525(7567), 47–55.

Bellprat, O., Guemas, V., Doblas-Reyes, F. & Donat, M.G. (2019) Towards reliable extreme weather and climate event attribution. *Nature Communications*, 10, 1732 Available from: http://www.nature.com/articles/s41467-019-09729-2

Berghuijs, W.R., Aalbers, E.E., Larsen, J.R., Trancoso, R. & Woods, R.A. (2017) Recent changes in extreme floods across multiple continents. *Environmental Research Letters*, 12, 114035.

Bevacqua, E., De, M.C., Manning, C., Couasnon, A., Ribeiro, A. F.S., Ramos, A.M. et al. (2021) Guidelines for studying diverse types of compound weather and climate events. *Earth's Future*, 9(11), e2021EF002340. https://doi.org/10.1029/2021EF002340

Bhatia, U. & Ganguly, A.R. (2019) Precipitation extremes and depth-duration-frequency under internal climate variability. *Scientific Reports*, 9, 1–9. https://doi.org/10.1038/s41598-019-45673-3

Breivik, Ø., Aarnes, O.J., Abadalla, S., Bidlot, J.-R. & Janssen, P. (2014) Wind and wave extremes over the world oceans from very large ensembles. *Geophysical Research Letters*, 41, 5122–5131.

Breivik, Ø., Aarnes, O.J., Bidlot, J.-R., Carrasco, A. & Saetra, Ø. (2013) Wave extremes in the Northeast Atlantic from ensemble forecasts. *Journal of Climate*, 26, 7525–7540.

Brunner, M.I. & Gilleland, E. (2020) Stochastic simulation of streamflow and spatial extremes: a continuous, wavelet-based approach. *Hydrology and Earth System Sciences*, 24, 3967–3982 Available from: https://hess.copernicus.org/articles/24/3967/2020/

Brunner, M.I. & Slater, L.J. (2022) Extreme floods in Europe: going beyond observations using reforecast ensemble pooling. *Hydrology and Earth System Sciences*, 26, 469–482 Available from: https://hess.copernicus.org/articles/26/469/2022/

Buontempo, C., Hutjes, R., Beavis, P., Berckmans, J., Cagnazzo, C., Vamborg, F. et al. (2020) Fostering the development of climate services through Copernicus Climate Change Service (C3S) for agriculture applications. *Weather and Climate Extremes*, 27, 100226 Available from: https://linkinghub.elsevier.com/retrieve/pii/S2212094719300994

Cannon, A.J., Piani, C. & Sippel, S. (2020) *Bias correction of climate model output for impact models. Climate Extremes and Their Implications for Impact and Risk Assessment.* Amsterdam: Elsevier, pp. 77–104.

Casanueva, A., Herrera, S., Iturbide, M., Lange, S., Jury, M., Dosio, A. et al. (2020) Testing bias adjustment methods for regional climate change applications under observational uncertainty and resolution mismatch. *Atmospheric Science Letters*, 21, e978. https://doi.org/10.1002/asl.978

Ciavarella, A., Cotterill, D., Stott, P., Kew, S., Philip, S., van Oldenborgh, G.J. et al. (2021) Prolonged Siberian heat of 2020 almost impossible without human influence. *Climatic Change*, 166, 9. https://doi.org/10.1007/s10584-021-03052-w

Coles, S. (2001) *An introduction to statistical modeling of extreme values*, Vol. 208. London: Springer London. https://doi.org/10.1007/978-1-4471-3675-0

Copernicus EMS (2020) UK and Ireland floods, February 2020. Copernicus EMS - European Flood Awareness System. Available at: https://www.efas.eu/en/news/uk-and-ireland-floods-february-2020.

Cornes, R.C., van der Schrier, G., van den Besselaar, E.J.M. & Jones, p.D. (2018) An ensemble version of the E-OBS temperature and precipitation data sets. *Journal of Geophysical Research – Atmospheres*, 123, 9391–9409. https://doi.org/10.1029/2017JD028200

Coumou, D., Robinson, A. & Rahmstorf, S. (2013) Global increase in record-breaking monthly-mean temperatures. *Climatic Change*, 118, 771–782. https://doi.org/10.1007/s10584-012-0668-1

Courty, L.G., Wilby, R.L., Hillier, J.K. & Slater, L.J. (2019) Intensity-duration-frequency curves at the global scale. *Environmental Research Letters*, 14, 084045. https://doi.org/10.1088/1748-9326/ab370a

Covey, C., Gleckler, p.J., Phillips, T.J. & Bader, D.C. (2006) Secular trends and climate drift in coupled ocean-atmosphere general circulation models. *Journal of Geophysical Research*, 111, D03107. https://doi.org/10.1029/2005JD006009

Cowan, T., Undorf, S., Hegerl, G.C., Harrington, L.J. & Otto, F.E.L. (2020) Present-day greenhouse gases could cause more frequent and longer dust bowl heatwaves. *Nature Climate Change*, 10, 505–510. https://doi.org/10.1038/s41558-020-0771-7

Davies, p.A., McCarthy, M., Christidis, N., Dunstone, N., Fereday, D., Kendon, M. et al. (2021) The wet and stormy UK winter of 2019/2020. *Weather*, 76, 396–402. https://doi.org/10.1002/wea.3955

De Luca, P., Messori, G., Pons, F.M.E. & Faranda, D. (2020) Dynamical systems theory sheds new light on compound climate extremes in Europe and Eastern North America. *Quarterly Journal of the Royal Meteorological Society*, 146(729), 1636–1650 Available from: https://www.cpc.ncep.noaa

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S. et al. (2011) The ERA-interim reanalysis: configuration and performance of the data assimilation system *Q. Journal of the Royal Meteorological Society*, 137, 553–597.

Deser, C., Lehner, F., Rodgers, K.B., Ault, T., Delworth, T.L., DiNezio, p.N. et al. (2020) Insights from earth system model initial-condition large ensembles and future prospect. *Nature Climate Change*, 10, 277–286 Available from: http://www.nature.com/articles/s41558-020-0731-2

Diffenbaugh, N.S., Singh, D., Mankin, J.S., Horton, D.E., Swain, D. L., Touma, D. et al. (2017) Quantifying the influence of global warming on unprecedented extreme climate events. *Proceedings of the National Academy of Sciences of the United States of America*, 114, 4881–4886 Available from: https://www.pnas.org/content/114/19/4881

Dunstone, N., Smith, D., Scaife, A., Hermanson, L., Eade, R., Robinson, N. et al. (2016) Skilful predictions of the winter North Atlantic oscillation one year ahead. *Nature Geoscience*, 9, 809–814 Available from: https://www.nature.com/articles/ngeo2824

ECMWF 2018 Climate reanalysis. ECMWF. Available at: https://www.ecmwf.int/en/research/climate-reanalysis.

Eyring, V., Cox, p.M., Flato, G.M., Gleckler, p.J., Abramowitz, G., Caldwell, P. et al. (2019) Taking climate model evaluation to the next level. *Nature Climate Change*, 9, 102–110 Available from: https://www.nature.com/articles/s41558-018-0355-y

Eyring, V., Righi, M., Lauer, A., Evaldsson, M., Wenzel, S., Jones, C. et al. (2016) ESMValTool (v1.0) – a community diagnostic and performance metrics tool for routine evaluation of earth system models in CMIP. *Geoscientific Model Development*, 9, 1747–1802 Available from: https://pure.mpg.de/pubman/faces/ViewItemOverviewPage.jsp?itemId=item_2283051

Faranda, D. (2020) An attempt to explain recent changes in European snowfall extremes. *Weather and Climate Dynamics*, 1, 445–458 Available from: https://wcd.copernicus.org/articles/1/445/2020/

Faranda, D., Messori, G., Carmen Alvarez-Castro, M. & Yiou, P. (2017) Dynamical properties and extremes of northern hemisphere climate fields over the past 60 years. *Nonlinear Processes in Geophysics*, 24, 713–725.

Fischer, E.M., Sippel, S. & Knutti, R. (2021) Increasing probability of record-shattering climate extremes. *Nature Climatic Change*, 118(11), 689–695 Available from: https://www.nature.com/articles/s41558-021-01092-9

Flannigan, M., Cantin, A.S., De Groot, W.J., Wotton, M., Newbery, A. & Gowman, L.M. (2013) Global wildland fire season severity in the 21st century. *Forest Ecology and Management*, 294, 54–61.

Geirinhas, J.L., Russo, A., Libonati, R., Sousa, p.M., Miralles, D. G. & Trigo, R.M. (2021) Recent increasing frequency of compound summer drought and heatwaves in Southeast Brazil. *Environmental Research Letters*, 16, 34036. https://doi.org/10.1088/1748-9326/abe0eb

Gessner, C., Fischer, E.M., Beyerle, U. & Knutti, R. (2021) Very rare heat extremes: quantifying and understanding using ensemble re-initialization. *Journal of Climate*, 34(16), 6619–6634 Available from: https://journals.ametsoc.org/view/journals/clim/aop/JCLI-D-20-0916.1/JCLI-D-20-0916.1.xml

Gilleland, E. & Katz, R.W. (2016) extRemes 2.0: an extreme value analysis package in R. *Journal of Statistical Software*, 72, 1–39.

Goss, M., Swain, D.L., Abatzoglou, J.T., Sarhadi, A., Kolden, C.A., Williams, A.P. et al. (2020) Climate change is increasing the likelihood of extreme autumn wildfire conditions across California. *Environmental Research Letters*, 15, 094016. https://doi.org/10.1088/1748-9326/ab83a7

Grazzini, F., Fragkoulidis, G., Pavan, V. & Antolini, G. (2020) The 1994 Piedmont flood: an archetype of extreme precipitation events in Northern Italy. *Bulletin of Atmospheric Science and Technology*, 1, 283–295. https://doi.org/10.1007/s42865-020-00018-1

Guillod, B.P., Jones, R.G., Bowery, A., Haustein, K., Massey, N.R., Mitchell, D.M. et al. (2017) Weather@home 2: validation of an improved global-regional climate modelling system. *Geoscientific Model Development*, 10, 1849–1872.

Guillod, B.P., Jones, R.G., Dadson, S.J., Coxon, G., Bussi, G., Freer, J. et al. (2018) A large set of potential past, present and future hydro-meteorological time series for the UK. *Hydrology and Earth System Sciences*, 22, 611–634.

Hall, J.W., Borgomeo, E., Bruce, A., Di Mauro, M. & Mortazavi-Naeini, M. (2019) Resilience of water resource systems: lessons from England. *Water Security*, 8, 100052.

Hall, J.W., Mortazavi-Naeini, M., Borgomeo, E., Baker, B., Gavin, H., Gough, M. et al. (2020) Risk-based water resources planning in practice: a blueprint for the water industry in England. *Water and Environment Journal*, 34, 441–454. https://doi.org/10.1111/wej.12479

Hansen, J., Ruedy, R., Sato, M. & Lo, K. (2010) Global surface temperature change. *Reviews of Geophysics*, 48, 4004 Available from: https://www.giss.nasa.gov

Hardiman, S.C., Dunstone, N.J., Scaife, A.A., Smith, D.M., Knight, J.R., Davies, P. et al. (2020) Predictability of European winter 2019/20: Indian Ocean dipole impacts on the NAO. *Atmospheric Science Letters*, 21, e1005. https://doi.org/10.1002/asl.1005

Hawkins, E., Burt, S., Brohan, P., Lockwood, M., Richardson, H., Roy, M. et al. (2019) Hourly weather observations from the Scottish highlands (1883–1904) rescued by volunteer citizen scientists. *Geoscience Data Journal*, 6, 160–173. https://doi.org/10.1002/gdj3.79

Hempel, S., Frieler, K., Warszawski, L., Schewe, J. & Piontek, F. (2013) A trend-preserving bias correction – the ISI-MIP approach. *Earth System Dynamics*, 4, 219–236 Available from: https://esd.copernicus.org/articles/4/219/2013/

Hermanson, L., Ren, H.-L., Vellinga, M., Dunstone, N.D., Hyder, P., Ineson, S. et al. (2018) Different types of drifts in two seasonal forecast systems and their dependence on ENSO. *Climate Dynamics*, 51, 1411–1426. https://doi.org/10.1007/s00382-017-3962-9

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J. et al. (2020) The ERA5 global reanalysis *Q. Journal of the Royal Meteorological Society*, 146, 1999–2049. https://doi.org/10.1002/qj.3803

Hillier, J.K. & Dixon, R.S. (2020) Seasonal impact-based mapping of compound hazards. *Environmental Research Letters*, 15, 114013. https://doi.org/10.1088/1748-9326/abbc3d

Hollis, D., McCarthy, M., Kendon, M., Legg, T. & Simpson, I. (2019) Had UK-grid—a new UK dataset of gridded climate observations. *Geoscience Data Journal*, 6, 151–159. https://doi.org/10.1002/gdj3.78

Hoskins, B. (2013) The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science *Q. Journal of the Royal Meteorological Society*, 139, 573–584. https://doi.org/10.1002/qj.1991

Hoyer, S. & Hamman, J.J. (2017) Xarray: N-D labeled arrays and datasets in python. *Journal of Open Research Software*, 5(1), 10. https://doi.org/10.5334/jors.148

Jain, S., Scaife, A.A., Dunstone, N., Smith, D. & Mishra, S.K. (2020) Current chance of unprecedented monsoon rainfall over India using dynamical ensemble simulations. *Environmental Research Letters*, 15, 94095. https://doi.org/10.1088/1748-9326/ab7b98

Johnson, S.J., Stockdale, T.N., Ferranti, L., Balmaseda, M.A., Molteni, F., Magnusson, L. et al. (2019) SEAS5: the new ECMWF seasonal forecast system. *Geoscientific Model Development*, 12, 1087–1117 Available from: https://gmd.copernicus.org/articles/12/1087/2019/

Katz, R.W. (2013) Statistical methods for nonstationary extremes. In: *Extremes in a changing climate*. New York City: Springer, pp. 15–37.

Kay, A.L., Bell, V.A., Guillod, B.P., Jones, R.G. & Rudd, A.C. (2018) National-scale analysis of low flow frequency: historical trends and potential future changes. *Climatic Change*, 147, 585–599. https://doi.org/10.1007/s10584-018-2145-y

Kay, G., Dunstone, N., Smith, D., Dunbar, T., Eade, R. & Scaife, A. (2020) Current likelihood and dynamics of hot summers in the UK. *Environmental Research Letters*, 15, 094099. https://doi.org/10.1088/1748-9326/abab32/meta

Kelder, T., Müller, M., Slater, L.J., Marjoribanks, T.I., Wilby, R.L., Prudhomme, C. et al. (2020) Using UNSEEN trends to detect decadal changes in 100-year precipitation extremes. *npj Climate and Atmospheric Science*, 3, 47 Available from: https://www.nature.com/articles/s41612-020-00149-4

Kennedy-Asser, A.T., Andrews, O., Mitchell, D.M. & Warren, R.F. (2021) Evaluating heat extremes in the UK climate projections (UKCP18). *Environmental Research Letters*, 16, 14039. https://doi.org/10.1088/1748-9326/abc4ad

Kent, C., Dunstone, N., Tucker, S., Scaife, A.A., Brown, S., Kendon, E.J. et al. (2022) Estimating unprecedented extremes in UK summer daily rainfall. *Environmental Research Letters*, 17, 014041. https://doi.org/10.1088/1748-9326/ac42fb

Kent, C., Pope, E., Dunstone, N., Scaife, A.A., Tian, Z., Clark, R. et al. (2019) Maize drought hazard in the northeast farming region of China: unprecedented events in the current climate. *Journal of Applied Meteorology and Climatology*, 58, 2247–2258.

Kent, C., Pope, E., Thompson, V., Lewis, K., Scaife, A.A. & Dunstone, N. (2017) Using climate model simulations to assess the current climate risk to maize production. *Environmental Research Letters*, 12, 054012. https://doi.org/10.1088/1748-9326/aa6cb9

Kim, D.-I., Han, D. & Lee, T. (2021) Reanalysis product-based nonstationary frequency analysis for estimating extreme design rainfall. *Atmosphere*, 12, 191 Available from: https://www.mdpi.com/2073-4433/12/2/191

King, A.D., Karoly, D.J. & Henley, B.J. (2017) Australian climate extremes at 1.5°C and 2°C of global warming. *Nature Climate Change*, 7, 412–416 Available from: www.nature.com/natureclimatechange

Kirchmeier-Young, M.C. & Zhang, X. (2020) Human influence has intensified extreme precipitation in North America. *Proceedings of the National Academy of Sciences*, 117, 13308–13313. https://doi.org/10.1073/pnas.1921628117

Lange, S. (2019) Trend-preserving bias adjustment and statistical downscaling with ISIMIP3BASD (v1.0). *Geoscientific Model Development*, 12, 3055–3070.

Lavers, D.A., Pappenberger, F. & Zsoter, E. (2014) Extending medium-range predictability of extreme hydrological events in Europe. *Nature Communications*, 5, 5382.

Leach, N.J., Li, S., Sparrow, S., van Oldenborgh, G.J., Lott, F.C., Weisheimer, A. et al. (2020) Anthropogenic influence on the 2018 summer warm spell in Europe: the impact of different spatio-temporal scales. *Bulletin of the American Meteorological Society*, 101, S41–S46 Available from: https://journals.ametsoc.org/view/journals/bams/101/1/bams-d-19-0201.1.xml

Lehner, F., Coats, S., Stocker, T.F., Pendergrass, A.G., Sanderson, B.M., Raible, C.C. et al. (2017) Projected drought risk in 1.5°C and 2°C warmer climates. *Geophysical Research Letters*, 44, 7419–7428. https://doi.org/10.1002/2017GL074117

Lewis, S.C., King, A.D. & Mitchell, D.M. (2017) Australia's unprecedented future temperature extremes under Paris limits to warming. *Geophysical Research Letters*, 44, 9947–9956.

Liepert, B.G. & Previdi, M. (2012) Inter-model variability and biases of the global water cycle in CMIP3 coupled climate models. *Environmental Research Letters*, 7, 12. https://doi.org/10.1088/1748-9326/7/1/014006

Lorenz, E.N. (1963) Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20, 130–141.

Lucarini, V. & Ragone, F. (2011) Energetics of climate models: net energy balance and meridional enthalpy transport. *Reviews of Geophys*, 49, RG1001. https://doi.org/10.1029/2009RG000323

Mankin, J.S., Lehner, F., Coats, S. & McKinnon, K.A. (2020) The value of initial condition large ensembles to robust adaptation decision-making. *Earth's Future*, 8, e2012EF001610. https://doi.org/10.1029/2020EF001610

Maraun, D. & Widmann, M. (2018) *Statistical downscaling and bias correction for climate research*. Cambridge: Cambridge University Press. Available from: https://www.cambridge.org/core/books/statistical-downscaling-and-bias-correction-for-climate-research/4ED479BAA8309C7ECBE6136236E3960F

Matthews, T., Mullan, D., Wilby, R.L., Broderick, C. & Murphy, C. (2016) Past and future climate change in the context of memorable seasonal extremes. *Climate Risk Management*, 11, 37–52.

Meehl, G.A., Richter, J.H., Teng, H., Capotondi, A., Cobb, K., Doblas-Reyes, F. et al. (2021) Initialized earth system prediction from subseasonal to decadal timescales. *Nature Reviews Earth & Environment*, 2, 340–357.

Meucci, A., Young, I.R. & Breivik, Ø. (2018) Wind and wave extremes from atmosphere and wave model ensembles. *Journal of Climate*, 31, 8819–8842.

Miralles, D.G., Gentine, P., Seneviratne, S.I. & Teuling, A.J. (2019) Land-atmospheric feedbacks during droughts and heatwaves: state of the science and current challenges. *Annals of the New York Academy of Sciences*, 1436, 19–35. https://doi.org/10.1111/nyas.13912

Mitchell, D., AchutaRao, K., Allen, M., Bethke, I., Beyerle, U., Ciavarella, A. et al. (2017) Half a degree additional warming, prognosis and projected impacts (HAPPI): background and experimental design. *Geoscientific Model Development*, 10, 571–583 Available from: https://gmd.copernicus.org/articles/10/571/2017/

Moore, P., Hannah, B., de Vries, J., Poortvliet, M., Steffens, R. & Stoof, C.R. (2020) *Wildland fire management under COVID-19. Brief 1, review of materials*. Wageningen: Wageningen University. Available from: https://research.wur.nl/en/publications/wildland-fire-management-under-covid-19-brief-1-review-of-materia

Murphy, C., Wilby, R.L., Matthews, T., Horvath, C., Crampsie, A., Ludlow, F. et al. (2020) The forgotten drought of 1765–1768: reconstructing and re-evaluating historical droughts in the British and Irish Isles. *International Journal of Climatology*, 40, 5329–5351. https://doi.org/10.1002/joc.6521

Osinski, R., Lorenz, P., Kruschke, T., Voigt, M., Ulbrich, U., Leckebusch, G.C. et al. (2016) An approach to build an event set of European windstorms based on ECMWF EPS. *Natural Hazards and Earth System Sciences*, 16, 255–268 Available from: https://www.nat-hazards-earth-syst-sci.net/16/255/2016/

Overland, J.E. & Wang, M. (2021) The 2020 Siberian heat wave. *International Journal of Climatology*, 41(S1), E2341–E2346. https://doi.org/10.1002/joc.6850

Palmer, T. (2019) The ECMWF ensemble prediction system: looking back (more than) 25 years and projecting forward 25 years Q. *Journal of the Royal Meteorological Society*, 145, 12–24. https://doi.org/10.1002/qj.3383

Palmer, T., Doblas-Reyes, F., Hagedorn, R. & Weisheimer, A. (2005) Probabilistic prediction of climate using multi-model ensembles: from basics to applications. *Philosophical Transactions of the Royal Society B*, 360, 1991–1998. https://doi.org/10.1098/rstb.2005.1750

Palmer, T.N. & Weisheimer, A. (2018) A simple pedagogical model linking initial-value reliability with trustworthiness in the forced climate response. *Bulletin of the American Meteorological Society*, 99, 605–614. https://doi.org/10.1175/BAMS-D-16-0240.1

Pappenberger, F. & Palkovic, M. (2020) Progress towards a European Weather Cloud. ECMWF. *ECMWF Newsletter*, 165, 24–27 Available from: https://doi.org/10.21957/9pft4uy055

Pappenberger, F., Rabier, F. & Venuti, F. (2021) Invited perspectives: the ECMWF strategy 2021–2030 challenges in the area of natural hazards. *Natural Hazards and Earth System Sciences*, 21, 2163–2167 Available from: https://doi.org/10.5194/nhess-21-2163-2021

Parker, W.S. (2016) Reanalyses and observations: What's the difference? *Bulletin of the American Meteorological Society*, 97, 1565–1572.

Pascale, S., Kapnick, S.B., Delworth, T.L. & Cooke, W.F. (2020) Increasing risk of another Cape Town “day zero” drought in the 21st century. *Proceedings of the National Academy of Sciences of the United States of America*, 117, 29495–29503 Available from: https://www.pnas.org/content/117/47/29495

Philip, S., Kew, S., van Oldenborgh, G.J., Otto, F., Vautard, R., van der Wiel, K. et al. (2020) A protocol for probabilistic extreme event attribution analyses. *Advances in Statistical Climatology, Meteorology and Oceanography*, 6, 177–203 Available from: https://ascmo.copernicus.org/articles/6/177/2020/

Pickrell, J. & Pennisi, E. (2020) Record U.S. and Australian fires raise fears for many species. *Science*, 370, 18–19 Available from: https://science.sciencemag.org/content/370/6512/18

Poschlod, B., Ludwig, R. & Sillmann, J. (2021) Ten-year return levels of sub-daily extreme precipitation over Europe. *Earth System Science Data*, 13, 983–1003.

Power, S.B. & Delage, F.p.D. (2019) Setting and smashing extreme temperature records over the coming century. *Nature Climatic Change*, 97(9), 529–534 Available from: https://www.nature.com/articles/s41558-019-0498-5

Rahmstorf, S. (1995) Climate drift in an ocean model coupled to a simple, perfectly matched atmosphere. *Climate Dynamics*, 11, 447–458. https://doi.org/10.1007/BF00207194

Robinson, A., Lehmann, J., Barriopedro, D., Rahmstorf, S. & Coumou, D. (2021) Increasing heat and rainfall extremes now far outside the historical climate. *npj Climate and Atmospheric Sciences*, 41(4), 1–4 Available from: https://www.nature.com/articles/s41612-021-00202-w

Salas, J.D., Obeysekera, J. & Vogel, R.M. (2018) Techniques for assessing water infrastructure for nonstationary extreme events: a review. *Hydrological Science Journal*, 63, 325–352. https://doi.org/10.1080/02626667.2018.1426858

Sen, G.A., Jourdain, N.C., Brown, J.N. & Monselesan, D. (2013) Climate drift in the CMIP5 models. *Journal of Climate*, 26, 8597–8615.

Sen, G.A., Santoso, A., Taschetto, A.S., Ummenhofer, C.C., Trevena, J. & England, M.H. (2009) Projected changes to the Southern Hemisphere ocean and sea ice in the IPCC AR4 climate models. *Journal of Climate*, 22, 3047–3078 Available from: https://journals.ametsoc.org/view/journals/clim/22/11/2008jcli2827.1.xml

Sillmann, J., Shepherd, T.G., van den Hurk, B., Hazeleger, W., Martius, O., Slingo, J. et al. (2021) Event-based storylines to address climate risk. *Earth's Future*, 9, e2020EF001783. https://doi.org/10.1029/2020EF001783

Sillmann, J., Thorarinsdottir, T., Keenlyside, N., Schaller, N., Alexander, L.V., Hegerl, G. et al. (2017) Understanding, modeling and predicting weather and climate extremes: challenges and opportunities. *Weather and Climate Extremes*, 18, 65–74.

Slater, L.J., Anderson, B., Buechel, M., Dadson, S., Han, S., Harrigan, S. et al. (2021) Nonstationary weather and water extremes: a review of methods for their detection, attribution, and management. *Hydrology and Earth System Sciences*, 25, 3897–3935 Available from: https://hess.copernicus.org/articles/25/3897/2021/

Squire, D.T., Richardson, D., Risbey, J.S., Black, A.S., Kitsios, V., Matear, R.J. et al. (2021) Likelihood of unprecedented drought and fire weather during Australia's 2019 megafires. *npj Climate and Atmospheric Sciences*, 41(4), 1–12 Available from: https://www.nature.com/articles/s41612-021-00220-8

Stevenson, S., Timmermann, A., Chikamoto, Y., Langford, S. & DiNezio, P. (2015) Stochastically generated North American megadroughts. *Journal of Climate*, 28, 1865–1880 Available from: https://journals.ametsoc.org/view/journals/clim/28/5/jcli-d-13-00689.1.xml

Stott, p.A., Christidis, N., Otto, F.E.L., Sun, Y., Vanderlinden, J.-P., van Oldenborgh, G.J. et al. (2016) Attribution of extreme weather and climate-related events. *Wiley Interdisciplinary Reviews: Climate Change*, 7, 23–41. https://doi.org/10.1002/wcc.380

Suarez-Gutierrez, L., Milinski, S. & Maher, N. (2021) Exploiting large ensembles for a better yet simpler climate model evaluation. *Climate Dynamics*, 1, 2557–2580. https://doi.org/10.1007/s00382-021-05821-w

Suarez-Gutierrez, L., Müller, W.A., Li, C. & Marotzke, J. (2020a) Dynamical and thermodynamical drivers of variability in European summer heat extremes. *Climate Dynamics*, 54, 4351–4366. https://doi.org/10.1007/s00382-020-05233-2

Suarez-Gutierrez, L., Müller, W.A., Li, C. & Marotzke, J. (2020b) Hotspots of extreme heat under global warming. *Climate Dynamics*, 55, 429–447. https://doi.org/10.1007/s00382-020-05263-w

Swain, D.L., Wing, O.E.J., Bates, p.D., Done, J.M., Johnson, K.A. & Cameron, D.R. (2020) Increased flood exposure due to climate change and population growth in the United States. *Earth's Future*, 8, e2020EF001778. https://doi.org/10.1029/2020EF001778

Swindles, G.T., Morris, p.J., Mullan, D., Watson, E.J., Turner, T.E., Roland, T.P. et al. (2015) The long-term fate of permafrost peatlands under rapid climate warming. *Scientific Report*, 51(5), 1–6 Available from: https://www.nature.com/articles/srep17951

Tebaldi, C. & Knutti, R. (2007) The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365, 2053–2075.

Thepaut, J.-N., Dee, D., Engelen, R. & Pinty, B. (2018) The Copernicus programme and its climate change service. *IGARSS 2018–2018 IEEE international geoscience and remote sensing symposium*, 2018, pp. 1591–1593. Available at: https://ieeexplore.ieee.org/document/8518067/.

Thompson, V., Dunstone, N.J., Scaife, A.A., Smith, D.M., Hardiman, S.C., Ren, H.-L. et al. (2019) Risk and dynamics of unprecedented hot months in south East China. *Climate Dynamics*, 52, 2585–2596. https://doi.org/10.1007/s00382-018-4281-5

Thompson, V., Dunstone, N.J., Scaife, A.A., Smith, D.M., Slingo, J. M., Brown, S. et al. (2017) High risk of unprecedented UK rainfall in the current climate. *Nature Communications*, 8, 107.

Tompkins, A.M., De Záratate, M.I.O., Saurral, R.I., Vera, C., Saulo, C., Merryfield, W.J. et al. (2017) The climate-system historical forecast project: providing open access to seasonal forecast ensembles from centers around the globe. *Bulletin of the American Meteorological Society*, 98, 2293–2301 Available at: https://journals.ametsoc.org/view/journals/bams/98/11/bams-d-16-0209.1.xml

Uhe, P., Otto, F.E.L., Haustein, K., van Oldenborgh, G.J., King, A. D., Wallom, D.C.H. et al. (2016) Comparison of methods: attributing the 2014 record European temperatures to human influences. *Geophysical Research Letters*, 43, 8685–8693. https://doi.org/10.1002/2016GL069568

UK Met Office (2020) Record breaking rainfall. Available at: https://www.metoffice.gov.uk/about-us/press-office/news/weather-and-climate/2020/2020-winter-february-stats

van den Brink, H.W., Können, G.P., Opsteegh, J.D., van Oldenborgh, G.J. & Burgers, G. (2004) Improving 10 4-year surge level estimates using data of the ECMWF seasonal prediction system. *Geophysical Research Letters*, 31, L17210. https://doi.org/10.1029/2004GL020610

van den Brink, H.W., Können, G.P., Opsteegh, J.D., van Oldenborgh, G.J. & Burgers, G. (2005) Estimating return periods of extreme events from ECMWF seasonal forecast ensembles. *International Journal of Climatology*, 25, 1345–1354. https://doi.org/10.1002/joc.1155

Van der Wiel, K., Kapnick, S.B., van Oldenborgh, G.J., Whan, K., Philip, S., Vecchi, G.A. et al. (2017) Rapid attribution of the August 2016 flood-inducing extreme precipitation in South Louisiana to climate change. *Hydrology and Earth System Sciences*, 21, 897–921 Available from: https://hess.copernicus.org/articles/21/897/2017/

Van der Wiel, K., Kapnick, S.B., Vecchi, G.A., Smith, J.A., Milly, p.C.D. & Jia, L. (2018) 100-year lower Mississippi floods in a global climate model: characteristics and future changes. *Journal of Hydrometeorology*, 19, 1547–1563 Available from: https://journals.ametsoc.org/view/journals/hydr/19/10/jhm-d-18-0018_1.xml

Van der Wiel, K., Selten, F.M., Bintanja, R., Blackport, R. & Screen, J.A. (2020) Ensemble climate-impact modelling: extreme impacts from moderate meteorological conditions. *Environmental Research Letters*, 15, 034050. https://doi.org/10.1088/1748-9326/ab7668

Van der Wiel, K., Wanders, N., Selten, F.M. & Bierkens, M.F.P. (2019) Added value of large ensemble simulations for assessing extreme river discharge in a 2°C warmer world. *Geophysical Research Letters*, 46, 2093–2102. https://doi.org/10.1029/2019GL081967

van Kempen, G., van der Wiel, K. & Melsen, L.A. (2021) The impact of hydrological model structure on the simulation of extreme runoff events. *Natural Hazards and Earth System Sciences*, 21, 961–976 Available from: https://nhess.copernicus.org/articles/21/961/2021/

van Oldenborgh, G.J., van der Wiel, K., Kew, S., Philip, S., Otto, F., Vautard, R. et al. (2021) Pathways and pitfalls in extreme event attribution. *Climatic Change*, 166, 13–27. https://doi.org/10.1007/s10584-021-03071-7

Vautard, R., Christidis, N., Ciavarella, A., Alvarez-Castro, C., Bellprat, O., Christiansen, B. et al. (2019) Evaluation of the HadGEM3-a simulations in view of detection and attribution of human influence on extreme events in Europe. *Climate Dynamics*, 52, 1187–1210.

Vitolo, C., Di Giuseppe, F., Barnard, C., Coughlan, R., San-Miguel-Ayanz, J., Libertá, G. et al. (2020) ERA5-based global meteorological wildfire danger maps. *Scientific Data*, 7, 1–11. www.nature.com/scientificdata

Wagemann, J., Clements, O., Marco Figuera, R., Rossi, A.P. & Mantovani, S. (2018) Geospatial web services pave new ways for server-based on-demand access and processing of big earth data. *International Journal of Digital Earth*, 11, 7–25 Available at: https://www.tandfonline.com/action/journalInformation?journalCode=tjde20

Walz, M.A. & Leckebusch, G.C. (2019) Loss potentials based on an ensemble forecast: how likely are winter windstorm losses similar to 1990? *Atmospheric Science Letters*, 20, e891. https://doi.org/10.1002/asl.891

Warszawski, L., Frieler, K., Huber, V., Piontek, F., Serdeczny, O. & Schewe, J. (2014) The inter-sectoral impact model intercomparison project (ISI–MIP): project framework. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 3228–3232 Available from: https://www.pnas.org/content/111/9/3228

Weber, E.U. (2006) Experience-based and description-based perceptions of long-term risk: why global warming does not scare us (yet). *Climatic Change*, 771(77), 103–120. https://doi.org/10.1007/s10584-006-9060-3

Wehner, M., Gleckler, P. & Lee, J. (2020) Characterization of long period return values of extreme daily temperature and precipitation in the CMIP6 models: part 1, model evaluation. *Weather and Climate Extremes*, 30, 100283.

Weigel, K., Bock, L., Gier, B.K., Lauer, A., Righi, M., Schlund, M. 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. *Geoscientific Model Development*, 14, 3159–3184 Available from: https://gmd.copernicus.org/articles/14/3159/2021/

Weisheimer, A., Befort, D.J., MacLeod, D., Palmer, T., O'Reilly, C. & Strømmen, K. (2021) Seasonal forecasts of the twentieth century. *Bulletin of the American Meteorological Society*, 101, E1413–E1426. https://doi.org/10.1175/BAMS-D-19-0019.1

Weisheimer, A., Decremer, D., MacLeod, D., O'Reilly, C., Stockdale, T.N., Johnson, S. et al. (2019) How confident are predictability estimates of the winter North Atlantic oscillation? *Quarterly Journal of the Royal Meteorological Society*, 145, 140–159. https://doi.org/10.1002/qj.3446

Weisheimer, A. & Palmer, T.N. (2014) On the reliability of seasonal climate forecasts. *Journal of the Royal Society Interface*, 11, 20131162. https://doi.org/10.1098/rsif.2013.1162

Weisheimer, A., Schaller, N., O'Reilly, C., MacLeod, D.A. & Palmer, T. (2017) Atmospheric seasonal forecasts of the twentieth century: multi-decadal variability in predictive skill of the winter North Atlantic oscillation (NAO) and their potential value for extreme event attribution *Q. Journal of the Royal Meteorological Society*, 143, 917–926. https://doi.org/10.1002/qj.2976

Wilby, R.L., Clifford, N.J., De Luca, P., Harrigan, S., Hillier, J.K., Hodgkins, R. et al. (2017) The 'dirty dozen' of freshwater science: detecting then reconciling hydrological data biases and errors. *Wiley Interdisciplinary Reviews: Water*, 4, e1209.

Wilcke, R.A.I., Kjellström, E., Lin, C., Matei, D., Moberg, A. & Tyrlis, E. (2020) The extremely warm summer of 2018 in Sweden—set in a historical context. *Earth System Dynamics*, 11, 1107–1121 Available from: https://esd.copernicus.org/articles/11/1107/2020/

Wilks, D.S. (2011) *Statistical methods in the atmospheric sciences*, Vol. 100. Cambridge, Massachusetts: Academic Press.

Wilks, D.S. & Wilby, R.L. (1999) The weather generation game: a review of stochastic weather models. *Progress in Physical Geography: Earth and Environment*, 23, 329–357. https://doi.org/10.1177/030913339902300302

Yan, H., Liu, C., An, Z., Yang, W., Yang, Y., Huang, P. et al. (2020) Extreme weather events recorded by daily to hourly resolution biogeochemical proxies of marine giant clam shells. *Proceedings of the National Academy of Sciences of the United States of America*, 117, 7038–7043 Available from: https://www.pnas.org/content/117/13/7038

Yiou, P. (2014) AnaWEGE: a weather generator based on analogues of atmospheric circulation. *Geoscientific Model Development*, 7, 531–543.

Zadra, A., Williams, K., Frassoni, A., Rixen, M., Adames, Á.F., Berner, J. et al. (2018) Systematic errors in weather and climate models: nature, origins, and ways forward. *Bulletin of the American Meteorological Society*, 99, ES67–ES70 Available from: http://collaboration.cmc.ec.gc.ca

Zuo, H., Alonso-Balmaseda, M.A., Mogensen, K. & Tietsche, S. (2018) *OCEAN5: the ECMWF ocean reanalysis system and its real-time analysis component*. Reading: European Centre for Medium-Range Weather Forecasts. http://dx.doi.org/10.21957/la2v0442

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Kelder, T., Marjoribanks, T. I., Slater, L. J., Prudhomme, C., Wilby, R. L., Wagemann, J., & Dunstone, N. (2022). An open workflow to gain insights about low-likelihood high-impact weather events from initialized predictions. *Meteorological Applications*, *29*(3), e2065. https://doi.org/10.1002/met.2065