Global Extreme Heat Forecasting Using Neural Weather Models🖉

Ignacio Lopez-Gomez🄳,[a,b] Amy McGovern,[a,c] Shreya Agrawal,[a] and Jason Hickey[a]

[a] *Google Research, Mountain View, California*
[b] *California Institute of Technology, Pasadena, California*
[c] *University of Oklahoma, Norman, Oklahoma*

ABSTRACT: Heatwaves are projected to increase in frequency and severity with global warming. Improved warning systems would help reduce the associated loss of lives, wildfires, power disruptions, and reduction in crop yields. In this work, we explore the potential for deep learning systems trained on historical data to forecast extreme heat on short, medium and subseasonal time scales. To this purpose, we train a set of neural weather models (NWMs) with convolutional architectures to forecast surface temperature anomalies globally, 1 to 28 days ahead, at ~200-km resolution and on the cubed sphere. The NWMs are trained using the ERA5 reanalysis product and a set of candidate loss functions, including the mean-square error and exponential losses targeting extremes. We find that training models to minimize custom losses tailored to emphasize extremes leads to significant skill improvements in the heatwave prediction task, relative to NWMs trained on the mean-square-error loss. This improvement is accomplished with almost no skill reduction in the general temperature prediction task, and it can be efficiently realized through transfer learning, by retraining NWMs with the custom losses for a few epochs. In addition, we find that the use of a symmetric exponential loss reduces the smoothing of NWM forecasts with lead time. Our best NWM is able to outperform persistence in a regressive sense for all lead times and temperature anomaly thresholds considered, and shows positive regressive skill relative to the ECMWF subseasonal-to-seasonal control forecast after 2 weeks.

SIGNIFICANCE STATEMENT: Heatwaves are projected to become stronger and more frequent as a result of global warming. Accurate forecasting of these events would enable the implementation of effective mitigation strategies. Here we analyze the forecast accuracy of artificial intelligence systems trained on historical surface temperature data to predict extreme heat events globally, 1 to 28 days ahead. We find that artificial intelligence systems trained to focus on extreme temperatures are significantly more accurate at predicting heatwaves than systems trained to minimize errors in surface temperatures and remain equally skillful at predicting moderate temperatures. Furthermore, the extreme-focused systems compete with state-of-the-art physics-based forecast systems in the subseasonal range, while incurring a much lower computational cost.

KEYWORDS: Forecasting; Forecasting techniques; Artificial intelligence; Deep learning; Machine learning; Neural networks

---

1. Introduction

An important consequence of anthropogenic radiative forcing is the robust increase in heatwave days per year, both globally and at a regional level (Perkins-Kirkpatrick and Lewis 2020). Heatwaves pose a significant health risk, as evidenced by the more than 70 000 excess deaths that occurred during the 2003 European heatwave (Robine et al. 2008). More frequent heatwaves will also lead to higher wildfire risk (Parente et al. 2018; Ruffault et al. 2020), stress on the power grid (Ke et al. 2016), and loss of agricultural crops (Brás et al. 2021). These trends underscore the importance of developing effective mitigation strategies to reduce the negative impacts of extreme heat. Accurate forecasts with sufficient lead time are a stepping stone in the development of such strategies (Lin et al. 2022). Current physics-based models, however, can only provide accurate forecasts of extreme heat events a few days in advance, which may not be sufficient to deploy effective mitigation strategies (White et al. 2017; Wulff and Domeisen 2019).

There is mounting evidence of heatwave predictors on weekly to subseasonal time scales. These include large-scale quasi-stationary atmospheric Rossby waves (Teng et al. 2013; Mann et al. 2018; White et al. 2022), negative soil moisture anomalies (Vautard et al. 2007; Benson and Dirmeyer 2021), and anomalous Pacific Ocean sea surface temperature (SST) gradients in the case of North American heatwaves (Deng et al. 2018; Miller et al. 2021). Recently, Miller et al. (2021) used a linear regression model based on empirical orthogonal functions of the North Pacific SST and soil moisture over the United States to predict the weekly frequency of extremely warm days in the United States, 1–4 weeks ahead. They show that their statistical model outperforms the operational NCEP CFSv2 model in this task over the eastern United States after the second week, which suggests that purely data-

---

driven forecasting may provide a path forward in extreme heat prediction beyond the 10-day horizon.

In this context, deep neural networks represent a natural extension of the data-driven approach, given their remarkable success in image segmentation and forecasting tasks (Ronneberger et al. 2015; Sønderby et al. 2020; Ravuri et al. 2021). Different methods to classify heatwaves leveraging deep learning have recently been proposed. Chattopadhyay et al. (2020) trained a capsule neural network on midtropospheric geopotential and surface temperature from a large ensemble of climate model runs to classify several future days into five different classes, each representing either a specific heatwave pattern over North America, or the absence of extreme temperatures. Jacques-Dumas et al. (2022) trained a convolutional neural network on wavenumber space to forecast the occurrence of heatwaves in France with a 15-day lead time. They used the same predictors as Chattopadhyay et al. (2020), and data from a 1000-yr cyclic climate model simulation.

These studies showcase the potential of deep learning to forecast extreme heat events as a classification problem, in particular regions, and trained on very large datasets sampled from quasi-stationary distributions. Here, we tackle some of the practical questions left unanswered by previous work:

- Can deep learning models forecast extreme heat events when trained on limited historical data? The use of observations or reanalysis data is crucial for systems to improve upon existing physics-based models, since deep learning models trained solely on synthetic data will at best inherit the biases of the numerical models they are trying to substitute.
- Can general purpose neural weather models (NWMs) be used to predict extreme heat events? By general purpose, we refer to deep learning systems trained to minimize errors in the underlying fields, such as temperature, and not on extreme classification explicitly (e.g., Rasp et al. 2020; Weyn et al. 2020; Pathak et al. 2022; Keisler 2022).
- Can NWMs improve their extreme prediction skill through the use of custom losses while retaining their skill in a general weather forecasting setting?

To answer these questions, we frame extreme heat prediction as a regression problem and restrict our training data to a large subset of the ERA5 reanalysis product. Framing the forecast problem as a regression task bridges the gap with NWMs that act as time integrators and require reliable previous forecasts as inputs to generate a new prediction (Weyn et al. 2020). The regression problem is also more robust to the definition of heatwaves, diverse in the literature (e.g., Chattopadhyay et al. 2020; Wulff and Domeisen 2019; Miller et al. 2021), and allows learning about the nuances of target states that may otherwise be masked under the same class in a classification problem. The advent of skillful regression-based NWMs would democratize the use of ensemble-based weather forecasting for targeted applications, which requires enormous computational resources when realized through state-of-the-art physics-based models (Palmer 2017). In contrast, deep learning systems designed to forecast a few fields of interest (e.g., surface temperature) only incur

high computational costs during training, but not during inference (Scher and Messori 2021; Weyn et al. 2021).

To address the second question, we make use of a state-of-the-art convolutional architecture on the cubed sphere, following Weyn et al. (2020), so that results can be extrapolated to similar NWMs described in the literature. To explore the last question, we compare forecasts of NWMs trained with the general-purpose mean-square-error loss with forecasts from NWMs trained to minimize custom losses that emphasize extremes. All results presented are contextualized through comparison with the European Centre for Medium-Range Weather Forecasts (ECMWF) subseasonal-to-seasonal (S2S) operational forecast system (Vitart et al. 2017).

The paper is organized as follows. In section 2, we define the forecasting task and describe the data and losses used to train the NWMs. In section 3, the model architecture is discussed. Section 4 explores the skill of NWMs trained using different loss functions in tasks varying from extreme heat to general surface temperature prediction, including example forecasts for the 2017 Iberian heatwave and the 2021 western North American heatwave. In section 5, the relevance of the different NWM inputs is explored using integrated gradients (Sundararajan et al. 2017). Section 6 ends with a discussion of the results and potential future research directions.

2. The forecasting problem

The forecasting task, given a set of observed input and target pairs $(\mathbf{x}, \mathbf{y})$ with $\mathbf{y} \in \mathbf{\Omega} \subset \mathbb{R}^d$ and $\mathbf{x} \in \mathbf{V} \subset \mathbb{R}^q$, can be framed as the following minimization problem (Lopez-Gomez et al. 2022),

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \int_{\mathbf{\Omega}} L[\boldsymbol{\Psi}(\boldsymbol{\theta}, \mathbf{x}), \mathbf{y}] \, d\mathbf{y}, \qquad (1)$$

where $\boldsymbol{\Psi} : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^d$ is a mapping from parameter and input space to target space, $\boldsymbol{\theta} \in \mathbb{R}^p$ are model parameters, $\mathbf{x} \in \mathbb{R}^q$ is an input data vector, and $L(,)$ is some loss that we seek to minimize over a target set $\mathbf{\Omega} \subset \mathbb{R}^d$. In this paper, we consider $\boldsymbol{\Psi}$ to be a convolutional neural network operating on a gnomonic equiangular cubed sphere, described in detail in section 3. Architecture exploration for NWMs is an active area of research (Pathak et al. 2022; Keisler 2022), but it is not the emphasis of this work, so we keep the architecture $\boldsymbol{\Psi}$ fixed. Instead, we are interested in comparing the usefulness of NWMs $\boldsymbol{\Psi}(\boldsymbol{\theta}^*,)$ that result from the minimization (1) when varying the definition of the loss. The NWM parameter vector $\boldsymbol{\theta}^*$ is obtained through minibatch gradient descent, using the Adam algorithm (Kingma and Ba 2017).

a. Data, predictors, and targets

The targets $\mathbf{y}$ are constructed from the daily average of the standardized climatological anomalies of the temperature 2 m above the surface $T_{2\mathrm{m}}$, which we denote $\tilde{T}_{2\mathrm{m}}$. We recognize that air temperature is a suboptimal indicator of heat-related illness (Xu et al. 2016; Heo et al. 2019), but it enables comparison with other models in the literature (Weyn et al. 2021;
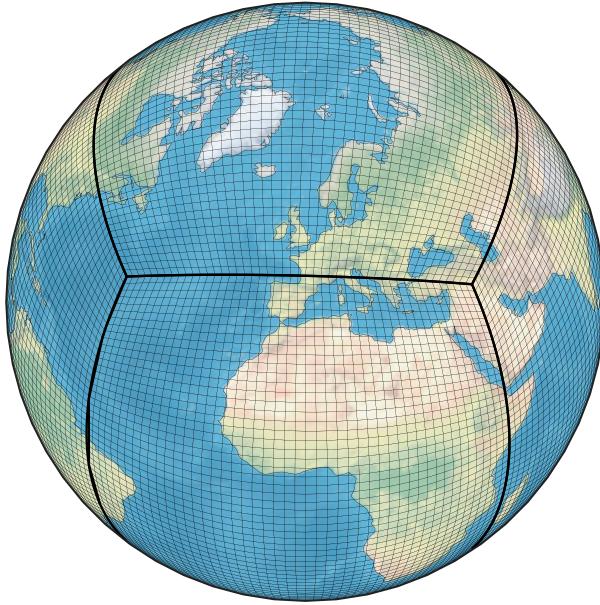
FIG. 1. Depiction of the gnomonic cubed-sphere grid onto which predictors and targets are projected. The cubed sphere is composed of 6 faces with $48^2 = 2304$ cells each.

Wulff and Domeisen 2019; Lin et al. 2022). Each target vector **y** includes $\tilde{T}_{2m}$ for a set of lead times $\tau = 1, \ldots, \tau_l$ days and for all tiles of a gnomonic equiangular grid of Earth's surface (Ronchi et al. 1996). Thus, the target size d is the flattened length of the tensor $\mathbf{y}_{cs} \in \mathbb{R}^{\tau_l \times f \times h \times w}$, where $f = 6$ is the number of faces of the cubed sphere and $h = 48$ and $w = 48$ are the number of meridional and zonal tiles of each face, respectively. The grid is shown in Fig. 1 for reference; the surface area of each tile is approximately 192^2 km^2. To assess the skill of the model from the short to the subseasonal range, we aim to predict temperature anomalies for the next $\tau_l = 28$ days.

The inputs $\mathbf{x} \in \mathbf{V} \subset \mathbb{R}^q$ contain υ daily averaged surface fields on the cubed sphere, such that q is the flattened length of the tensor $\mathbf{x}_{cs} \in \mathbb{R}^{\upsilon \times f \times h \times w}$. Here, $\upsilon = \tau_p \upsilon_t + \upsilon_i$ is the number of input fields and υ_t and υ_i are the number of time-dependent and independent fields, respectively. For time-dependent fields, daily averages from the last $\tau_p = 7$ days are included. We consider as time-dependent fields the geopotential height and potential vorticity at 300, 500 and 700 hPa, the temperature T_{2m}, the top net outgoing longwave radiative flux (OLR), and the volume of soil water. To these fields we append as auxiliary time-independent variables the latitude, longitude, topography, and a land–sea mask. We also include the present insolation and date as inputs. The full list of predictors is summarized in Table 1.

The choice of predictors is informed by studies linking extreme heat events to midtropospheric geopotential height (Teng et al. 2013; Mann et al. 2018), soil moisture (Vautard et al. 2007; Benson and Dirmeyer 2021) and large-scale phenomena with characteristic OLR signatures, like the Madden–Julian oscillation (MJO; Jacques-Coper et al. 2015; Maloney et al. 2019) or the boreal summer intraseasonal oscillation (Lin et al. 2022).

TABLE 1. Predictors used for the temperature anomaly forecasting task. Anomalies are standardized (std) with respect to climatology. Heights are specified as above ground level (a.g.) and below ground level (b.g.).

Predictor	Levels
Temperature anomaly (std)	2 m (a.g.)
Temperature	2 m (a.g.)
Geopotential height anomaly (std)	300, 500, and 700 hPa
Geopotential height	300, 500, and 700 hPa
Potential vorticity anomaly (std)	300, 500, and 700 hPa
Potential vorticity	300, 500, and 700 hPa
Outgoing longwave radiation	Top of atmosphere
Incoming shortwave radiation	Top of atmosphere
Surface soil moisture	0–7 and 7–28 cm (b.g.)
Deep soil moisture	28–100 and 100–289 cm (b.g.)
Topography	—
Land–sea mask	—
Latitude, longitude	—
Date	—

In addition, we include T_{2m} to learn about transport processes such as advection; and its standardized anomaly to facilitate improving upon a persistence model. Both T_{2m} and $\tilde{T}_{2m}$ capture the signature of large-scale oscillations such as El Niño–Southern Oscillation or the North Atlantic Oscillation (Ogi et al. 2003; Wang et al. 2011; Wright et al. 2014).

All data are daily averages from the ERA5 reanalysis product (Hersbach et al. 2020), downloaded at $2° \times 2°$ resolution from the Copernicus Climate Data Store (CDS) and projected onto the cubed sphere following Ullrich and Taylor (2015), Ullrich et al. (2016). The climatology is computed for the time period 1979–2019. All days of the year, not only the summer days, are used to train the model. We do this to learn about physical processes that are season independent, like advection. When framed as a classification task, extreme heat prediction may require undersampling of nonextreme samples during training (Jacques-Dumas et al. 2022). Here, we make use of all available information with no explicit undersampling; class imbalance is dealt with through the use of custom losses as discussed in section 2b. In addition, we perform a sequential split of the data into training (1979–2012), validation (2013–16), and test sets (2017–21). The limited amount of historical data available means that the influence of longer modes of climate variability (e.g., the Pacific decadal oscillation) is unlikely to be robustly captured. Furthermore, a shift in the target distribution from training to testing sets is implicit with this split, due to climate change (White et al. 2022; Chan et al. 2020). This shift in the temperature distribution is, however, representative of situations in which a warning system might be used in practice, since both data-driven and NWP models are calibrated using historical observations.

b. Losses considered

Unless the optimal parameter vector $\boldsymbol{\theta}^*$ is able to yield a perfect model [i.e., $\boldsymbol{\Psi}(\boldsymbol{\theta}^*, \mathbf{x}) = \mathbf{y} \; \forall \; (\mathbf{x}, \mathbf{y})$], the optimum will depend on the definition of the loss. This is the case in the

extreme heat prediction task, since the chaotic nature of the atmosphere precludes a perfect forecast of trajectories from inexact initial conditions (Lorenz 1969a,b; Slingo and Palmer 2011). For this reason, we can expect models that minimize generic losses to be suboptimal for the extreme prediction task.

To study the potential benefits of training NWMs on losses targeting extreme prediction, we consider two losses: the mean-square error (MSE) and a custom loss L_e based on the exponential of targets and forecasts,

$$L_e(\mathbf{y}', \mathbf{y}) = a\text{MSE}(e^{\mathbf{y}'}, e^{\mathbf{y}}) + b\text{MSE}(e^{-\mathbf{y}'}, e^{-\mathbf{y}}), \qquad (2)$$

where $\mathbf{y}$ are the targets and $\mathbf{y}'$ are the forecasts. In (2), the choice $(a, b) = (1, 0)$ emphasizes the correct prediction of positive extremes, $(a, b) = (0, 1)$ emphasizes negative extremes, and the midpoint $(a, b) = (0.5, 0.5)$ emphasizes both extremes. In the context of temperature anomaly prediction, these losses emphasize heatwaves, cold spells and extreme deviations from climatology, respectively. This custom loss is motivated by the reported success of neural networks in extreme prediction tasks when using target transformations involving the softmax and softmin functions (Qi and Majda 2020). However, we exclusively take the numerator of the suggested transformations because the softmax and softmin functions are invariant under translational shifts (Lopez-Gomez et al. 2020, their appendix A), which in this application means that climatological biases would not be penalized.

In the following, we denote models trained with the loss (2) as HeatNet for $(a, b) = (1, 0)$ and ExtNet for $(a, b) = (0.5, 0.5)$. The model trained with the MSE loss, representative of general neural weather prediction systems (Rasp et al. 2020; Weyn et al. 2020, 2021), is denoted GenNet. We trained HeatNet, ExtNet, and GenNet models using a hyperparameter search over the learning rate, and the magnitude of L1 and L2 norm regularization. All models were trained until they started overfitting to the training set, evidenced by an increase in validation loss persistent over many epochs. Notably, our best HeatNet and ExtNet models were obtained through transfer learning, by retraining our best GenNet model for a few (<3) epochs on the custom exponential loss. This implies that any performance improvements of HeatNet or ExtNet with respect to generic models trained on the MSE loss can be realized efficiently through transfer learning from the original models. Our transfer learning methodology relies on early stopping to retain an inductive bias toward the GenNet parameters (Yosinski et al. 2014; Li et al. 2018). The implementation of other transfer learning techniques, like Bayesian regularization toward the original (i.e., prior) model parameters (Li et al. 2018; Inubushi and Goto 2020), may result in further skill improvements and will be explored in the future.

3. Neural weather model architecture

We employ a convolutional architecture to construct the neural network $\boldsymbol{\Psi}$, which maps the input fields at all past times $\tau = -6, \ldots, 0$ days to the daily averaged temperature anomaly $\tilde{T}_{2\text{m}}$ at all lead times $\tau = 1, \ldots, 28$ days. Consistent with our projection of the data, convolutions are performed on each of the cubed sphere faces, using halo exchange at the borders (Weyn et al. 2020). Kernel weights are shared among all four equatorial faces, and a different set of kernel weights is used for the polar faces. This enables learning about different processes governing on one hand tropical and subtropical dynamics, and on the other mid and high-latitude dynamics. The northern polar face is mirrored before each convolution to align cyclonic and anticyclonic motions in each hemisphere, following Weyn et al. (2020).

a. Receptive field

Because of the nonrecurrent nature of the architecture and the lead times considered, it is crucial to achieve a fully receptive field if we want to capture long-range dependencies and teleconnections (Espeholt et al. 2022). A fully receptive field is realized through two design characteristics of the proposed architecture, which is sketched in Fig. 2. The first one is the use of dilated convolutions, which rapidly increase the receptive field of any location on the cubed sphere as information traverses the network (Yu and Koltun 2016). The second one is the use of a UNet-type architecture (Ronneberger et al. 2015) with 3 resolution levels going from the data resolution to the synoptic scale: $\sim200^2$, $\sim400^2$, and $\sim800^2$ km². Coarser-resolution levels increase the receptive field proportionally to their downsampling rate, allowing one to achieve larger receptive fields with fewer layers.

b. Encoder and decoder architecture

The architecture of our model is based on the UNet 3+ architecture (Huang et al. 2020) with a few modifications. All nonlinearities consist of parametric rectified linear units (PReLUs) that share parameters across all dimensions except the channels (He et al. 2015). We use dilated convolutions, as previously mentioned, with dilation factors r that increase geometrically with network depth at every resolution level. The first two levels have encoder and decoder stacks with two layers each, and the synoptic-scale level is composed of an encoder stack with four layers.

Each encoder layer $l = 0, 1, \ldots$ applies 2D 3×3 dilated convolutions with dilation factor $r = 2^l$ and a PReLU nonlinearity. The decoder layers in the first two levels apply 2D 3×3 convolutions with dilation factors $r = 2^3$ and 2^4, respectively, and are each followed by a PReLU nonlinearity. In addition, we include a nonlinear skip connection at the finest resolution level to easily capture persistence. Downsampling between levels is performed using max-pooling. In the decoder, upsampling is followed by 2D 3×3 convolutions. The number of layers per level was obtained through cross-validation from a small set of architectures that achieved a full receptive field.

All layers at 200^2 km² are composed of 32 convolutional filters, and layers at 400^2- and 800^2-km² resolution apply 64 and 128 filters to their inputs, respectively. The skip connections between encoder and decoder stacks, as well as the upsampling layers, have 32 filters each. In each layer of the network, we use two independent convolutional kernels for each filter: one covering all four equatorial faces of the cubed sphere, and the other covering the polar faces. As shown in section 4,
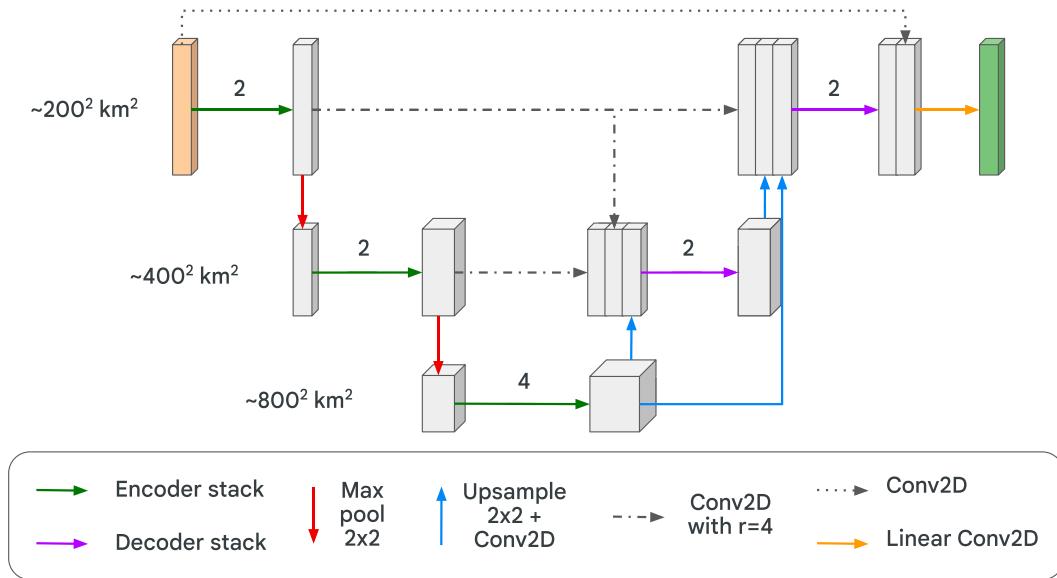
FIG. 2. Neural weather model architecture, modified from a UNet 3+ architecture (Huang et al. 2020). The number of layers of each encoder and decoder stack is as indicated in the schematic. Encoder convolutional layers have geometrically growing dilation factors $r = 2^l$, where $l = 0, 1, \ldots$ is the layer number within the stack from inputs to outputs, and decoder layers have dilation factors $r = 8$ and 16. The layers connecting same-level encoders and decoders have convolutions with 32 filters and dilation factor $r = 4$. All other layers have dilation factor $r = 1$, and all layers have convolutional kernels of size 3×3. Same-level layers implement 32, 64, and 128 filters in the first, second and third levels, respectively.

4. Results

a. Reference models considered

We assess the skill of HeatNet, ExtNet, and GenNet against persistence and the ECMWF S2S forecast system (Vitart et al. 2017). The ECMWF S2S system is an operational model that provides real-time 46-day forecasts 2 times per week. For dates in the test set (2017–21), real-time forecasts used ECMWF's IFS cycles CY43R1, CY46R1, and CY47R2 (ECMWF 2016, 2019, 2020). The S2S system employed 91 vertical levels until May 2021, and 137 levels after that. All versions are coupled to an ocean model at 0.25° resolution with an interactive sea ice model and use a triangular-cubic-octahedral horizontal discretization with 16-km resolution for days 1–15 and 32 km after that (Malardel et al. 2016).

To allow comparison with our deep learning systems, the ECMWF forecasts are bilinearly interpolated to the same resolution as the input ERA5 data (2°), subtracting the same ERA5 mean climatology to produce climatological anomalies. Then, the results are mapped to be cubed sphere using the conservative remapping of Ullrich and Taylor (2015) and Ullrich et al. (2016); all skill metrics are computed on this grid. To assess potential errors due to the spherical harmonics truncation employed by ECMWF's

there is no discernible imprint of the cubed sphere edges on the model forecasts. In the end, the model architecture has about 1.8 million parameters, halfway between the complexity of the models used in Weyn et al. (2020, 2021).

Meteorological Archival and Retrieval System (MARS), we downloaded the forecasts at 2° and 0.25° resolutions, and compared the forecasts after bilinearly interpolating the latter to the 2° grid. The root-mean-square difference between the forecasts is $\sim 10^{-3}$ K, much lower than typical forecast errors.

For all comparisons in this study, we employ both the real-time daily averaged ECMWF control and perturbed ensemble forecasts (Vitart et al. 2017). Model drift is removed from the real-time ECMWF forecasts using 660 reforecasts covering the past 20 years and initialized from ERA5 data (Vitart et al. 2017). Comparisons with the ECMWF control assess the skill of NWMs against a deterministic "best guess" physics-based forecast. The ECMWF S2S ensemble prediction system employs 50 additional ensemble members, perturbing both their initial conditions and model physics to capture forecast uncertainty (Buizza et al. 1999). Operational warning systems typically use perturbed ensembles, which have been shown to yield a higher economic value than high-resolution deterministic forecasts (Richardson 2000; Palmer 2017). For this reason, we include the ECMWF ensemble mean forecast for comparison. Information beyond the first moment of the ensemble statistics is also valuable (Molteni et al. 1996; Zhu et al. 2002; Palmer 2017). However, we limit our comparison to the ensemble mean in this study, since we only consider NWM point forecasts. Even though our models yield a single deterministic output, direct NWM forecasts more closely resemble an ensemble mean prediction than a physical trajectory of the system; this interpretation is supported by the results in sections 4b–f.

Two additional points should be considered when interpreting the relative skill of the ECMWF forecasts. First, the real-time ECMWF system is initialized from the operational IFS analysis, not ERA5, which leads to reduced accuracy at short lead times. Second, the native resolution of the ECMWF system is higher than the resolution of the NWMs. This is both an asset and a liability when evaluating pointwise objective scores; higher resolution reduces structural model errors, but inevitable errors in the timing and location of sharper resolved features can result in lower skill (Mass et al. 2002). Nevertheless, negative impacts of resolution on forecast skill are reduced in our study through the smoothing induced by bilinear interpolation and daily averaging (Accadia et al. 2003).

b. Forecast skill for summer over land

Although we train the NWMs using global data from all seasons, we evaluate here the performance of the forecast systems exclusively for summer over land, where heatwave prediction is most relevant. We define summer as the June–August trimester for the Northern Hemisphere and December–February for the Southern Hemisphere.

To assess model skill during increasingly hot summer days, we evaluate forecasts using two different temperature anomaly percentiles: the 75th (hot) and 95th (extremely hot) percentiles. Setting these thresholds allows assessing the forecast systems as binary classifiers. When evaluating regressive skill, conditioning on the target distribution can confront forecasters with the dilemma of overforecasting a rare event to improve their scores. There is no obvious way to avoid this problem when evaluating the regressive skill of deterministic forecasts at predicting extremes (Lerch et al. 2017). We verified that this dilemma is not a concern for the models we evaluate, after their global bias is subtracted, since they either underpredict extreme anomalies (NWMs), or are well-calibrated (ECMWF control); results conditioned on the union of forecast and target values, which account for false alarms, are included in appendix B.

The regressive skill of the models is characterized in this study through the debiased root-mean-square error (RMSE_d) and the centered anomaly correlation coefficient (AnCC) of standardized temperature anomalies. The RMSE_d is defined as the RMSE of forecasts with respect to targets after removing the global mean bias per lead time of forecasts with respect to targets in the entire test set. We choose to debias the forecasts to prevent forecast bias from positively affecting the skill metric, since the mean target above the temperature thresholds is nonzero (Lerch et al. 2017). (The subtracted bias is shown for all models in Fig. 5 for reference; it is clear that subtracting the bias prevents HeatNet from hedging.)

The centered anomaly correlation coefficient for a given lead time i is defined as (Wulff and Domeisen 2019)

$$\mathrm{AnCC}_i = \frac{\sum_{k=1}^{N_i} (y_{ik} - \overline{y}_i)(y'_{ik} - \overline{y}'_i)}{\sqrt{\sum_{k=1}^{N_i} (y_{ik} - \overline{y}_i)^2 \sum_{k=1}^{N_i} (y'_{ik} - \overline{y}'_i)^2}}, \qquad (3)$$

where $\overline{y}_i \in \mathbb{R}$ is the temporal and spatial average of the target temperature anomaly $\tilde{T}_{2\mathrm{m}}$ at lead time i for summer over land over the entire test set; $y_{ik} \in \mathbb{R}$ are individual values of $\tilde{T}_{2\mathrm{m}}$ at lead time i and at a given location and summer day, indexed by k; and the sums are over the N_i summer targets y_{ik} above the considered standardized temperature anomaly threshold. The forecast counterparts $\overline{y}'_i$ and y'_{ik} are defined similarly based on the forecast temperature anomaly, but the sum over indices k is still conditioned on the anomaly threshold of the targets. The AnCC is a useful metric of the potential of a forecast system, measuring the correlation between the target and forecast outputs (Wilks 2019, chapter 9). Our definition (3) takes into account the dynamic temperature anomalies, filtering out the thermodynamic shift of temperature anomalies over land with respect to the 1979–2019 climatology due to global warming, as well as lead-time dependent model biases. This is not the case for the noncentered anomaly correlation coefficient, which does not filter out forecast biases with respect to climatology (e.g., Weyn et al. 2020).

The classification skill of the models is evaluated through the extremal dependence index (EDI; Ferro and Stephenson 2011) and the equitable threat score (ETS). The EDI is defined as

$$\mathrm{EDI} = \frac{\log F - \log H}{\log F + \log H}, \quad \text{with} \quad H = \frac{a}{a+c} \quad \text{and} \quad F = \frac{b}{b+d}, \qquad (4)$$

where F is the false alarm rate, H is the hit rate, a are the hits, b are the false alarms, c are the misses, and d are the correct negatives (Wilks 2019). Positive values of EDI indicate higher skill than a random forecast. We choose this metric because it is base-rate independent, equitable, and it does not degenerate for rare event classifiers. Thus, the EDI between both thresholds considered can be compared, which is not the case for base-rate dependent measures (Wulff and Domeisen 2019). The equitable threat score is defined as

$$\mathrm{ETS} = \frac{\mathrm{TS} - \mathrm{TS}_{\mathrm{ref}}}{1 - \mathrm{TS}_{\mathrm{ref}}}, \quad \text{with} \quad \mathrm{TS} = \frac{a}{a+b+c}, \qquad (5)$$

where $\mathrm{TS}_{\mathrm{ref}}$ is the threat score of a random forecast and higher ETS is representative of higher skill.

The skill of the different models over land is shown in Fig. 3 for the summers of 2017–21. The lead time is shown in a logarithmic scale to differentiate between three different time scales: the short range (<3 days), the medium range (3–10 days), and the extended or subseasonal range (11–28 days). In the short-range errors are dominated by the initialization, which is more precise for the NWMs, since ERA5 data are fed as predictors. The medium range is characterized by predictable trajectories of the atmospheric state, whereas forecasting a single physical trajectory in the extended range typically adds little value over climatology (Lorenz 1969b). Predictive power in the subseasonal range is associated with slower dynamical modes of the climate system, like the MJO or those arising from ocean–atmosphere interactions (Palmer 1993; Zhou et al. 2019).
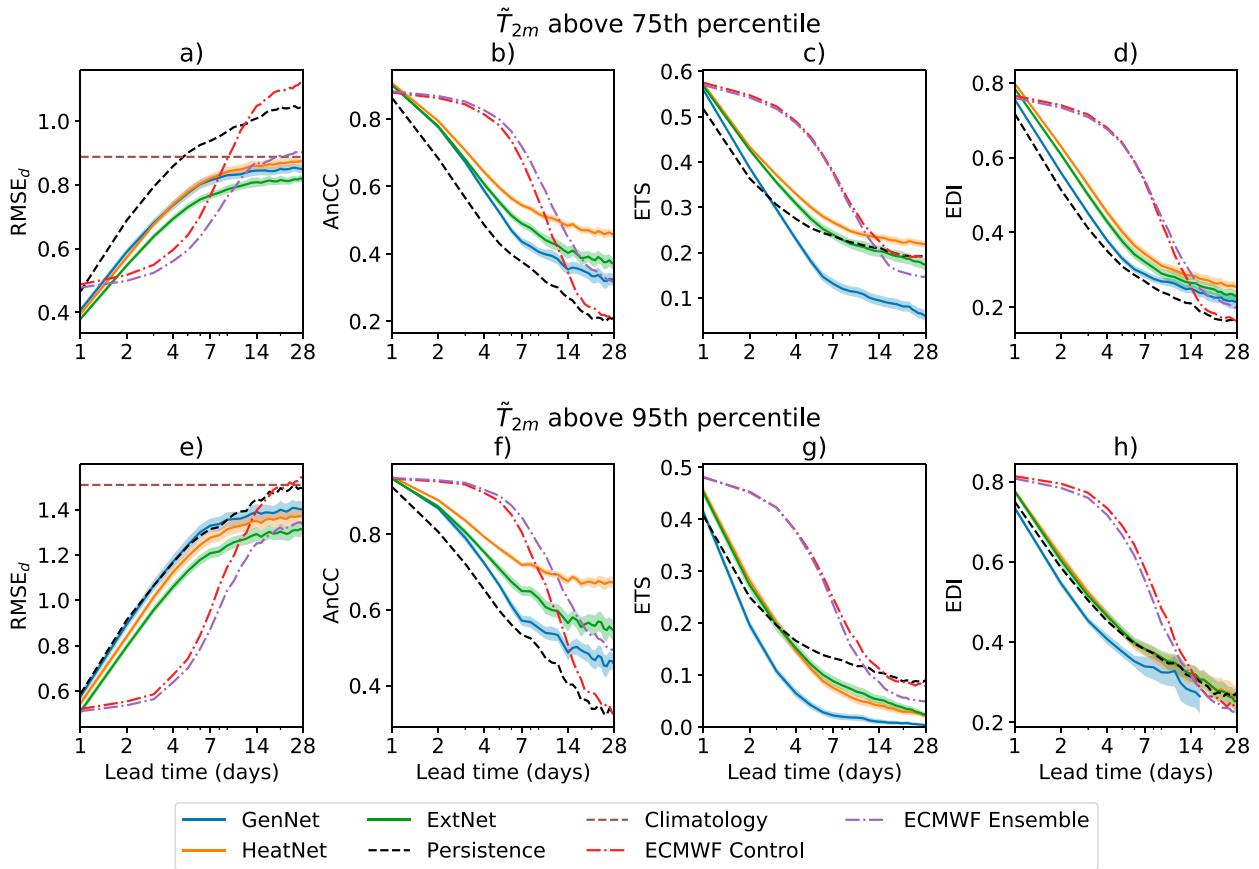
FIG. 3. Forecast metrics for different models during the summer months of 2017–21 and over land. Metrics are shown for forecasts conditioned on target standardized temperature anomalies being above the (a)–(d) 75th and (e)–(h) 95th percentiles. Shown are the (left) debiased root mean square error (RMSE$_d$), (left center) centered anomaly correlation coefficient (AnCC), (right center) equitable threat score (ETS) and (right) extremal dependence index (EDI). Uncertainty bands, shown for the NWMs as a reference, represent 1 standard deviation. Results are only shown for metrics with a robust uncertainty estimate; details may be found in appendix A.

The extreme-focused HeatNet and ExtNet outperform GenNet for both temperature thresholds and all metrics considered, highlighting the usefulness of the exponential loss (2) in the extreme prediction task. All NWMs maintain a higher anomaly correlation with the targets than persistence, but only the models trained on the exponential loss improve upon persistence in a mean-square-error sense during extremely hot days ($\tilde{T}_{2m} > 95$th percentile). HeatNet, which is trained to emphasize positive extremes exclusively, yields forecasts with higher AnCC than the symmetric ExtNet during hot summer days. Although the AnCC difference indicates higher predictive potential relative to ExtNet in this task, its one-sided emphasis on heatwaves leads to a significant positive bias, as shown in Figs. 4 and 5. This bias is detrimental to the prediction of dynamic temperature anomalies, increasing the RMSE$_d$, and does not lead to significant classification skill improvements over ExtNet (Figs. 3g,h).

The skill of all models is comparable for day-ahead forecasting. In the medium range, the control and ensemble ECMWF forecasts remain superior, but their skill drops significantly faster than that of the NWMs beyond the first week. After the

second week, the extreme-focused NWMs have higher regressive skill than the physics-based models. The RMSE$_d$ skill of NWMs relative to the ECMWF system is significantly higher when considering all hot days ($\tilde{T}_{2m} > 75$th percentile) than during extremely hot days ($\tilde{T}_{2m} > 95$th percentile). Positive skill in the extended range is enabled by the use of the exponential loss, without which the NWM forecasts cannot improve upon the AnCC of the operational ensemble system.

The ECMWF ensemble mean forecast substantially improves upon the regressive skill of the control run in the extended range, with RMSE$_d$ and AnCC metrics closer to ExtNet. However, higher regressive skill in the extended range does not translate into higher classification skill, as shown by the ETS and EDI diagnostics. As classifiers, HeatNet and ExtNet have slightly higher skill than the ECMWF models beyond the medium range for hot days ($\tilde{T}_{2m} > 75$th percentile). During extremely hot days ($\tilde{T}_{2m} > 95$th percentile), all NWMs fail to improve upon the ECMWF system as classifiers, although the use of the exponential loss significantly increases the skill of the NWMs for all regression and classification metrics considered.
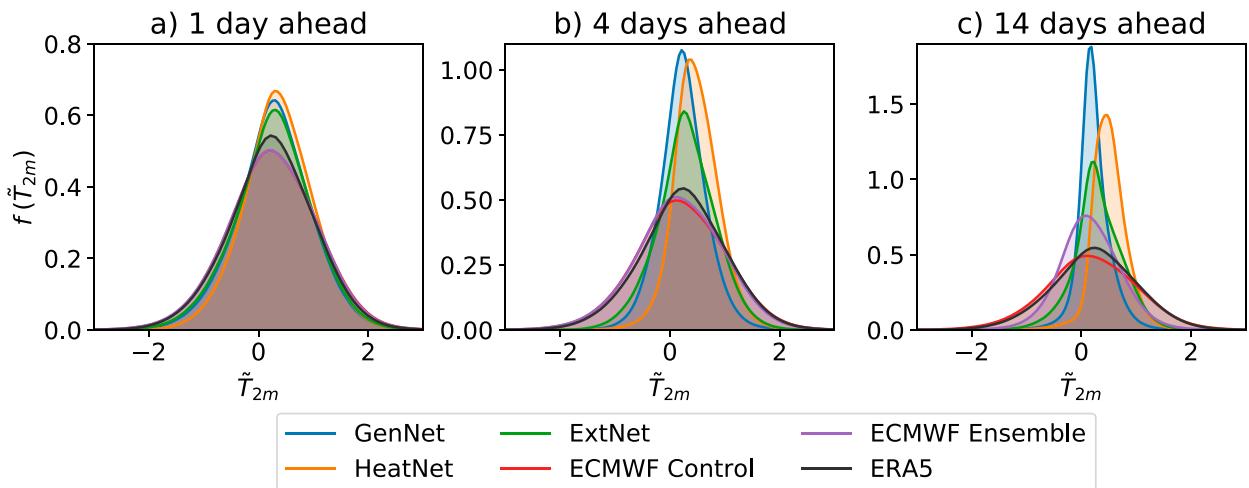
FIG. 4. PDFs [here $f()$] of forecast global standardized temperature anomalies during the period 2017–21. Results are shown for all NWMs, the control and ensemble forecasts from the ECMWF S2S system, and the true target distribution (ERA5). Note that the PDFs are not centered about zero, indicating a prediction of the shift in climatology from the 1979–2019 mean.

We also analyzed interhemispheric differences in skill for summer over land using the same thresholds and found that all models have higher skill in the Northern Hemisphere. The interhemispheric contrast is higher for the NWMs than for the physics-based models; results are included in the online supplemental material.

c. Smoothing of forecasts with lead time

The contrast between regressive and classifying skill of NWMs and the ECMWF ensemble is due to a smoothing of their forecasts as lead time progresses, and the predictability of the targets diminishes. Here, we define smoothing as loss of sharpness, or loss of ability to predict events far from climatology. This smoothing is illustrated in Fig. 4 through the evolution of the forecast probability density functions (PDFs)

with lead time for all models considered. Smoothing leads to a density concentration near the mean, as the probability of strong temperature anomalies decreases.

In the case of the ECMWF ensemble, lower predictability reduces the correlation between individual forecasts with lead time. This leads to a variance reduction in the ensemble mean distribution. Smoothing is also typical of data-driven methods, although in this case it is the result of forecast error minimization under uncertainty (e.g., Sønderby et al. 2020). While the PDF of individual (hindcast corrected) physics-based forecasts remains relatively constant, data-driven forecasts shift toward distributions closer to the target mean, with fewer extreme events.

Notably, this smoothing is slowed down through the use of the exponential loss (2), particularly in ExtNet. The use of the
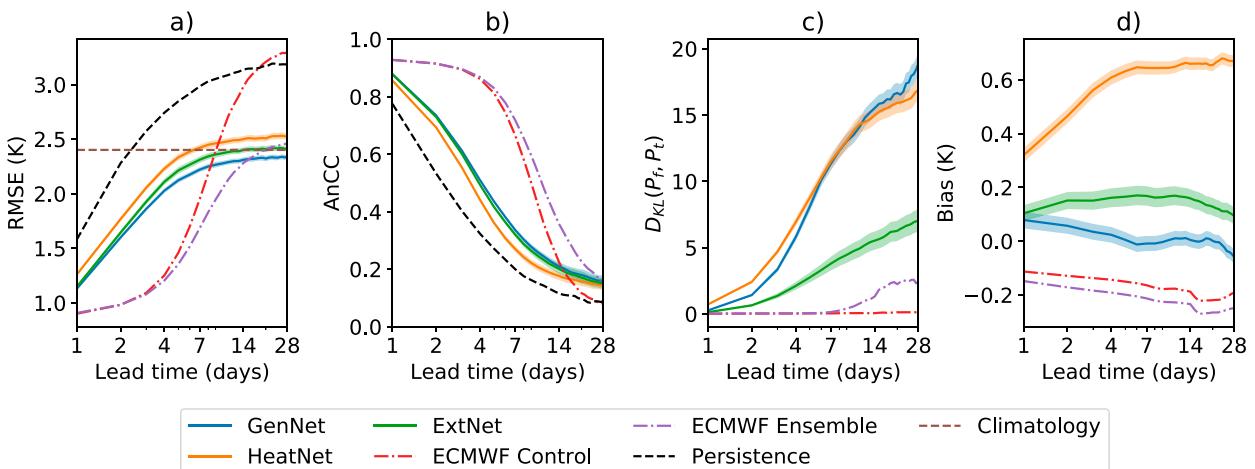

FIG. 5. Forecast metrics for the general 2-m temperature T_{2m} prediction task, for different models as a function of lead time, showing the (a) RMSE, (b) AnCC, (c) Kullback–Leibler divergence D_{KL} of $\tilde{T}_{2m}$ with respect to the targets, and (d) unconditional T_{2m} bias. All results are global and temporal averages over the period 2017–21. NWM names follow section 2b. Uncertainty is defined as in Fig. 3.

symmetric exponential loss increases the probability of significant deviations from climatology: ExtNet forecasts deviations above the 95th target percentile 14 days ahead 4.5 times more frequently than GenNet, and only 25% less frequently than the ECMWF ensemble mean. Minimizing the positive exponential loss also reduces forecast smoothing, but it leads to a positive bias and makes HeatNet forecasts of negative anomalies extremely unlikely. The deviation of the forecast distribution from the true target PDF is further quantified in Fig. 5 through the Kullback–Leibler (KL) divergence, which is an information-based measure of the difference between probability distributions (Kullback and Leibler 1951; Joyce 2011). The use of the symmetric exponential loss reduces the divergence of ExtNet to less than half of the GenNet divergence for all lead times, whereas the bias induced by the positive extreme loss results in a similar KL divergence when compared with GenNet.

Although ExtNet does not manage to capture the same sharpness as the ECMWF ensemble, it is closer in KL divergence to it than to the MSE-trained GenNet model, highlighting the effectiveness of the exponential loss (2) in retaining forecast sharpness for a given architecture. Interestingly, the difference in probability of strong positive anomaly forecasts between ExtNet and the ECMWF ensemble mean in the extended range is significantly smaller than the difference in negative anomaly probabilities (Fig. 4c), even though the loss used to train ExtNet is symmetric. This suggests that positive anomalies are easier to capture than negative anomalies given our predictors.

d. Global surface temperature prediction skill

To further assess the effect of the exponential loss (2) on the general temperature prediction problem, we include in Fig. 5 the RMSE and AnCC of T_{2m} (i.e., not standardized) for all dates in the test set, over both land and oceans. Note that the RMSE in this case is not debiased. Remarkably, ExtNet shows a very small reduction in forecast skill in the general temperature prediction problem with respect to GenNet. All NWMs beat persistence for all lead times and remain skillful with respect to the ECMWF control beyond the medium range; the ECMWF ensemble mean remains the most skillful model in the general prediction task. Although the RMSE of ExtNet forecasts converges to that of climatology after 3 weeks, the model can forecast strong deviations from climatology (as shown in Fig. 7 for an individual forecast at 23 days of lead time). Finally, the forecast biases of GenNet and ExtNet are similar in magnitude to those of the ECMWF model (Fig. 5d), even though the neural weather predictions are not bias-corrected by reforecasts. HeatNet does suffer from a significant positive bias, which explains its loss of skill with respect to ExtNet. From Figs. 3 to 5, it is evident that ExtNet provides the best compromise between extreme heat forecasting skill, forecast reliability and general prediction accuracy among the NWMs considered.

Figure 5a also allows comparison with other NWMs in the literature. Weyn et al. (2021) use a neural network with a simpler albeit similar architecture as a time integrator, forecasting fields 6 and 12 h into the future with each inference step. They show that only when taking an ensemble mean of such models can they beat the RMSE of the ECMWF S2S control forecast in the extended range. In contrast, producing all lead time predictions at once, a single ExtNet forecast is able to improve upon the ECMWF control forecast both in RMSE and AnCC in the extended range. This is consistent with studies comparing direct and iterative forecasting using NWMs, which show that the former configuration leads to enhanced regressive skill (Rasp et al. 2020). The similarities between the ensemble forecast of Weyn et al. (2021), the ECMWF ensemble, and our results, suggest that NWMs outputting longer lead times yield forecasts more similar to the ensemble mean of physics-driven forecasts than to a given physical trajectory. The similarities include the smoothing of forecasts with lead time, and the saturation of the RMSE in the extended range around the climatological error.

The results in Figs. 3–5 yield important insights into the questions that we posed in the introduction. NWMs trained on limited historical data can improve upon persistence in the prediction of out-of-sample rare events, in a regressive sense. As classifiers of extreme events, they only remain skillful with respect to persistence in the short range, due to their loss of sharpness with lead time. For our chosen architecture, positive regressive and classifier skill can only be achieved for extreme events when employing the exponential loss (2). Furthermore, training on the symmetric exponential loss, ExtNet is able to reduce the prediction error for extreme events and slow down the distributional shift with lead time, all while maintaining an unconditional regressive skill practically indistinguishable from models trained on the MSE. The extreme-focused models improve upon the ECMWF models in the prediction of rate events in a regressive sense after 2 weeks; in the medium range the physics-based models remain vastly superior. We now explore two specific heatwave events as forecast by the ECMWF model and the NWMs to illustrate the implications of these results.

e. Analysis of the 2017 western European heatwave

Sections 4b–d highlight the different ways in which uncertainty affects physics-based and NWM forecasts. These differences are further explored here at the regional scale by considering the western European heatwave of June 2017. The 2017 heatwave resulted in the hottest June on record in Spain and the Netherlands, and the second warmest in France and Switzerland. It was associated with northward warm air intrusions fostered by a subtropical ridge over western Europe, as shown by Sánchez-Benítez et al. (2018).

Forecasts of the standardized temperature anomaly on 20 June 2017 are shown in Fig. 6 for several lead times. The ECMWF S2S forecasts initialized 5 days prior accurately predicted the spatial anomaly patterns over western Europe, slightly overpredicting their magnitude over coastal regions and Morocco. In contrast, the control forecast initialized 15 days prior projected important negative temperature anomalies over most of western Europe, the opposite of what was observed. It also failed to predict the warm air intrusion from the Saharan coast. Only about 10 of the 50 ECMWF perturbed ensemble members predicted warm temperature anomalies over France
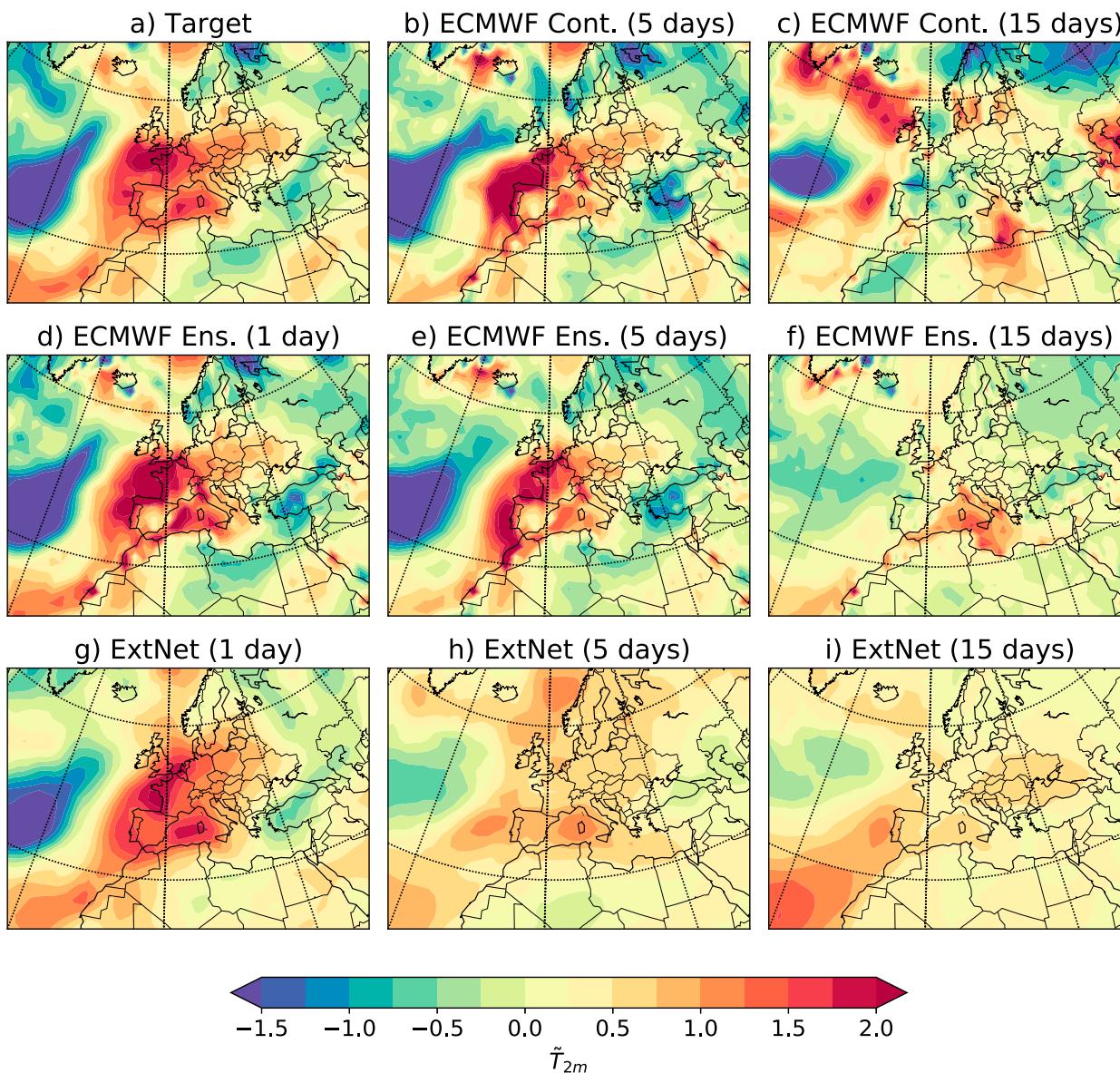
FIG. 6. Daily averaged standardized 2-m temperature anomaly over Europe on 20 Jun 2017 from (a) the ERA5 reanalysis product and as forecast by (b),(c) the ECMWF S2S control (ECMWF Cont.), (d)–(f) the ECMWF perturbed ensemble mean (ECMWF Ens.), and (g)–(i) ExtNet. The lead time of the forecasts is given in parentheses in the title of each panel.

and Spain, and a higher fraction predicted negative anomalies; forecasts from 22 of these members are shown in appendix C for reference. As a result, the ensemble mean forecast was close to climatology outside the Mediterranean Sea.

On the other hand, ExtNet robustly forecast the warm air intrusion for the same lead time (15 days), but not its northward penetration into France and the "Benelux" countries. At 5 days of lead time, ExtNet predicted positive temperature anomalies over Europe, although the forecast was too mild and inferior to the ECMWF forecasts. Overall, the NWM forecasts track well both the magnitude and patterns of temperature anomaly in the short range. In the medium and extended ranges, the forecasts match the temperature anomaly

patterns well, but underestimate their magnitude. Figure 6 is consistent with our hypothesis that, contrary to physics-based models, forecasts by direct NWMs do not represent trajectories of the system. They are more closely related to the mean projection of an ensemble of physics-based forecasts, or NWMs acting as time integrators (Slingo and Palmer 2011; Weyn et al. 2021; Scher and Messori 2021).

f. Analysis of the 2021 western North American heatwave

To showcase the benefits of the symmetric exponential loss, we compare forecasts of the 2021 western North American (WNA) heatwave provided by ExtNet, GenNet, and the ECMWF ensemble in Fig. 7. We consider the WNA heatwave
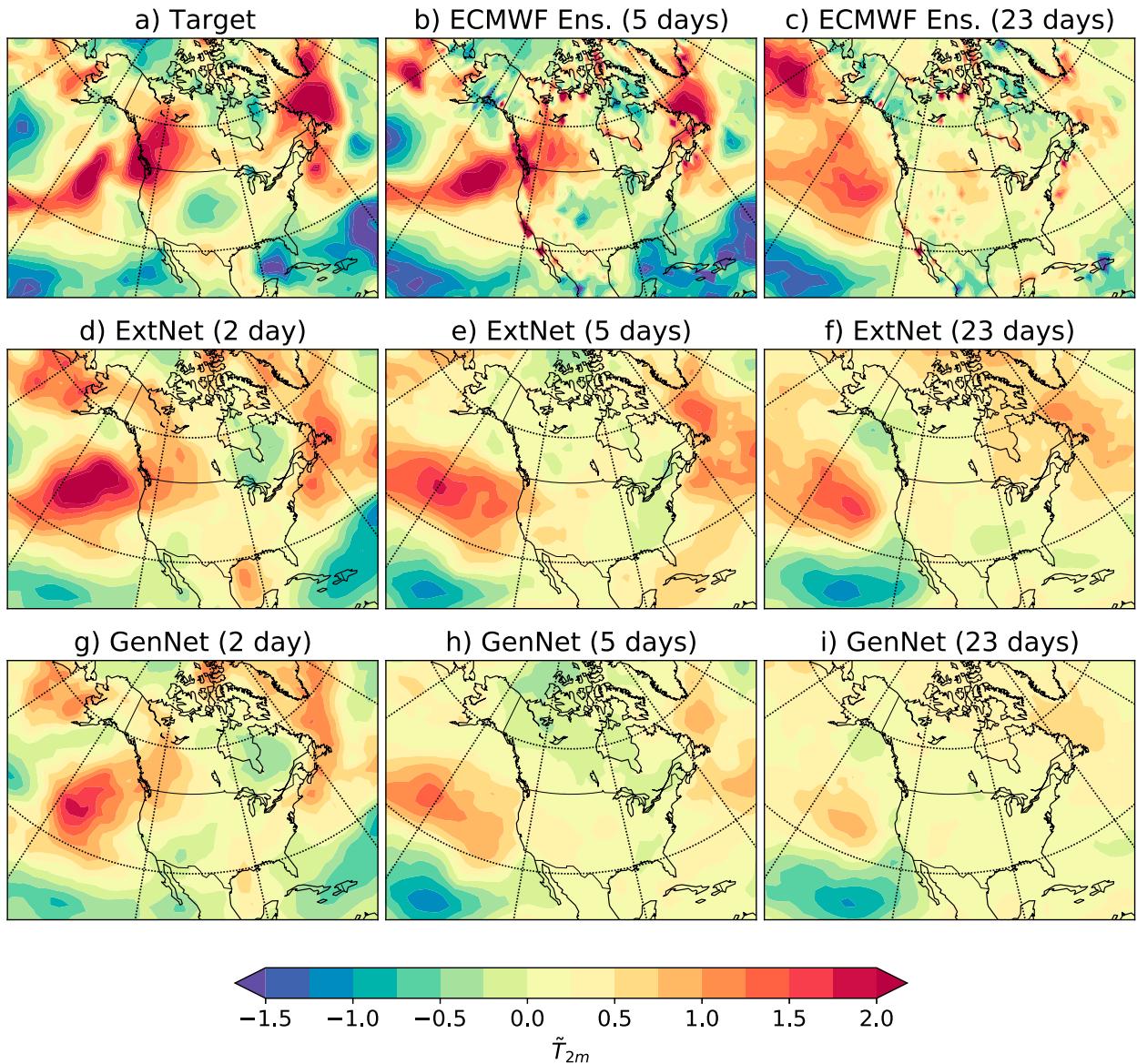
FIG. 7. Daily averaged standardized 2-m temperature anomaly over North America on 26 Jun 2021 from (a) the ERA5 reanalysis product and as forecast by (b),(c) ECMWF Ens.; (d)–(f) ExtNet; and (g)–(i) GenNet. The lead time of the forecasts is given in parentheses in the title of each panel.

because its forecast using operational systems is well characterized in the literature (Lin et al. 2022). Several phenomena have been suggested as causes of the WNA heatwave. Lin et al. (2022) note the eastward propagation of a Rossby wave train from the tropical western Pacific that may have favored the formation of a heat dome over western North America. Mo et al. (2022) and Lin et al. (2022) also show that the heatwave was preceded by a strong atmospheric river transporting warm moist air from Southeast Asia into the region.

We focus on the heatwave onset, which took place 25–26 June 2021. The actual temperature anomaly on 26 June was characterized by heatwave conditions over Washington, Oregon, and British Columbia (Canada). Extreme temperature anomalies

were also observed over the northeastern Pacific and the Labrador Sea (Fig. 7a). The ECMWF ensemble forecast from 21 June correctly predicted warm temperature anomalies over western North America. The forecast from 3 June, more than 3 weeks ahead, failed to predict positive anomalies over western North America or the Labrador Sea. This loss of predictive skill over land has been linked to the inability to forecast both the continental penetration of the atmospheric river (Mo et al. 2022), and the eastward shift of the atmospheric ridge over western Canada (Lin et al. 2022).

ExtNet forecast the anomaly pattern correctly with 2 days of lead time, but only predicted significant positive anomalies over Washington, Oregon, and the Labrador Sea in the forecast

5 days prior. Relative to GenNet, ExtNet provides significantly better and sharper forecasts for all lead times considered, confirming the results in sections 4b–d. GenNet underpredicted the extent of the heatwave over North America even in the short range, and failed to predict continental penetration 5 days prior to the event. At 23 days of lead time, the ExtNet forecast closely resembles that of the ECMWF ensemble, exhibiting an anomaly dipole over the eastern Pacific (Figs. 7c,f). The data-driven model exhibits better correlation with the target over the Labrador and Bering Seas at this lead time, highlighting the skill of the model in the extended range.

5. Model interpretation

The NWMs presented here leverage a wider range of predictors than other extreme heat forecasting systems in the literature (Chattopadhyay et al. 2020; Jacques-Dumas et al. 2022). Here we assess the importance of the additional input fields using integrated gradients for feature attribution (Sundararajan et al. 2017; Sundararajan and Agrawal 2021).

a. Feature attribution through integrated gradients

The attribution for each feature $\mathbf{x}^{(i)}$ is defined as the mean absolute value of its contribution to the model forecast $\mathbf{y}' = \mathbf{\Psi}(\boldsymbol{\theta}^*, \mathbf{x})$ with respect to a null-contribution baseline forecast $\mathbf{y}'_b = \mathbf{\Psi}(\boldsymbol{\theta}^*, \mathbf{x}_b)$. We formulate the baseline input $\mathbf{x}_b$ to be the feature vector on the linear path between $\mathbf{x}_{min}$ and $\mathbf{x}_{max}$ that results in a forecast closest to global climatology ($\mathbf{y}' \approx 0$), where $\mathbf{x}_{min}$ and $\mathbf{x}_{max}$ are feature vectors constructed using the minimum and maximum values of the features found in the evaluation set, respectively.

For each actual forecast $\mathbf{y}'_a = \mathbf{\Psi}(\boldsymbol{\theta}^*, \mathbf{x}_a)$, the contribution is computed as the partial derivative of the forecast with respect to $\mathbf{x}^{(i)}$, integrated along a linear path from the baseline $\mathbf{x}_b = [\mathbf{x}_b^{(1)}, \mathbf{x}_b^{(2)}, ...]^T$ to the actual input value $\mathbf{x}_a = [\mathbf{x}_a^{(1)}, \mathbf{x}_a^{(2)}, ...]^T$,

$$\text{Att}[\mathbf{x}^{(i)}]|_{\mathbf{y}'_a} = \frac{1}{N} \left\| \int_0^1 \frac{\delta \mathbf{\Psi}[\boldsymbol{\theta}^* \mathbf{x}(\alpha)]}{\delta \mathbf{x}^{(i)}} d\alpha \, [\mathbf{x}_a^{(i)} - \mathbf{x}_b^{(i)}] \right\|_1, \quad (6)$$

where N is the number of pixels over which the L1 norm is computed, $\delta \mathbf{\Psi}/\delta \mathbf{x}^{(i)}$ is a discretized approximation of the partial derivative, and $\alpha \in [0, 1]$ parameterizes the linear path from baseline to actual feature values, such that $\mathbf{x}(\alpha = 0) = \mathbf{x}_b$ and $\mathbf{x}(\alpha = 1) = \mathbf{x}_a$. Last, we compute the mean attribution for each feature over dates in the evaluation set.

b. Relevance of model inputs

We apply the integrated gradients methodology described in section 5a to ExtNet forecasts for summer over land during the 5-yr period from 2015 to 2019. Feature attributions are shown in Fig. 8 for the extreme and the general prediction task, and for lead times spanning the short, medium and extended range. The contributions from the most recent data, data from the previous 2 days, and data from the first 4 days of the week preceding the forecast are shown in different colors to quantify the relevance of past history as a predictor of future states.

We find that $\tilde{T}_{2m}$ itself is the most important feature in all cases, while the relative relevance of other features increases with forecast lead time and depends strongly on the task considered (i.e., extreme or general prediction). In addition, the relevance of feature history robustly increases with lead time, suggesting that NWM forecasts in the extended range learn to rely on lower frequency signals. These shifts in relevance with lead time may be used to optimize the NWM architecture, for instance by pruning connections between short lead time forecasts and predictors more than a few days old.

In the extreme prediction task, the temperature anomaly, 700-hPa geopotential height anomaly, topography, and OLR are important predictors at all lead times. Soil moisture below 7 cm gains relevance with lead time, consistent with the characteristic low frequency of land–atmosphere coupling processes and the memory of root-zone soil moisture (Wu et al. 2002). Additional relevant predictors include the potential vorticity at 500 hPa, and the geopotential height anomaly at 300 and 500 hPa. The potential vorticity at 700 hPa and the surface soil moisture above 7 cm are irrelevant to ExtNet predictions at all time scales.

The most important predictors for extreme prediction are also the most important ones for the general forecasting task. However, the relative importance of the temperature anomaly is significantly greater in the general problem, dominating the total attribution. Soil moisture plays a much smaller role in this task relative to heatwave prediction, consistent with observations of a much stronger land–atmosphere coupling under extreme conditions (Orth and Seneviratne 2012; Benson and Dirmeyer 2021). Overall, Fig. 8 suggests that generic estimates of the relevance of alternative features for NWM forecasting may underestimate the contribution of such predictors to extreme event forecasting. For the auxiliary features, we find topography to be the most relevant predictor, followed by the land–sea mask. Finally, the global attribution decreases with lead time, consistent with the progressive loss of predictive information and forecast sharpness.

6. Discussion

Referencing the first question that we posed in the introduction, we find that deep learning systems trained on limited historical data can forecast out-of-sample extreme heat events with positive regressive skill above persistence for lead times between 1 and 28 days. This is remarkable given the length of the reanalysis record, and potentially indicative of the ability of regression-based neural weather models to learn about causal physical mechanisms that are common to both the extreme and general forecasting tasks. The rare nature of heatwaves implies that this learning process occurs in the low data regime, and that improved models may be obtained through data augmentation techniques (Miloshevich et al. 2022). In this context, an interesting research direction would be to train deep learning models using a much larger synthetic dataset as a first step (Chattopadhyay et al. 2020; Jacques-Dumas et al. 2022), and then leverage reanalysis products like ERA5 to fine-tune the model through transfer learning. This technique
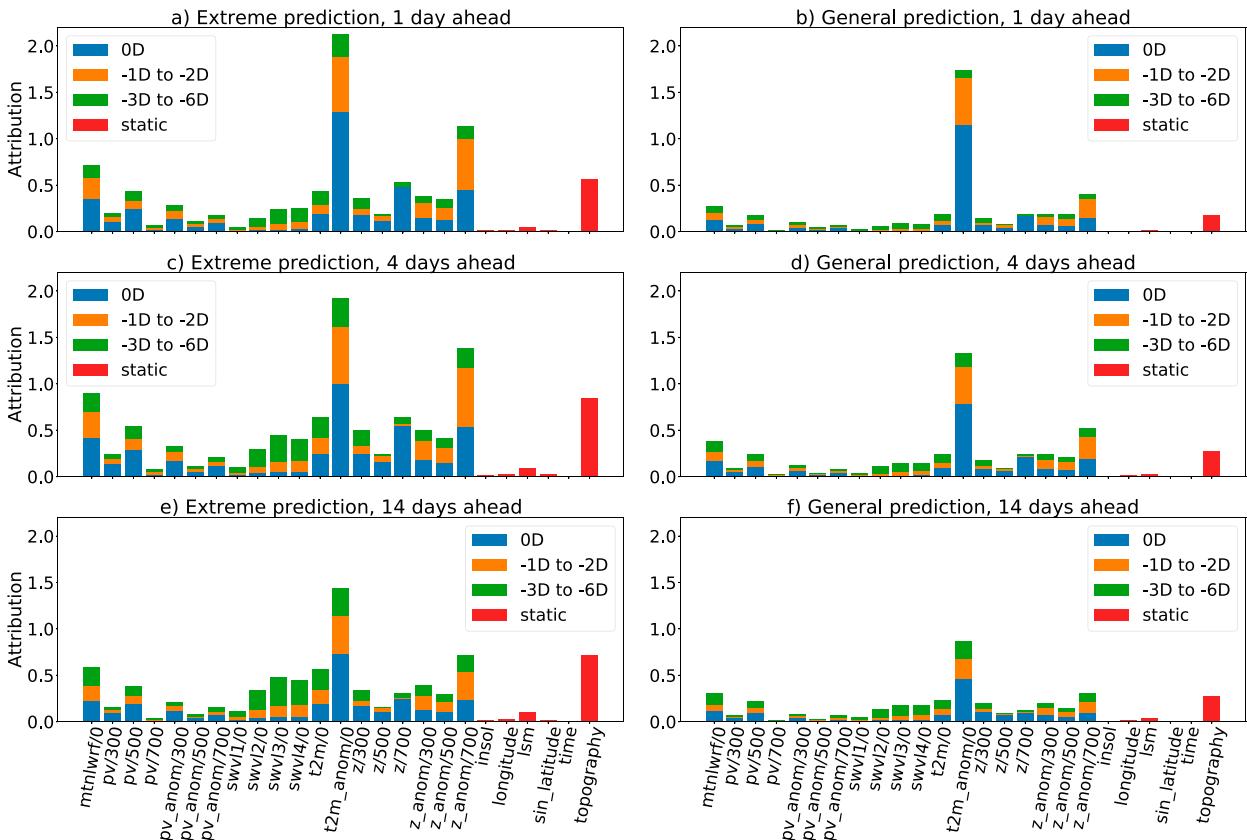
FIG. 8. Forecast feature attribution for summer over land, using data spanning the period 2015–19. Attribution is shown for (a),(c),(e) extreme ($\tilde{T}_{2m} > 2$) and (b),(d),(f) general prediction, and for different lead times. Feature names follow the notation *name/pressure level*, where each *name* follows ECMWF notation: mtnlwrf is outgoing longwave radiation, pv is potential vorticity, *z* is geopotential, swvli is soil water volume at level *i* below the ground, and t2m is temperature 2 m above the surface. Pressure level 0 represents the land surface. The legend label $-i$D defines the contribution of data *i* days before forecast initialization.

has already resulted in remarkable achievements in other fields of science, such as organic synthesis (Pesciullesi et al. 2020).

For the second question, we find that NWMs trained on the mean-square-error loss fail to yield skillful forecasts of extremely hot days at any lead time considered, at least with our architecture and only using historical data. Our results suggest that it is crucial to train models using losses that emphasize extremes to achieve positive skill in this task, which has been shown before for idealized dynamical systems (Qi and Majda 2020). Moreover, the switch to the proposed symmetric exponential loss results in negligible skill loss in the general temperature prediction problem and yields more reliable and sharper forecast distributions farther into the future. Thus, the answer to our third question, whether NWMs trained to predict extremes retain skill in more general settings, is positive.

Our best neural weather model (ExtNet) compares favorably to the ECMWF S2S control forecast in the subseasonal range, yielding lower errors and higher correlations with the target both in the general and extreme heat prediction tasks. In the medium range, the ECMWF model remains the most powerful forecast system. The ECMWF ensemble pushes the dominance of physics-based forecasts to longer lead times, but

even then ExtNet retains regressive skill in the extreme prediction task after 2 weeks. This, however, does not fully translate into higher skill as a binary classifier due to the smoothing of forecasts with lead time, which also results in reduced effective resolution. Although the symmetric exponential loss reduces the distributional shift of the forecasts, additional modifications to NWMs, such as the use of generative modeling (Kingma and Welling 2013; Rezende and Mohamed 2015), may be necessary to further increase forecast sharpness beyond the short range. This requirement is particularly important for the prediction of extremes. In addition, many practical applications require higher-resolution forecasts than those provided by the neural weather models analyzed here. Higher sharpness and effective resolution at long lead times are some of the specifications that neural weather models will need to meet before they can be used to produce actionable information; we expect the results in this paper to inform the design of such models.

Operational warning systems achieve maximum economic value when they can represent the space of possible trajectories as a probability density function, such that the occurrence of extreme events can be treated probabilistically, not as a binary
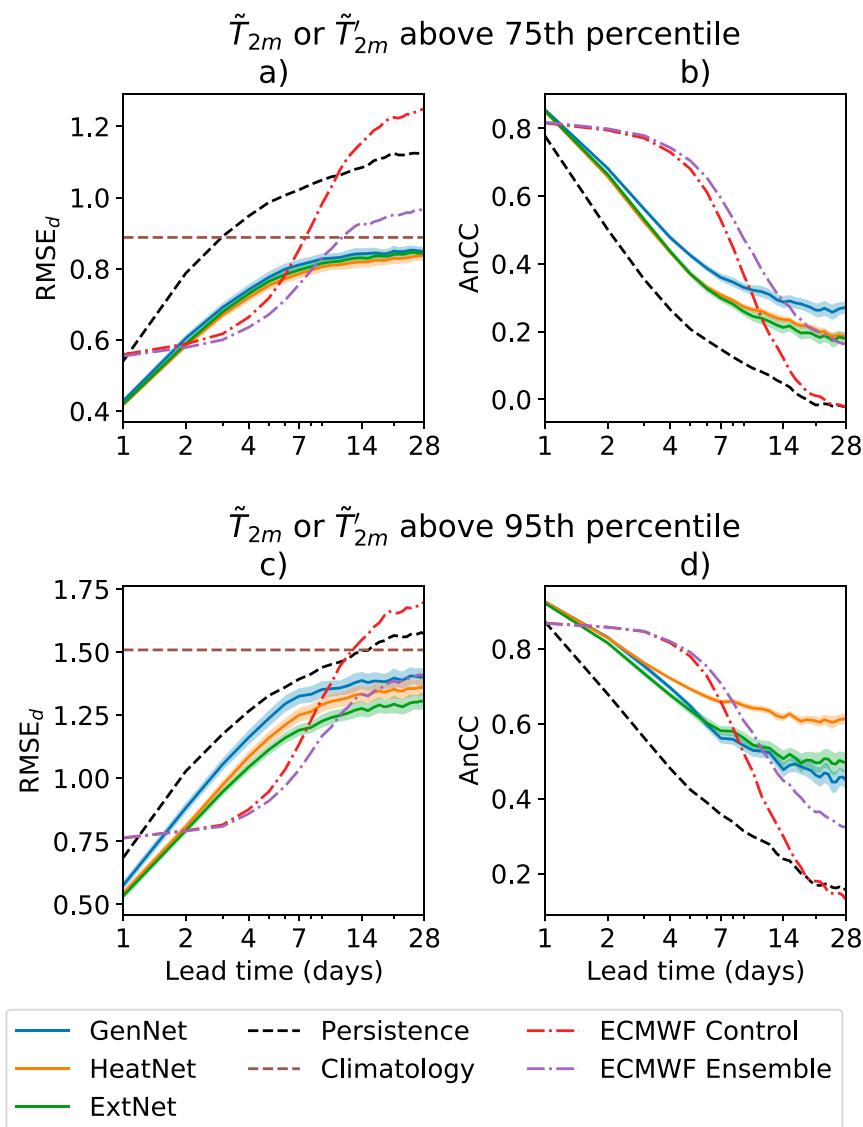
FIG. B1. Forecast metrics for different models during the summer months of 2017–21 and over land. Metrics are shown for forecasts conditioned on either target or forecast standardized temperature anomalies being above the (a),(b) 75th and (c),(d) 95th percentiles of values in the test set. The legend is as in Fig. 3.

problem (Palmer 2017). This is done in practice through the use of perturbed ensembles. The use of perturbed ensembles has recently been explored for NWMs acting as time integrators (Scher and Messori 2021), which still show a moderate distributional shift with lead time (Weyn et al. 2021). The use of our proposed exponential loss may enable the use of longer time steps in iterative NWMs while preserving forecast sharpness.

An alternative avenue of research that may prove fruitful is the direct prediction of the probability distribution of trajectories (Sønderby et al. 2020), or some parametric approximation of it. In the context of climate modeling, Guillaumin and Zanna (2021) trained a convolutional neural network to predict the

mean and standard deviation of subgrid-scale momentum fluxes in the ocean, which they parameterized as Gaussian. Similar approaches could be taken to predict the ensemble distribution in temperature anomaly projections, retaining the regressive skill of direct NWM forecasts while correcting their underdispersion and smoothness. We hope that these or other methodologies, combined with the use of extreme-focused loss functions such as the one we propose, can enable reliable, actionable and efficient forecasting of extreme events using neural weather models in the near future.
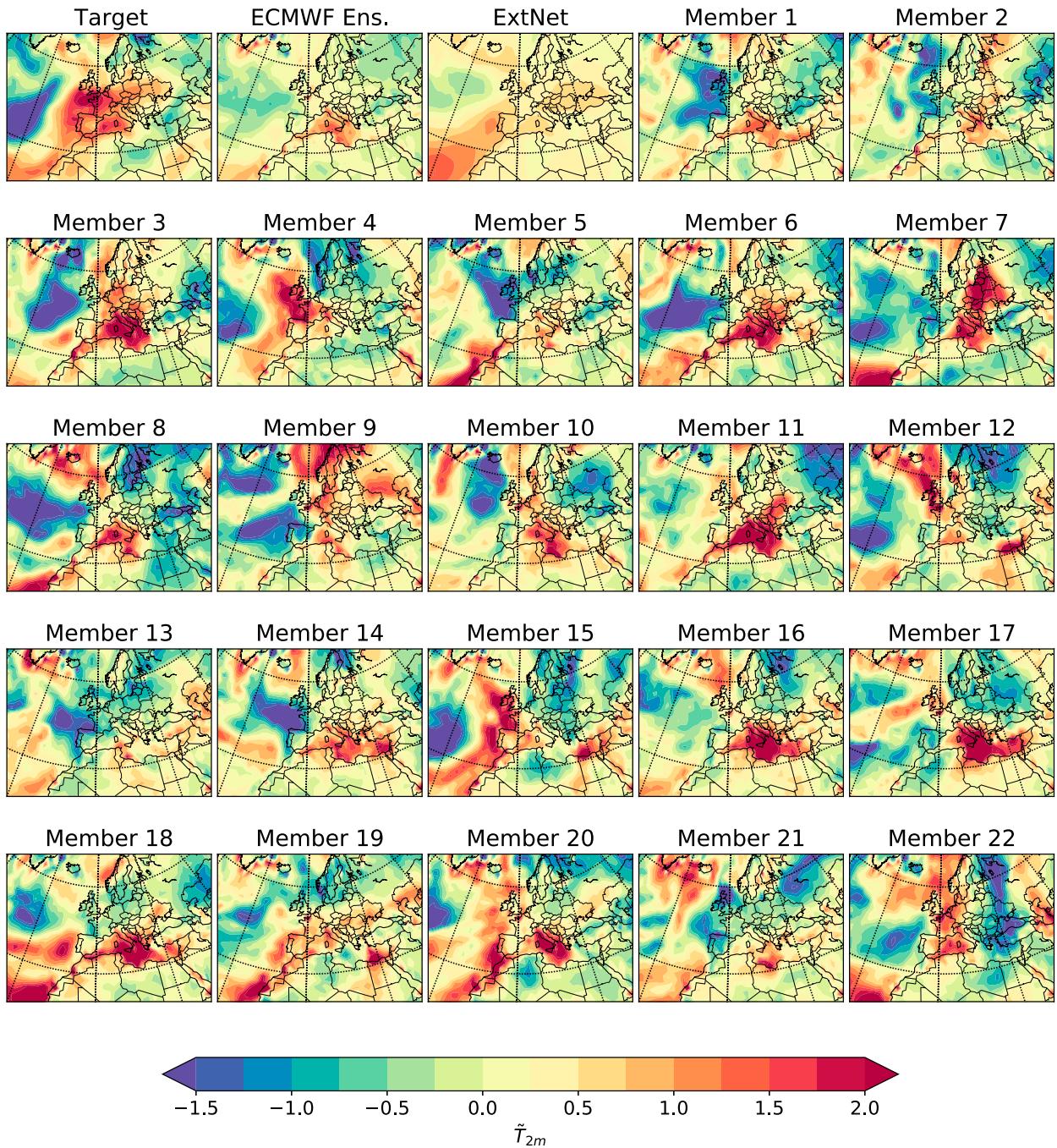
FIG. C1. Daily-averaged standardized 2-m temperature anomaly over Europe on 20 Jun 2017 from the ERA5 reanalysis product ("Target"), the ECMWF Ens., the ExtNet forecast, and the forecasts by 22 individual members of the operational ECMWF perturbed ensemble. The lead time for all forecasts is 15 days.

discussions that helped to improve this paper, as well as Peter Düben and two anonymous reviewers for insightful comments on an earlier version of this work. The authors also acknowledge the use of the DLWP-CS open source package developed by Jonathan Weyn as a starting point for this project.

Data availability statement. The ERA5 reanalysis data are freely available at the Copernicus Climate Data Store (https://cds.climate.copernicus.eu). The ECMWF S2S control and perturbed forecasts can be obtained from the Meteorological Archival and Retrieval System of ECMWF (https://apps.ecmwf.int/datasets/data/s2s). The software used to train the

neural weather models is available on GitHub (https://github.com/google-research/heatnet).

APPENDIX A

Metric Uncertainty Estimation

The test set used in this article contains about 7 million samples, >170 000 samples of hot days over land ($\tilde{T}_{2m} > 75$th percentile), and >34 000 samples of extremely hot days over land ($\tilde{T}_{2m} > 95$th percentile). To determine the variance of the sample mean due to finite sample size, we use block bootstrapping with yearlong disjoint blocks (Hall et al. 1995). We construct the empirical distribution function of the sample mean from 10^6 bootstrap samples and use its standard deviation as a measure of uncertainty. All uncertainty estimates proved robust to block size reduction except for EDI during extremely hot days and after a certain lead time for GenNet. Results are omitted for these EDI estimates.

APPENDIX B

Regressive Skill Conditioned on Target and Forecast Values

To verify that the evaluated models do not suffer from the forecaster's dilemma (Lerch et al. 2017), the debiased RMSE and the centered anomaly correlation coefficient are evaluated here over all dates and locations where either the target or the forecast temperature anomalies were above a certain percentile of values in the test set. This conditioning assesses the skill over false alarms, as well as over hits and misses, penalizing models that overforecast extremes. As shown in Fig. B1, differences with respect to Fig. 3 are most prominent for the ECMWF and persistence forecasts, which are well calibrated. The skill reduction is smaller for NWMs, which tend to underpredict extremes, and insignificant for GenNet. This pushes the threshold above which ExtNet improves upon GenNet to a higher percentile. Above the 95th percentile, ExtNet still outperforms GenNet.

APPENDIX C

Stamp Plot of the 2017 European Heatwave from the ECMWF Ensemble

Figure C1 shows a stamp plot of 22 random individual 15-day forecasts of the heatwave described in section 4e, from the ECMWF ensemble. Several members (e.g., 4, 15, and 20) capture elements of the heatwave, but many others show similar shortcomings to the control forecast.

REFERENCES

Accadia, C., S. Mariani, M. Casaioli, A. Lavagnini, and A. Speranza, 2003: Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Wea. Forecasting*,

18, 918–932, https://doi.org/10.1175/1520-0434(2003)018<0918:SOPFSS>2.0.CO;2.

Benson, D. O., and P. A. Dirmeyer, 2021: Characterizing the relationship between temperature and soil moisture extremes and their role in the exacerbation of heat waves over the contiguous United States. *J. Climate*, **34**, 2175–2187, https://doi.org/10.1175/JCLI-D-20-0440.1.

Brás, T. A., J. Seixas, N. Carvalhais, and J. Jägermeyr, 2021: Severity of drought and heatwave crop losses tripled over the last five decades in Europe. *Environ. Res. Lett.*, **16**, 065012, https://doi.org/10.1088/1748-9326/abf004.

Buizza, R., M. Milleer, and T. N. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908, https://doi.org/10.1002/qj.49712556006.

Chan, D., A. Cobb, L. R. V. Zeppetello, D. S. Battisti, and P. Huybers, 2020: Summertime temperature variability increases with local warming in midlatitude regions. *Geophys. Res. Lett.*, **47**, e2020GL087624, https://doi.org/10.1029/2020GL087624.

Chattopadhyay, A., E. Nabizadeh, and P. Hassanzadeh, 2020: Analog forecasting of extreme-causing weather patterns using deep learning. *J. Adv. Model. Earth Syst.*, **12**, e2019MS001958, https://doi.org/10.1029/2019MS001958.

Deng, K., M. Ting, S. Yang, and Y. Tan, 2018: Increased frequency of summer extreme heat waves over Texas area tied to the amplification of Pacific zonal SST gradient. *J. Climate*, **31**, 5629–5647, https://doi.org/10.1175/JCLI-D-17-0554.1.

ECMWF, 2016: IFS Documentation CY43R1–Part V: Ensemble prediction system. ECMWF IFS Doc. 5, 23 pp., https://doi.org/10.21957/6fm80smm.

——, 2019: IFS Documentation CY46R1–Part V: Ensemble prediction system. ECMWF IFS Doc. 5, 23 pp., https://doi.org/10.21957/38yug0cev.

——, 2020: IFS Documentation CY47R1–Part V: Ensemble prediction system. ECMWF IFS Doc. 5, 23 pp., https://doi.org/10.21957/d7e3hrb.

Espeholt, L., and Coauthors, 2022: Deep learning for twelve hour precipitation forecasts. *Nat. Commun.*, **13**, 5145, https://doi.org/10.1038/s41467-022-32483-x.

Ferro, C. A. T., and D. B. Stephenson, 2011: Extremal dependence indices: Improved verification measures for deterministic forecasts of rare binary events. *Wea. Forecasting*, **26**, 699–713, https://doi.org/10.1175/WAF-D-10-05030.1.

Guillaumin, A. P., and L. Zanna, 2021: Stochastic-deep learning parameterization of ocean momentum forcing. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002534, https://doi.org/10.1029/2021MS002534.

Hall, P., J. L. Horowitz, and B.-Y. Jing, 1995: On blocking rules for the bootstrap with dependent data. *Biometrika*, **82**, 561–574, https://doi.org/10.1093/biomet/82.3.561.

He, K., X. Zhang, S. Ren, and J. Sun, 2015: Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. arXiv, 1502.01852v1, https://doi.org/10.48550/arXiv.1502.01852.

Heo, S., M. L. Bell, and J.-T. Lee, 2019: Comparison of health risks by heat wave definition: Applicability of wet-bulb globe temperature for heat wave criteria. *Environ. Res.*, **168**, 158–170, https://doi.org/10.1016/j.envres.2018.09.032.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Huang, H., and Coauthors, 2020: UNet 3+: A full-scale connected UNet for medical image segmentation. arXiv, 2004.08790v1, https://doi.org/10.48550/arXiv.2004.08790.

Inubushi, M., and S. Goto, 2020: Transfer learning for nonlinear dynamics and its application to fluid turbulence. *Phys. Rev. E*, **102**, 043301, https://doi.org/10.1103/PhysRevE.102.043301.

Jacques-Coper, M., S. Brönnimann, O. Martius, C. S. Vera, and S. B. Cerne, 2015: Evidence for a modulation of the intraseasonal summer temperature in eastern Patagonia by the Madden-Julian oscillation. *J. Geophys. Res. Atmos.*, **120**, 7340–7357, https://doi.org/10.1002/2014JD022924.

Jacques-Dumas, V., F. Ragone, P. Borgnat, P. Abry, and F. Bouchet, 2022: Deep learning-based extreme heatwave forecast. *Front. Climate*, **4**, 789641, https://doi.org/10.3389/fclim.2022.789641.

Joyce, J. M., 2011: Kullback-Leibler divergence. *International Encyclopedia of Statistical Science*, M. Lovric, Ed., Springer, 720–722, https://doi.org/10.1007/978-3-642-04898-2_327.

Ke, X., D. Wu, J. Rice, M. Kintner-Meyer, and N. Lu, 2016: Quantifying impacts of heat waves on power grid operation. *Appl. Energy*, **183**, 504–512, https://doi.org/10.1016/j.apenergy.2016.08.188.

Keisler, R., 2022: Forecasting global weather with graph neural networks. arXiv, 2202.07575v1, https://doi.org/10.48550/arxiv.2202.07575.

Kingma, D. P., and M. Welling, 2013: Auto-encoding variational Bayes. arXiv, 1312.6114v11, https://doi.org/10.48550/arxiv.1312.6114.

——, and J. Ba, 2017: Adam: A method for stochastic optimization. arXiv, 1412.6980v9, https://doi.org/10.48550/arXiv.1412.6980.

Kullback, S., and R. A. Leibler, 1951: On information and sufficiency. *Ann. Math. Stat.*, **22**, 79–86, https://doi.org/10.1214/aoms/1177729694.

Lerch, S., T. L. Thorarinsdottir, F. Ravazzolo, and T. Gneiting, 2017: Forecaster's dilemma: Extreme events and forecast evaluation. *Stat. Sci.*, **32**, 106–127, https://doi.org/10.1214/16-STS588.

Li, X., Y. Grandvalet, and F. Davoine, 2018: Explicit inductive bias for transfer learning with convolutional networks. *Proc. 35th Int. Conf. on Machine Learning*, Stockholm, Sweden, PMLR, 2825–2834, https://proceedings.mlr.press/v80/li18a.html.

Lin, H., R. Mo, and F. Vitart, 2022: The 2021 western North American heatwave and its subseasonal predictions. *Geophys. Res. Lett.*, **49**, e2021GL097036, https://doi.org/10.1029/2021GL097036.

Lopez-Gomez, I., Y. Cohen, J. He, A. Jaruga, and T. Schneider, 2020: A generalized mixing length closure for eddy-diffusivity mass-flux schemes of turbulence and convection. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002161, https://doi.org/10.1029/2020MS002161.

——, C. Christopoulos, H. L. L. Ervik, O. R. A. Dunbar, Y. Cohen, and T. Schneider, 2022: Training physics-based machine-learning parameterizations with gradient-free ensemble Kalman methods. *J. Adv. Model. Earth Syst.*, **14**, e2022MS003105, https://doi.org/10.1029/2022MS003105.

Lorenz, E. N., 1969a: Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.*, **26**, 636–646, https://doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2.

——, 1969b: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307, https://doi.org/10.1111/j.2153-3490.1969.tb00444.x.

Malardel, S., N. Wedi, W. Deconinck, M. Diamantakis, C. Kuehnlein, G. Mozdzynski, M. Hamrud, and P. Smolarkiewicz, 2016: A new grid for the IFS. *ECMWF Newsletter*, No. 146, ECMWF, Reading, United Kingdom, 23–28, https://www.ecmwf.int/node/17262.

Maloney, E. D., A. F. Adames, and H. X. Bui, 2019: Madden–Julian oscillation changes under anthropogenic warming. *Nat. Climate Change*, **9**, 26–33, https://doi.org/10.1038/s41558-018-0331-6.

Mann, M. E., S. Rahmstorf, K. Kornhuber, B. A. Steinman, S. K. Miller, S. Petri, and D. Coumou, 2018: Projected changes in persistent extreme summer weather events: The role of quasi-resonant amplification. *Sci. Adv.*, **4**, eaat3272, https://doi.org/10.1126/sciadv.aat3272.

Mass, C. F., D. Ovens, K. Westrick, and B. A. Colle, 2002: Does increasing horizontal resolution produce more skillful forecasts?: The results of two years of real-time numerical weather prediction over the Pacific Northwest. *Bull. Amer. Meteor. Soc.*, **83**, 407–430, https://doi.org/10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2.

Miller, D. E., Z. Wang, B. Li, D. S. Harnos, and T. Ford, 2021: Skillful subseasonal prediction of U.S. extreme warm days and standardized precipitation index in boreal summer. *J. Climate*, **34**, 5887–5898, https://doi.org/10.1175/JCLI-D-20-0878.1.

Miloshevich, G., B. Cozian, P. Abry, P. Borgnat, and F. Bouchet, 2022: Probabilistic forecasts of extreme heatwaves using convolutional neural networks in a regime of lack of data. arXiv, 2208.00971v1, https://doi.org/10.48550/arxiv.2208.00971.

Mo, R., H. Lin, and F. Vitart, 2022: An anomalous warm-season trans-Pacific atmospheric river linked to the 2021 western North America heatwave. *Commun. Earth Environ.*, **3**, 127, https://doi.org/10.1038/s43247-022-00459-w.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.*, **122**, 73–119, https://doi.org/10.1002/qj.49712252905.

Ogi, M., Y. Tachibana, and K. Yamazaki, 2003: Impact of the wintertime north Atlantic oscillation (NAO) on the summertime atmospheric circulation. *Geophys. Res. Lett.*, **30**, 1704, https://doi.org/10.1029/2003GL017280.

Orth, R., and S. I. Seneviratne, 2012: Analysis of soil moisture memory from observations in Europe. *J. Geophys. Res.*, **117**, D15115, https://doi.org/10.1029/2011JD017366.

Palmer, T. N., 1993: Extended-range atmospheric prediction and the Lorenz model. *Bull. Amer. Meteor. Soc.*, **74**, 49–66, https://doi.org/10.1175/1520-0477(1993)074<0049:ERAPAT>2.0.CO;2.

——, 2017: The primacy of doubt: Evolution of numerical weather prediction from determinism to probability. *J. Adv. Model. Earth Syst.*, **9**, 730–734, https://doi.org/10.1002/2017MS000999.

Parente, J., M. Pereira, M. Amraoui, and E. Fischer, 2018: Heat waves in Portugal: Current regime, changes in future climate and impacts on extreme wildfires. *Sci. Total Environ.*, **631–632**, 534–549, https://doi.org/10.1016/j.scitotenv.2018.03.044.

Pathak, J., and Coauthors, 2022: FourCastNet: A global data-driven high-resolution weather model using adaptive Fourier neural operators. arXiv, 2202.11214v1, https://doi.org/10.48550/arxiv.2202.11214.

Perkins-Kirkpatrick, S. E., and S. C. Lewis, 2020: Increasing trends in regional heatwaves. *Nat. Commun.*, **11**, 3357, https://doi.org/10.1038/s41467-020-16970-7.

Pesciullesi, G., P. Schwaller, T. Laino, and J.-L. Reymond, 2020: Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. *Nat. Commun.*, **11**, 4874, https://doi.org/10.1038/s41467-020-18671-7.

Qi, D., and A. J. Majda, 2020: Using machine learning to predict extreme events in complex systems. *Proc. Natl. Acad. Sci. USA*, **117**, 52–59, https://doi.org/10.1073/pnas.1917285117.

Rasp, S., P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey, 2020: WeatherBench: A benchmark data set for data-driven weather forecasting. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002203, https://doi.org/10.1029/2020MS002203.

Ravuri, S., and Coauthors, 2021: Skilful precipitation nowcasting using deep generative models of radar. *Nature*, **597**, 672–677, https://doi.org/10.1038/s41586-021-03854-z.

Rezende, D., and S. Mohamed, 2015: Variational inference with normalizing flows. *Proc. 32nd Int. Conf. on Machine Learning*, Lille, France, PMLR, 1530–1538, https://proceedings.mlr.press/v37/rezende15.html.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649–667, https://doi.org/10.1002/qj.49712656313.

Robine, J.-M., S. L. K. Cheung, S. Le Roy, H. Van Oyen, C. Griffiths, J.-P. Michel, and F. R. Herrmann, 2008: Death toll exceeded 70,000 in Europe during the summer of 2003. *C. R. Biol.*, **331**, 171–178, https://doi.org/10.1016/j.crvi.2007.12.001.

Ronchi, C., R. Iacono, and P. S. Paolucci, 1996: The "cubed sphere": A new method for the solution of partial differential equations in spherical geometry. *J. Comput. Phys.*, **124**, 93–114, https://doi.org/10.1006/jcph.1996.0047.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-Net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*, N. Navab et al., Eds., Springer International Publishing, 234–241, https://doi.org/10.1007/978-3-319-24574-4_28.

Ruffault, J., and Coauthors, 2020: Increased likelihood of heat-induced large wildfires in the Mediterranean Basin. *Sci. Rep.*, **10**, 13790, https://doi.org/10.1038/s41598-020-70069-z.

Sánchez-Benítez, A., R. García-Herrera, D. Barriopedro, P. M. Sousa, and R. M. Trigo, 2018: June 2017: The earliest European summer mega-heatwave of reanalysis period. *Geophys. Res. Lett.*, **45**, 1955–1962, https://doi.org/10.1002/2018GL077253.

Scher, S., and G. Messori, 2021: Ensemble methods for neural network-based weather forecasts. *J. Adv. Model. Earth Syst.*, **13**, e2020MS002331, https://doi.org/10.1029/2020MS002331.

Slingo, J., and T. Palmer, 2011: Uncertainty in weather and climate prediction. *Philos. Trans. Roy. Soc.*, **A369**, 4751–4767, https://doi.org/10.1098/rsta.2011.0161.

Sønderby, C. K., and Coauthors, 2020: MetNet: A neural weather model for precipitation forecasting. arXiv, 2003.12140v2, https://doi.org/10.48550/ARXIV.2003.12140.

Sundararajan, M., and S. Agrawal, 2021: The rain check. http://raincheck.karyk.com/rain-check.

——, A. Taly, and Q. Yan, 2017: Axiomatic attribution for deep networks. arXiv, 1703.01365v2, https://doi.org/10.48550/arxiv.1703.01365.

Teng, H., G. Branstator, H. Wang, G. A. Meehl, and W. M. Washington, 2013: Probability of US heat waves affected by a subseasonal planetary wave pattern. *Nat. Geosci.*, **6**, 1056–1061, https://doi.org/10.1038/ngeo1988.

Ullrich, P. A., and M. A. Taylor, 2015: Arbitrary-order conservative and consistent remapping and a theory of linear maps: Part I. *Mon. Wea. Rev.*, **143**, 2419–2440, https://doi.org/10.1175/MWR-D-14-00343.1.

——, D. Devendran, and H. Johansen, 2016: Arbitrary-order conservative and consistent remapping and a theory of linear maps: Part II. *Mon. Wea. Rev.*, **144**, 1529–1549, https://doi.org/10.1175/MWR-D-15-0301.1.

Vautard, R., and Coauthors, 2007: Summertime European heat and drought waves induced by wintertime Mediterranean rainfall deficit. *Geophys. Res. Lett.*, **34**, L07711, https://doi.org/10.1029/2006GL028001.

Vitart, F., and Coauthors, 2017: The subseasonal to seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

Wang, G., A. J. Dolman, and A. Alessandri, 2011: A summer climate regime over Europe modulated by the North Atlantic oscillation. *Hydrol. Earth Syst. Sci.*, **15**, 57–64, https://doi.org/10.5194/hess-15-57-2011.

Weyn, J. A., D. R. Durran, and R. Caruana, 2020: Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere. *J. Adv. Model. Earth Syst.*, **12**, e2020MS002109, https://doi.org/10.1029/2020MS002109.

——, ——, ——, and N. Cresswell-Clay, 2021: Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *J. Adv. Model. Earth Syst.*, **13**, e2021MS002502, https://doi.org/10.1029/2021MS002502.

White, C. J., and Coauthors, 2017: Potential applications of sub-seasonal-to-seasonal (S2S) predictions. *Meteor. Appl.*, **24**, 315–325, https://doi.org/10.1002/met.1654.

White, R. H., K. Kornhuber, O. Martius, and V. Wirth, 2022: From atmospheric waves to heatwaves: A waveguide perspective for understanding and predicting concurrent, persistent, and extreme extratropical weather. *Bull. Amer. Meteor. Soc.*, **103**, E923–E935, https://doi.org/10.1175/BAMS-D-21-0170.1.

Wilks, D. S., 2019: *Statistical Methods in the Atmospheric Sciences.* 4th ed. Elsevier, 369–483, https://doi.org/10.1016/B978-0-12-815823-4.00009-2.

Wright, C. K., K. M. de Beurs, and G. M. Henebry, 2014: Land surface anomalies preceding the 2010 Russian heat wave and a link to the North Atlantic oscillation. *Environ. Res. Lett.*, **9**, 124015, https://doi.org/10.1088/1748-9326/9/12/124015.

Wu, W., M. A. Geller, and R. E. Dickinson, 2002: The response of soil moisture to long-term variability of precipitation. *J. Hydrometeor.*, **3**, 604–613, https://doi.org/10.1175/1525-7541(2002)003<0604:TROSMT>2.0.CO;2.

Wulff, C. O., and D. I. V. Domeisen, 2019: Higher subseasonal predictability of extreme hot European summer temperatures as compared to average summers. *Geophys. Res. Lett.*, **46**, 11 520–11 529, https://doi.org/10.1029/2019GL084314.

Xu, Z., G. FitzGerald, Y. Guo, B. Jalaludin, and S. Tong, 2016: Impact of heatwave on mortality under different heatwave definitions: A systematic review and meta-analysis. *Environ. Int.*, **89–90**, 193–203, https://doi.org/10.1016/j.envint.2016.02.007.

Yosinski, J., J. Clune, Y. Bengio, and H. Lipson, 2014: How transferable are features in deep neural networks? *Advances in Neural Information Processing Systems*, Z. Ghahramani et al., Eds., Vol. 27, Curran Associates, Inc., 2204–2212, https://proceedings.neurips.cc/paper/2014/hash/375c71349b295fbe2dcdca9206f20a06-Abstract.html.

Yu, F., and V. Koltun, 2016: Multi-scale context aggregation by dilated convolutions. *Fourth Int. Conf. on Learning Representations*, San Juan, Puerto Rico, ICLR, https://doi.org/10.48550/arXiv.1511.07122.

Zhou, Y., B. Yang, H. Chen, Y. Zhang, A. Huang, and M. La, 2019: Effects of the Madden–Julian oscillation on 2-m air temperature prediction over China during boreal winter in the S2S database. *Climate Dyn.*, **52**, 6671–6689, https://doi.org/10.1007/s00382-018-4538-z.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.*, **83**, 73–84, https://doi.org/10.1175/1520-0477(2002)083<0073:TEVOEB>2.3.CO;2.