ARTICLE OPEN

Evaluation of heat wave forecasts seamlessly across subseasonal timescales

Trent W. Ford[1], Paul A. Dirmeyer[2] and David O. Benson[3]

We develop an extreme heat validation approach for medium-range forecast models and apply it to the NCEP coupled forecast model, for which we also attempt to diagnose sources of poor forecast skill. A weighting strategy based on the Poisson function is developed to provide a seamless transition from short-term day-by-day weather forecasts to expanding time means across subseasonal timescales. The skill of heat wave forecasts over the conterminous United States is found to be rather insensitive to the choice of skill metric; however, forecast skill does display spatial patterns that vary depending on whether daily mean, minimum, or maximum temperatures are the basis of the heat wave metric. The NCEP model fails to persist heat waves as readily as is observed. This inconsistency worsens with longer forecast lead times. Land–atmosphere feedbacks appear to be a stronger factor for heat wave maintenance at southern latitudes, but the NCEP model seems to misrepresent those feedbacks, particularly over the Southwest United States, leading to poor skill in that region. The NCEP model also has unrealistically weak coupling over agricultural areas of the northern United States, but this does not seem to degrade model skill there. Overall, we find that the Poisson weighting strategy combined with a variety of deterministic and probabilistic skill metrics provides a versatile framework for validation of dynamical model heat wave forecasts at subseasonal timescales.

npj Climate and Atmospheric Science (2018)1:20 ; doi:10.1038/s41612-018-0027-7

INTRODUCTION

Heat waves have major implications for human health, particularly in urban areas where enhanced vulnerability to high temperature and humidity over the last few decades[1,2] is in direct response to increases in heat wave frequency, intensity, and duration.[3–5] Recent changes and projected future increases in heat wave frequency and intensity in many regions[6,7] have been attributed to increased greenhouse gas concentrations and their effect on climate, favoring warmer conditions.[8,9]

Heat waves are usually concurrent with persistent atmospheric circulation features: high pressure systems that force conditions favorable for extreme temperatures. Land surface moisture deficits and land–atmosphere feedbacks have been connected to the onset and maintenance of heat waves in many regions.[10–12] Memory in land surface states provides a potential source of prediction skill, as the slow manifold of land and ocean are primary sources of atmospheric predictability on subseasonal timescales. Positive land–atmosphere feedbacks can exacerbate and prolong temperature anomalies, providing a form of coupled memory.[13] For forecasting purposes, land surface memory may be defined by the temporal extent of improved forecast skill when realistic land surface initiation conditions are used in a model.[14–16] Because land surface memory is most relevant at subseasonal timescales, accurate land surface modeling is important for subseasonal forecasts. We use the term "subseasonal" to refer to timescales of less than 90 days, which encompass traditional short-range (1–5 days), medium-range (5–14 days), and longer (up to 2 months) forecasts. General circulation model intercomparison projects like the Global Land-Atmosphere Coupling Experiment (GLACE[17]) are leading to advancements in model simulation and forecast accuracy. Results from GLACE-2[18] demonstrated that model fidelity regarding key processes that communicate initial land surface anomalies to the atmosphere leads to significantly improved skill from better land surface initialization. Nevertheless, there remain knowledge gaps,[19] particularly concerning the influence of atmospheric and land surface model parameterizations and forecast initialization on model forecast reliability.[20] Therefore, more work is necessary to understand and quantify the value of the land surface and soil moisture memory, particularly for subseasonal forecasts of heat waves.

Additionally, an important component of model evaluation is the comparison of model forecast performance across lead times; this is particularly vital for forecasting heat waves and other events that manifest on subseasonal timescales. However, many forecast verification methods are based on a dichotomous outcome (0,1) or a deterministic value on a single forecast day, and thus do not consider the observed outcome on days surrounding the forecast day. This is sensible when applied at short lead times, as a model forecast should not be considered skillful if a 1-day lead forecast calls for a heat wave when the actual heat wave begins days later. However, as we go toward medium-range timescales, such a rigid deterministic approach makes less sense. If a 10-day lead forecast shows a heat wave beginning and lasting 3 days while the actual heat wave begins on day 12 and lasts 4 days, this should be considered useful and not be penalized for only overlapping on one day. This hypothetical situation exemplifies the challenge of applying the same model verification framework over a wide range of forecast lead times.

[1]Southern Illinois University, Carbondale, IL, USA; [2]Center for Ocean-Land-Atmosphere Studies, George Mason University, Fairfax, VA, USA and [3]George Mason University, Fairfax, VA, USA
Correspondence: Trent W. Ford (twford@siu.edu)

Received: 30 June 2017 Revised: 14 February 2018 Accepted: 20 February 2018
Published online: 15 October 2018

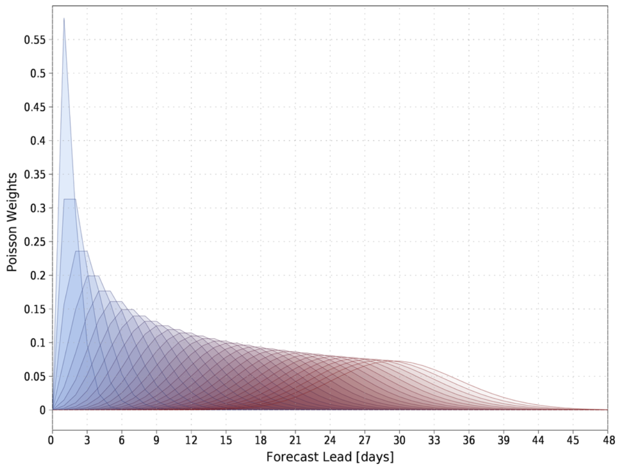

Fig. 1 Poisson weighting framework for seamless transition from deterministic to probabilistic forecasts from short- to medium-range timescales. The figure shows the shape of the weights with forecast leads from 1 to 30 days (blue to red)

In response to the aforementioned questions and challenges regarding subseasonal heat wave forecasting, the objectives of this paper are to (1) design and implement a framework for seamlessly evaluating heat wave forecasts spanning subseasonal timescales and (2) diagnose potential sources of forecast fidelity, specifically focused on the physical coupling between land surface and atmosphere. The larger objective of our project is to apply the heat wave validation framework developed here to a suite of medium-range forecast models.

We defer a full multi-model comparison and present here an evaluation framework using one model: the National Centers for Environmental Prediction (NCEP) Climate Forecast System, version 2 (CFSv2[21]). The NCEP model, like any single model, comes with a set of advantages and limitations. One of its limitations is that it only has four ensemble members; however, the members are initialized and forecasts issued daily at 6-h intervals, providing a continuous stream of forecasts representing every lead time and validating every day. This is advantageous as compared to models that may have more ensemble members, but from which forecasts are only issued every few days. The 11-year period (1999–2010) over which NCEP forecasts are available is shorter than ideal; however, given the objectives of this study, particularly the intent of developing a methodology that will be applied to a suite of medium-range forecast models, we argue that our focus on the NCEP model forecasts is justified. Predictions of extreme heat factor (see Methods) are evaluated with four skill metrics and a seamless methodology using temperature observations and reanalyses, the latter useful for further evaluating model representation of processes involved in heat waves.

RESULTS

Seamless transition across timescales

Model verification is based on four complimentary metrics (see Methods section): the area under the relative operating characteristic curve (AUC); reliability; the equitable threat score (ETS), and the Kullback-Leibler divergence (KLD). We selected these four model verification metrics because they each evaluate a different facet of forecast "skill". AUC is a measure of discrimination, meaning a high AUC maximizes forecast hits while minimizing false positives. Reliability and ETS both reward consistency among ensemble members, while penalizing false positives and false negatives; however, reliability assess each ensemble member while ETS assesses the ensemble mean. KLD does not treat the heat wave forecasts as dichotomous events, but instead accounts for the forecasted heat wave probability distribution. The metrics, taken together provide a comprehensive evaluation of model heat wave forecast skill. The range of forecast lead times examined, in the case of the NCEP model from 1 to 30 days, complicates model verification and precludes us from applying the same metrics to evaluate a single day's forecast value or dichotomous outcome. Our solution for seamlessly evaluating model forecasts across a wide range of lead times is to apply a Poisson weighting strategy (see Methods section) that accounts for the innate uncertainty accompanying longer lead-time forecasts. The Poisson distribution is selected because (1) distribution of weights expands with increasing forecast lead time, transitioning seamlessly from deterministic to time-averaged validation (2) its asymmetry allows for expansion to a range of maximal forecast lead times, improving the method's transferability between models, and (3) it is relatively easy to compute. These features allow the distribution in time of the weights broaden the event window as the lead time increases (from blue to red in Fig. 1), thereby including more surrounding days in the weighting of the validation state at longer forecast leads. To test the benefit of the weighting for verifying heat wave (extreme heat factor or EHF) forecasts, an idealized situation is created where the validation EHF also acts as the forecast, but shifted by 0–29 days. This simulates a phase error in an otherwise "perfect" prediction. The black line with circles in the inset of Fig. S1 represents the deterministic approach to forecast verification—single-day comparisons—where the skill (represented by the normalized KLD metric) degrades quickly with growing phase error. The normalization is by the skill of the 36-year climatology of EHF from observations; a value of 1 matches the skill of the climatology forecast, and skill better than climatology has a score < 1. In contrast to the deterministic approach, applying the Poisson window for forecast verification (array of colored lines in Fig. S1 inset) allows for the more intuitive notion that, for instance, a forecast of a heat wave at a 30-day lead that is shifted by <10 days is still better than a climatological forecast.

Further analysis using the KLD and ETS metrics reveals that the decline in model forecast skill from 1- to 30-day lead times is usually much faster for the deterministic forecasts than when the Poisson weighting method is used. Therefore, the effect of this weighting scheme—determined by the difference in skill between deterministic and weighted forecast verification—grows as the forecast lead time increases. This finding is noteworthy as the weighting scheme developed here produces similar results to deterministic verification at short leads while allowing a fair comparison of model forecast skill between short-range and medium-range lead times.

The Poisson weighting method does introduce one issue with forecast verification. The calculated forecast climatology and validation data also must be a weighted function of lead time. Because the number of days included in the weighting increases with lead time, the Poisson-weighted probability of a heat wave occurring in the observations also increases with lead time (Fig. S2, black line). In order to assuage this issue for model verification, the (0,1) (no heat wave, heat wave) observation for each day included in the weighting is multiplied by its respective weight. A heat wave is then in the validation dataset if the sum of all observation-weight products exceeds 0.5. This adjustment reduces the frequency of heat waves in the validation time series (Fig. S2, red line), and approaches the heat wave frequency of the model.

Forecast skill

Skill evaluation of NCEP subseasonal forecasts using the KLD metric is shown in Fig. 2, where the basis of skill is the longest lead time at which model forecasts are more skillful than a climatological forecast, calculated from 36 years of positive EHF in the observations. The Poisson weighting method clearly extends the forecast skill compared to the deterministic

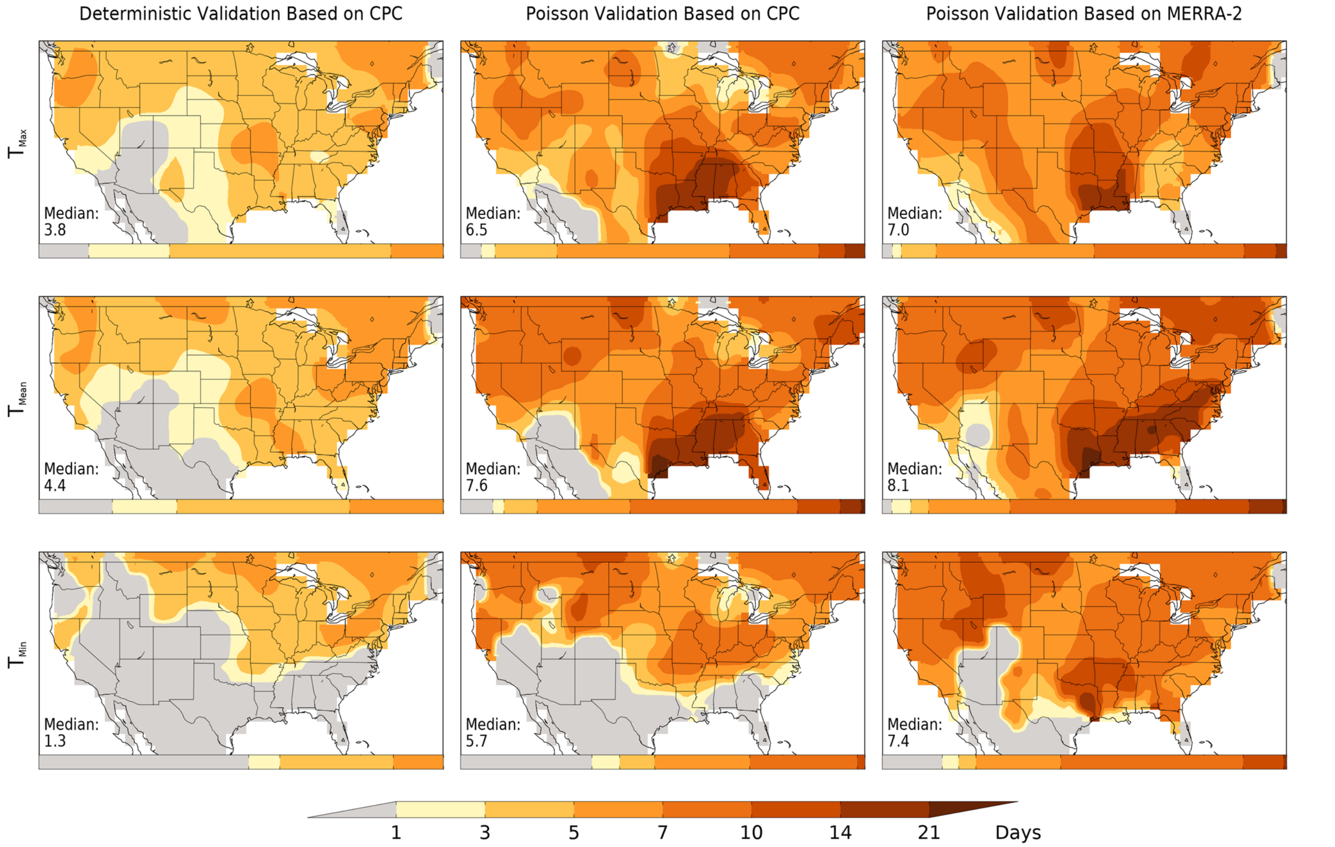

Fig. 2 Lead beyond which skill of NCEP model forecasts of EHF falls below that of a climatological forecast for $T_{\max}$ (top), T_{mean} (middle), and $T_{\min}$ (bottom) with different validation approaches against observations and MERRA-2 temperature analyses. Widths of colors in the bars at the bottom of each panel reflect the fractional area in each map; median value in days is given in the corner of each panel

approach—by more than 2 weeks in some parts of Southeast and South Central United States. The median lead for a forecast more skillful than climatology is extended by about 3 days in the case of EHF based on $T_{\max}$, 3.5 days for T_{mean}, and 4 days for $T_{\min}$. The extension of skill into longer lead times when applying the Poisson weighting scheme is not surprising, given that medium-range forecasts do not need to be as precise in their timing.

Concentrating on the Poisson-based validation, we see the lead time of skillful forecasts is slightly higher when Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2) reanalysis is used for validation, yet much higher across the southern United States for $T_{\min}$. This feature in the southern United States is a result of the NCEP model having a larger hit rate when validated with MERRA-2 than when validated with the NOAA Climate Prediction Center (CPC) global daily temperature data. This is most likely attributable to shared bias between the MERRA-2 reanalysis CFS reanalysis product, the latter of which is used for initial conditions in the NCEP model. Otherwise, features between the CPC and MERRA-2 products are similar: the lower Mississippi Valley is generally a region of long-duration skill, as are the northern Rockies and Northwest. Such similarity is used to justify MERRA-2 as a diagnostic dataset for NCEP model behavior, as described in the Methods section. Regions where the forecast model struggles include Florida and the Southwest. The skill minimum over Lake Michigan appears to be because the lake is not resolved in the NCEP forecast model.

Consistent with the KLD results, skill evaluation using the AUC, reliability, and ETS metrics shows the highest heat wave forecast skill in the northern Rockies and Southern Great Plains, but degrades at lead times beyond ~10–15 days (Figs. 3 and 4). This is especially true over the western United States, where AUC scores approach 0.5 (no skill) at lead times beyond 20 days. Spatial patterns of AUC are similar between $T_{\max}$, T_{mean}, and $T_{\min}$ forecasts (Fig. 3). Despite a general west-to-east decreasing gradient of skill at short lead times, AUC remains above 0.6 throughout much of the eastern United States even beyond 25-day leads (Fig. 3).

AUC is a measure of model discrimination, and is therefore dependent on both the frequency of true positive and false positive heat wave forecasts. The noticeable west-east decreasing gradient of skill (as determined by the AUC in Fig. 3) at short lead times is contrasted by a rapid decline in skill in the western half of the United States at 15- to 25-day lead times. The spread of positive heat wave forecasts between ensemble members in the central Midwest and Southeast regions increases as the lead time increases. This means that at longer lead times (>15 days), fewer heat wave events are simultaneously predicted by multiple ensemble members. The result is a larger rate of true positive forecasts when only one or two ensemble members had to indicate a heat wave, and a smaller rate of false positive forecasts when three or more members had to indicate a heat wave (Fig. S7). However, larger ensemble spread is also accompanied by an increase in false negative forecasts (misses), although the AUC metric does not penalize false negative forecasts. Consequently, the increase in hits and decrease in false positives that corresponds with larger ensemble member spread results in an AUC increase, when in reality the overall model skill is reduced. Therefore, although the Midwest and Southeast show higher overall "skill" according to the AUC at longer lead times, this is attributable to disagreement between ensemble members and should be carefully interpreted.

Model forecast skill assessed via the reliability metric shows similar results as those based on AUC (Fig. S9): a general west-east

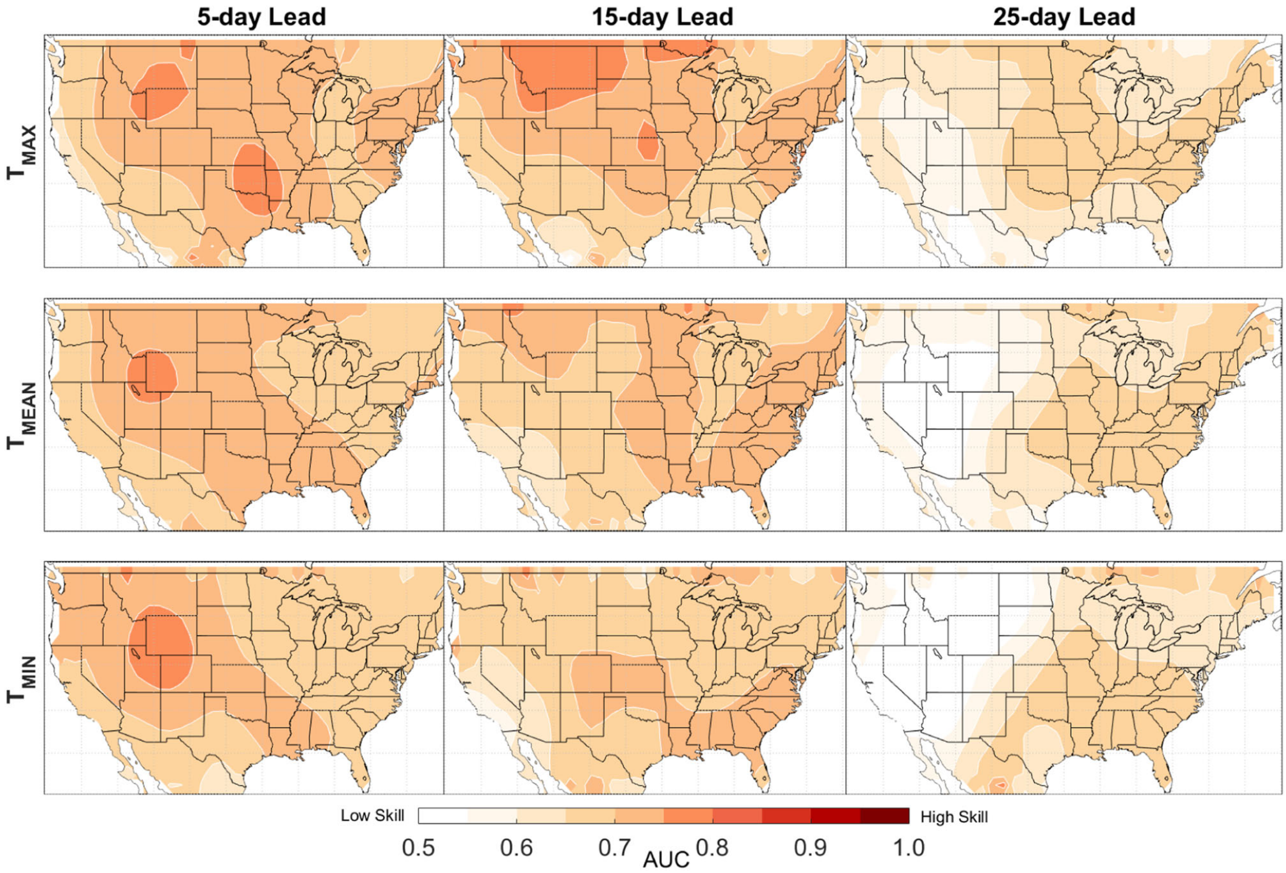

Fig. 3 Panels show areas under the ROC curve (AUC) for heat wave forecasts validated with observations. AUC is shown of $T_{\max}$ (top), T_{mean} (middle), and $T_{\min}$ (bottom) at 5-day (left), 15-day (middle), and 25-day (right) lead times

decreasing gradient of skill appears at shorter lead times. However, unlike the AUC patterns, this west-east reliability gradient tends to persist at longer lead times, although with an overall decrease in reliability (Fig. S9). Under-forecasting of heat waves in the Midwest that lead to maintenance of relatively high AUC (Fig. 3) causes a noticeable decline in reliability at a 25-day lead time (Fig. S9), as reliability accounts for false negatives.

Assessing model forecast skill with ETS results in similar spatial patterns to the AUC, reliability, and KLD validations. Most skill is found at smaller lead times in the northwest and inter-mountain west, with less skill in the southeast (Fig. 4). Unlike AUC and reliability, ETS values quickly diminish, and forecasts are devoid of skill with respect to a random forecast (ETS ~ 0), beyond a 25-day lead over much of the country. ETS is not sensitive to inter-member spread or disagreement like AUC, partly explaining its lack of skill in the Midwest at long lead times.

Sources of skill

One advantage of EHF as a heat wave metric is its mandate of the 3-day $T_{\max}$, T_{mean}, or $T_{\min}$ to exceed an annual climatological threshold, as opposed to identifying a heat wave during any day that exceeds a daily climatological threshold. This requirement is based on research linking multiple-day heat accumulation to excess mortality and other adverse physiological impacts.[21,22] When using EHF to monitor heat wave occurrence, it is critical to understand both the absolute exceedance of a temperature threshold and the duration of that anomaly. The multi-day persistence of high temperature anomalies during extreme heat events is partly attributed to surface moisture limitations and land–atmosphere feedbacks.[23] We examine the persistence of extreme heat, in this case the conditional probability (p_c) of a heat wave (EHF > 0) on day n followed by a heat wave on day $n+1$ within the NCEP and observed datasets. Figure 4 shows the NCEP-observation difference in p_c for $T_{\max}$, T_{mean}, and $T_{\min}$ heat waves at 5-, 15-, and 25-day lead times. The maps show a distinct west-to-east gradient, such that the conditional probability of a heat wave persisting is noticeably larger in observations than in the NCEP forecasts for most of the eastern United States, particularly in lower Mississippi Valley. These differences grow with lead time, as p_c is nearly 25% lower in the NCEP system than in observations in the lower Mississippi Valley at 25-day lead times.

The reduced conditional probability of two consecutive heat wave days in the NCEP forecasts can be explained by day-to-day changes in $T_{\max}$, T_{mean}, and $T_{\min}$ subsequent to a heat wave day. Fig. S8 shows the NCEP and observed temperature changes between day n and day $n+1$, given that day n is a heat wave day. The NCEP model exhibits a considerably larger temperature decrease on days immediately following heat wave days compared to observations (Fig. S8). Reduced temperature anomaly persistence in the NCEP model is apparent in the Southeast United States at 5-day lead times, and expands to nearly every region of the United States at longer lead times. The maintenance of high temperature anomalies is necessary in order to indicate a heat wave when using EHF, and therefore any model-observation differences in daily temperature persistence will affect the model's heat wave forecast skill. The mismatch of hot day persistence can partly explain the sharp decline in NCEP forecast skill in the Southeast and central Midwest regions (Fig. 5).

Finally, we attempt to understand the spatial pattern of skill in the NCEP forecasts. CFSv2 has well-documented biases over crop vegetation types[16,24] that manifest as excessive evaporation, near-zero sensible heat flux, and very shallow atmospheric boundary

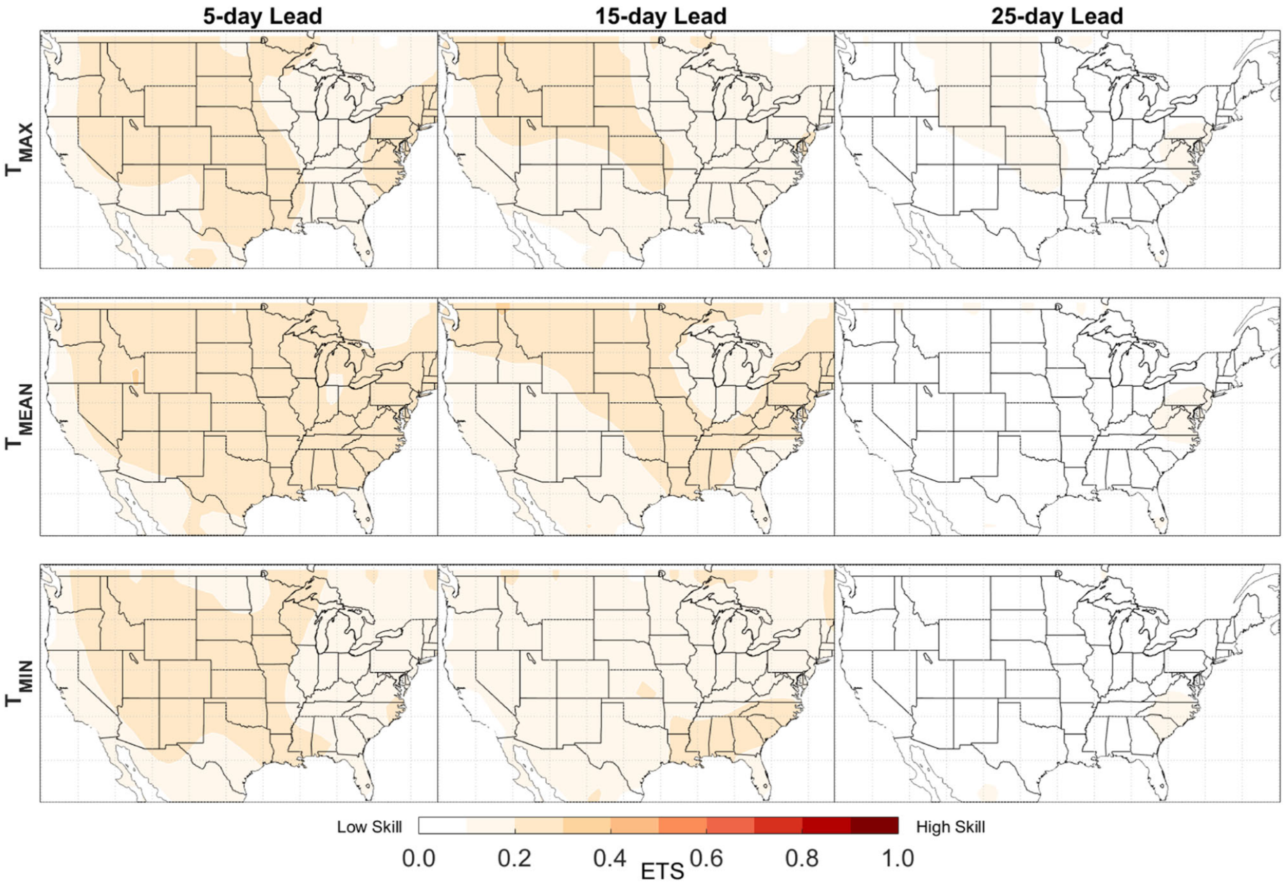

Fig. 4 Panels show areas under the equitable threat score (ETS) for heat wave forecasts validated with observations. ETS is shown of T_{max} (top), T_{mean} (middle), and T_{min} (bottom) at 5-day (left), 15-day (middle), and 25-day (right) lead times

layers. This may be a factor in the excessive model cooling during heat waves (Fig. S8). Figure 6 shows evidence of this bias: correlations between surface soil moisture and evaporative fraction during summer are generally high, indicating soil moisture's control on surface fluxes and the potential for land–atmosphere coupling,[25] decreasing from south to north, as seen in the MERRA-2 reanalysis (Fig. 6, top panels). For NCEP forecasts, that pattern is interrupted by a region of low correlation that corresponds to the crop region over the Ohio Valley, upper Mississippi and Missouri basins, and the lower Mississippi Delta region.

We use the Π feedback parameter[25] to diagnose consistencies in the regions where soil moisture can potentially exacerbate heat waves. Overall, spatial patterns of Π are quite consistent between the NCEP model and MERRA-2 (Fig. 6, middle panels); however, the overall strength of Π is lower in the forecast model, particularly in the central Rockies and Southwest. At longer forecast lead times the heat wave feedback index declines (bottom panels) over several regions, including the Southwest, suggesting there the NCEP model would have increasing difficulty maintaining land–atmosphere feedbacks that would support heat waves. This is consistent with the rapid drop in skill over the Southwest seen in Figs. 1 and 2. Meanwhile, the heat wave feedback strengthens with forecast lead over South Texas and parts of California.

DISCUSSION

Our results suggest that low heat wave forecast skill in some regions of the United States could be attributable to inconsistencies in the coupling between soil moisture and evapotranspiration (ET) and the coupling between ET and temperature. Further investigation is necessary to determine precisely the contribution of soil moisture–ET–temperature interactions to model forecast skill. A recent study found many climate models simulate too strong a linkage between low evaporation and temperatures, especially over semi-humid regions.[26] However, given widespread long-term drift in models, it is not clear if such biases manifest on subseasonal timescales where land and atmosphere initial conditions play a role. Finally, high humidity during heat waves plays an added role in human health and mortality,[27] but is not accounted for in this study. The framework developed here would be applicable for validating model forecasts of humid heat waves if applied to apparent temperature or equivalent temperature. The application of our framework for diagnosing model forecast fidelity for oppressive heat waves (i.e., hot and humid) is separate from the objectives of this particular study, but will be the focus of future research.

The results presented here demonstrate the advantages of implementing the Poisson weighting strategy for seamless model forecast validation across a range of lead times. This provides a framework for fair and consistent validation of model heat wave forecasts along the spectrum of short- to medium-range timescales, thereby affording a more comprehensive evaluation of model skill. An additional advantage of this method is its applicability to a wide range of skill metrics, as evidenced here with the combination of KLD, AUC, reliability, and ETS. This study demonstrates a robust framework for validating dynamic model heat wave forecasts across a range of timescales, and we will apply this framework for future multi-model validation and comparison.

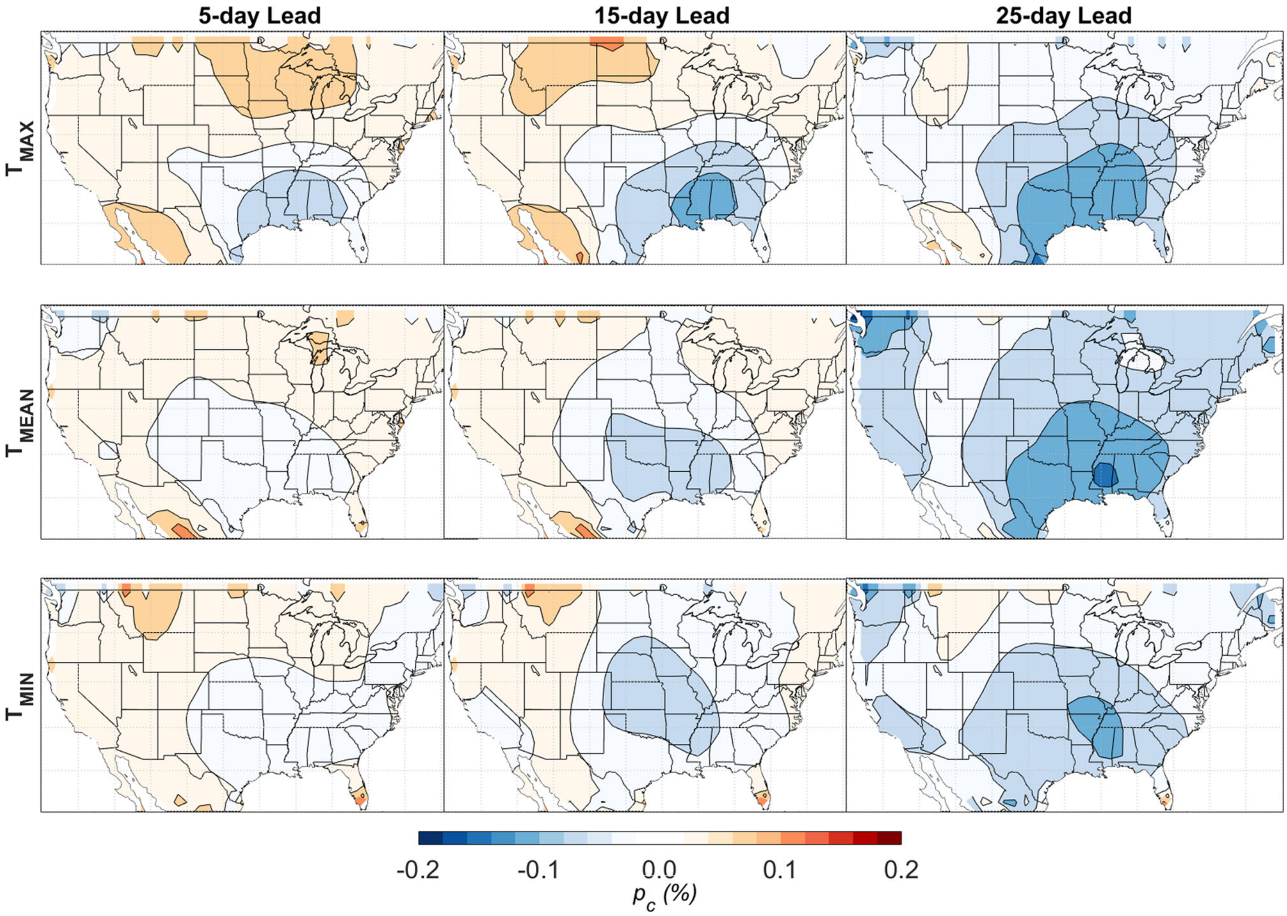

Fig. 5 Panels show NCEP-observed differences in the conditional probability of a heat wave on day $n+1$ given a heat wave on day n. Differences are shown as % probability, and separated by T_{max}, T_{mean}, and T_{min} at 5-, 15-, and 25-day NCEP forecast lead time

METHODS

Validation data

Model heat wave verification is based on 2-m maximum, mean, and minimum temperature from several sources. The CPC global daily temperature analysis at 0.5° resolution[28] for the period 1980–2015 is used as the observational validation source; reanalysis data are also used as they contain gridded fields in addition to temperature that can be compared to forecast model output in order to diagnose model behavior. The MERRA-2[29] is the primary atmospheric and land surface reanalysis product used to validate terrestrial and atmospheric segments of soil moisture–temperature coupling in subseasonal model forecasts. MERRA-2 fields are available at a 0.5° × 0.625° spatial resolution as daily data from 1980 to the present. The MERRA-2 system assimilates observations from atmospheric in situ and remote sensing sources; however, it does not assimilate 2-m temperature. One of many MERRA-2 updates over the original MERRA product is hourly precipitation correction guided by the National Oceanographic and Atmospheric Administration (NOAA) Climate Prediction Center's unified gauge-based analysis of global daily precipitation. This correction occurs within the atmosphere–land reanalysis system, thus also influences land surface states.[30] This results in improved correspondence between MERRA-2 and observations of soil moisture.[31] Soil moisture, sensible and latent heat flux, and 2-m temperature are used from MERRA-2.

Forecast model

The NCEP CFSv2 system includes an atmospheric model with 64 atmospheric sigma-pressure layers and a T126 horizontal resolution coupled with the NOAA GFDL Modular Ocean Model, version 4, including a sea ice model[32] and the four-layer Noah land surface model.[33] Model initial conditions are from the Climate Forecast System Reanalysis (CFSR), and one forecast is initialized every 6 h to produce a daily four-member ensemble. The land surface conditions are initialized from CFSR, which employs a parallel uncoupled Noah simulation to reset the CFSR land states every 24 h. Land states include realistic soil moisture (liquid + ice) for four layers down to 2-m below the surface. Vegetation is represented by a monthly climatology of green vegetation fraction, and soil properties are defined following.[34] Daily NCEP re-forecasts acquired from the Subseasonal-to-Seasonal (S2S) project database[35] for 1999 through 2010 are provided at 1.5° horizontal resolution. Daily model re-forecast data are examined at lead times ranging from 1 to 30 days.

Heat wave metrics

Extreme heat is identified using the EHF,[36] which has been widely used to identify heat waves and measure their intensity.[12,37] EHF quantifies heat wave intensity through two separate metrics, the significance excess heat index ($\mathrm{EHI_{sig}}$) and the acclimatization excess heat index ($\mathrm{EHI_{accl}}$), defined as:

$$\mathrm{EHI_{sig}} = \frac{T_{i-2}+T_{i-1}+T_i}{3} - T_{90} \tag{1}$$

$$EHI_{accl} = \frac{T_{i-2}+T_{i-1}+T_i}{3} - \frac{T_{i-32}+T_{i-31}+\cdots+T_{i-3}}{30}, \tag{2}$$

where T_i is daily 2-m maximum temperature (T_{max}), minimum temperature (T_{min}), or mean temperature (T_{mean}), and T_{90} is the climatological 90th percentile value of that particular metric. The indices are calculated over a sliding 3-day window over the duration of the study period and are set to zero if negative. The $\mathrm{EHI_{sig}}$ is a threshold index representing heat wave intensity with respect to T_{90}, while $\mathrm{EHI_{accl}}$ represents heat wave intensity with respect to human acclimatization based on conditions during the previous 30 days. The EHF is the product of the two indices:

$$\mathrm{EHF} = \max[1, \mathrm{EHI_{accl}}] \times \mathrm{EHI_{sig}}, \tag{3}$$

A heat wave is identified on day i if that day has a positive *EHF* value, thereby representing exceedance of T_{90} over a 3-day period, ending on day i. In order to fairly compare EHF between observations and the model,

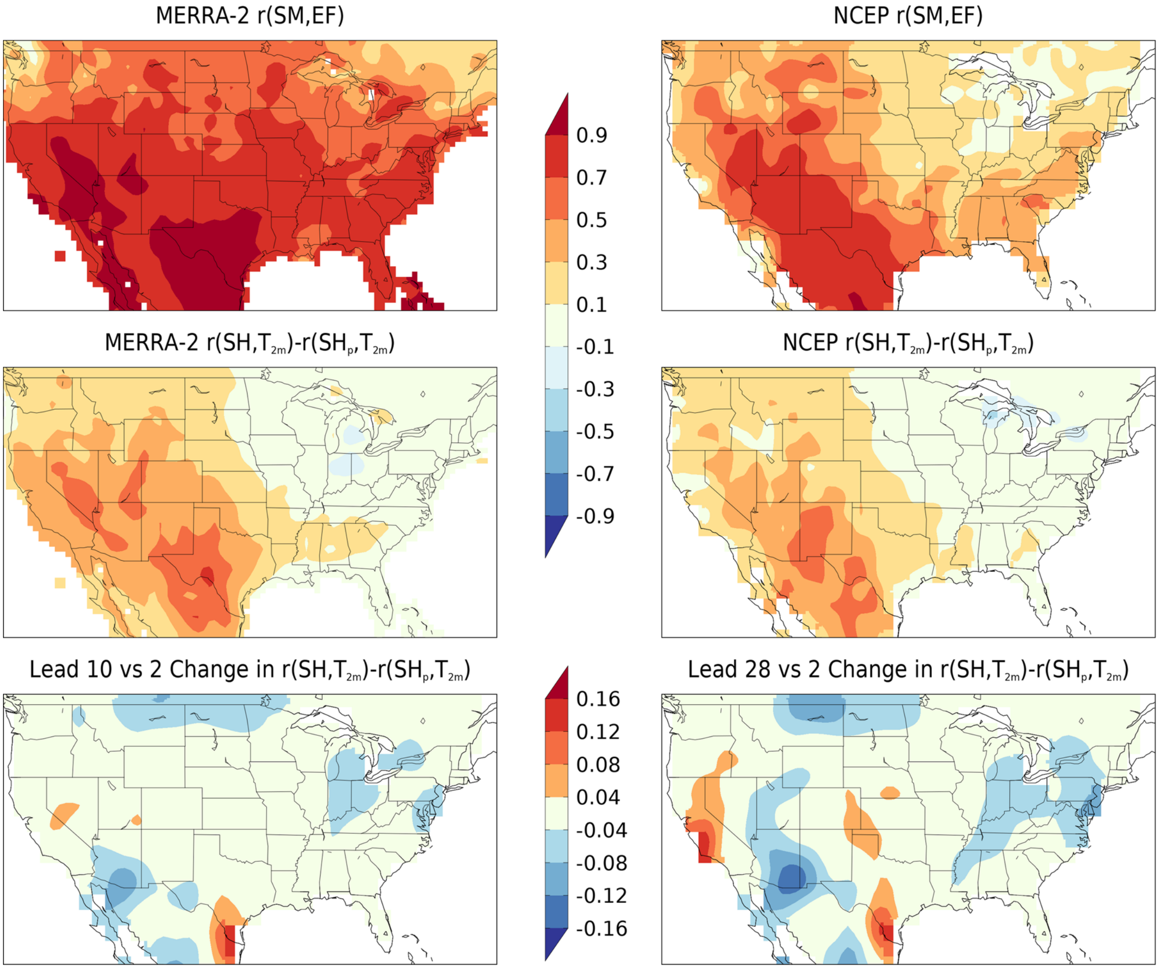
MERRA-2 r(SM,EF)
NCEP r(SM,EF)
0.9
0.7
0.5
0.3
0.1
-0.1
-0.3
-0.5
-0.7
-0.9
MERRA-2 r(SH,T2m)-r(SHp,T2m)
NCEP r(SH,T2m)-r(SHp,T2m)
Lead 10 vs 2 Change in r(SH,T2m)-r(SHp,T2m)
Lead 28 vs 2 Change in r(SH,T2m)-r(SHp,T2m)
0.16
0.12
0.08
0.04
-0.04
-0.08
-0.12
-0.16

Fig. 6 Correlation of daily near-surface soil moisture (SM) with evaporative fraction (EF) in MERRA-2 and the NCEP forecast model (top row); Miralles et al. heat wave index Π (middle row) and the change in Π in the NCEP forecast model between 10- and 2-day forecast lead (bottom left) and between 28- and 2-day forecast lead (bottom right)

daily observed $T_{\max}$, $T_{\min}$, and T_{mean} are aggregated to match the lower NCEP model resolution. Daily EHF is calculated from $T_{\max}$, $T_{\min}$, and T_{mean} within the observations and each of the four model ensemble members, separately. Daily EHF is also calculated from the four-member model ensemble mean $T_{\max}$, $T_{\min}$, and T_{mean}. This makes for a more equitable comparison, as the T_{90} threshold is relative to each individual realization.

Poisson weighting of forecasts

A series of weights in time are applied to the daily forecasts and observations (Fig. S1). Poisson function weights $W_{\lambda,k}$ are defined for validation of a forecast of a heat wave variable X as:

$$F_\lambda = \frac{\sum_{k=1}^{N} W_{\lambda,k} X_k}{\sum_{k=1}^{N} W_{\lambda,k}}\ ;\ W_{\lambda,k} = \frac{\lambda^k e^{-\lambda}}{k!} \qquad (4)$$

Where F is the time-weighted average forecast at lead λ composed from daily forecasts of X at leads $k = 1$, N. The Poisson function W also has a weight at $k = 0$, which corresponds to including the analysis of variable X as part of the weighted forecast. That is a valid approach in practice, analogous to incorporating an element of persistence, but it does not help us evaluate the forecast model behavior, so we leave that term out and renormalize F by the sum of the remaining weights. In fact, renormalization is only necessary for leads $\lambda \leq 7$; at longer leads the $k = 0$ weight becomes negligible. For $\lambda \leq 30$, $N \cong 45$ is an adequate limit, meaning at least 45-day model forecast output is needed to apply this approach for up to 30-day forecasts.

Forecast evaluations

Operational forecast skill is often evaluated using skill scores, which are relatively easy to compute and interpret. Certain skill scores are best implemented for probability forecasts of dichotomous events, such as heat wave/no heat wave, while others are better suited for verification of a deterministic prediction of heat wave/no heat wave.[38] To best evaluate model re-forecasts, both probabilistic and deterministic skill scores are employed, namely the AUC, reliability, and the ETS.

Calculation of both the AUC and ETS relies on contingency tables (supplementary Table S1), populated with hits (a), false positives (b), misses (c), and true negatives (d) from all samples. A contingency table is computed separately for each ensemble member in the AUC calculation. The hit rate (HR) and false alarm rate (FAR) for member i are computed as:

$$\text{HR}_i = \frac{a_i}{a_i + c_i} \qquad (5)$$

and

$$\text{FAR}_i = \frac{b_i}{b_i + d_i} \qquad (6)$$

A model with no skill will have an ROC curve that lies along the HR = FAR diagonal line, and therefore any positive area (AUC) between this diagonal and the actual ROC curve indicates forecast skill. Fig. S3 shows the ROC curve from the 15-day $T_{\max}$ forecast over Atlanta, Georgia; AUC is the integrated area between the model skill line and the "no skill" line. We adapt the AUC to verify deterministic (heat wave/no heat wave) forecasts from the NCEP model by establishing thresholds of the percent of model ensemble members that forecast a heat wave on a particular day at a particular lead time. We calculate HR and FAR when 1, 2, 3, and 4 out of 4 ensemble members forecast a heat wave, providing four (FAR, HR) points, to which we add a 0 and 1 to the ends. The AUC is then calculated as the integrated area between the model FAR:HR curve and the diagonal 1-to-1 line.

Model forecast reliability is assessed via the reliability diagram, which summarizes the forecast probabilities of dichotomous outcome E with the observed frequency of the occurrence of E.[39] In our case, we calculate the observed frequency of heat waves when 0, 1, 2, 3, and all 4 NCEP ensemble members forecast a heat wave, providing 5 points along which we can evaluate model reliability. Fig. S4 displays a reliability diagram for NCEP heat wave forecasts at a 10-day lead over Salt Lake City, Utah. The blue line shows the perfect forecast, when the forecasted probability and observed frequency have a 1-to-1 correspondence. We identify model forecast reliability as the integrated area between the “perfect forecast” line and the actual forecast line, shown as the thick, black line (Fig. S4). Therefore, model reliability improves as the area approaches 0; positive (negative) areas represent over- (under-) forecasting of heat waves.

The ETS, a deterministic prediction skill score, is used to evaluate the ensemble mean heat wave forecast. The ETS is calculated as:

$$\mathrm{ETS} = \frac{a - a_r}{a + b + c - a_r}, \tag{7}$$

where a_r is the expected fraction of hits for a random forecast:

$$a_r = \frac{(a + c)(a + b)}{a + b + c + d}, \tag{8}$$

We also employ KLD, also known as information divergence or relative entropy, which measures the difference between two probability distributions.[40] Here we apply it in the time dimension to determine how well model simulations of EHF agree with observations at each location, and how well the observation and reanalysis datasets agree with each other (Fig. S5). For two probability distributions p and q:

$$K_{p,q} = \sum_t p(t) ln \frac{p(t)}{q(t)} \tag{9}$$

provided the areas under the time series are normalized so that

$$\sum_t p(t) = \sum_t q(t) = 1 \tag{10}$$

and p and q are non-zero over the same domain. To fulfill this last requirement, values of the excessive heat metrics during the time domain of the heat wave season are set to 10^{-4} when there is no heat wave. $K_{p,q} \geq 0$ and $K_{p,q} = 0$ only if the distributions of p and q are identical. $K_{p,q} \neq K_{q,p}$ but ranking is preserved, as KLD is invariant to linear and nonlinear transformations. In our validations, p represents EHF from observations and q is EHF from the model forecast at any particular lead time. The duration of the heat wave season across CONUS (Fig. S6) is estimated as the difference in days between the earliest and latest date of occurrence in any year of a positive EHF value in T_{mean} across the 36 observation years.

Data availability

Model re-forecast datasets are available via the ECMWF S2S repository (http://apps.ecmwf.int/datasets/data/s2s-reforecasts-daily-averaged-ecmf/levtype=sfc/type=cf/). Validation datasets, CPC and MERRA-2, are available via the NOAA ESRL climate data repository (https://www.esrl.noaa.gov/psd/data/gridded/data.cpc.globaltemp.html) and NASA GES DISC repository (https://disc.gsfc.nasa.gov/datasets?page=1&keywords=MERRA-2), respectively.

ACKNOWLEDGEMENTS

This work is funded by the Modeling, Analysis, Predictions, and Projections (MAPP) Program of the National Oceanographic and Atmospheric Administration (NOAA) under its Subseasonal to Seasonal (S2S) Prediction effort. T.W.F. is supported by National Oceanic and Atmospheric Administration grant NA16OAR4310066, and P.A.D. and D.O.B. are supported by National Oceanic and Atmospheric Administration grant NA16OAR4310095. We thank Benjamin Cash and Jennifer Adams of COLA for their efforts toward retrieval and curation of the subseasonal prediction data.

AUTHOR CONTRIBUTIONS

T.W.F. and P.A.D. jointly designed the study. T.W.F. performed all ETS, AUC and conditional probability analyses. P.A.D. developed the Poisson weighting methodology and performed the KLD-based analyses. D.O.B. performed the correlation analysis. T.W.F. and P.A.D. wrote the manuscript and all co-authors participated in the discussion and interpretation of results.

ADDITIONAL INFORMATION

Supplementary information accompanies the paper on the *npj Climate and Atmospheric Science* website (https://doi.org/10.1038/s41612-018-0027-7).

Competing interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Semenze, J. C. et al. Heat-related deaths during the July 1995 heat wave in Chicago. *N. Engl. J. Med.* **335**, 84–90 (1996).
2. Vanos, J. K., Kalkstein, L. S. & Sanford, T. J. Detecting synoptic warming trends across the US Midwest and implications to human health and heat-related mortality. *Int. J. Climatol.* **35**, 85–96 (2015).
3. Della-Marta, P. M., Haylock, M. R., Luterbacher, J. & Wanner, H. Doubled length of western European summer heat waves since 1880. *J. Geophys. Res.* **112**, D15 (2007).
4. Kuglitsch, F. G. et al. Heat wave changes in the eastern Mediterranean since 1960. *Geophys. Res. Lett.* **37**, 4 (2010).
5. Keelings, D. & Waylen, P. Increased risk of heat waves in Florida: characterizing changes in bivariate heat wave risk using extreme value analysis. *Appl. Geogr.* **46**, 90–97 (2014).
6. Meehl, G. A. & Tebaldi, C. More intense, more frequent, and longer lastingheat waves in the 21st century. *Science* **305**, 994–997 (2004).
7. Perkins-Kirkpatrick, S. E. & Gibson, P. B. Changes in regional heatwave characteristics as a function of increasing global temperature. *Sci. Rep.* **7**, 12256 (2017).
8. Beniston, M. & Diaz, H. F. The 2003 heat wave as an example of summers in a greenhouse climate? Observations and climate model simulations for Basel, Switzerland. *Glob. Planet. Change* **44**, 73–81 (2004).
9. Fischer, E. M. & Schär, C. Consistent geographical patterns of changes in high-impact European heatwaves. *Nat. Geosci.* **3**, 398–403 (2010).
10. Mueller, B. & Seneviratne, S. I. Hot days induced by precipitation deficits at the global scale. *Proc. Natl Acad. Sci. USA* **109**, 12398–12403 (2012).
11. Hirsch, A. L., Pitman, A. J., Seneviratne, S. I., Evans, J. P. & Haverd, V. Summertime maximum and minimum temperature coupling asymmetry over Australia determined using WRF. *Geophys. Res. Lett.* **41**, 1546–1552 (2014).
12. Ford, T. W. & Schoof, J. T. Characterizing extreme and oppressive heat waves in Illinois. *J. Geophys. Res.* **122**, 682–698 (2017).
13. Quesada, B., Vautard, R., Yiou, P., Hirschi, M. & Seneviratne, S. I. Asymmetric European summer heat predictability from wet and dry southern winters and springs. *Nat. Clim. Change* **2**, 736–741 (2012).
14. Orth, R. & Seneviratne, S. I. Predictability of soil moisture and streamflow on subseasonal timescales: a case study. *J. Geophys. Res.* **118**, 10963–10979 (2013).
15. Dirmeyer, P. A. & Halder, S. Sensitivity of surface fluxes and atmospheric boundary layer properties to initial soil moisture variations in CFSv2. *Wea. Fcst.* **31**, 1973–1983 (2016).
16. Dirmeyer, P.A. & Halder, S. Sensitivity of surface fluxes and atmospheric boundary layer properties to initial soil moisture variations in CFSv2. *Wea. Fcst.* **31**, 1973–1983 (2016).
17. Koster, R. D. et al. GLACE: the global land–atmosphere coupling experiment. Part I: overview. *J. Hydrometeor.* **7**, 591–610 (2006).
18. Koster, R. D. et al. The second phase of the global land–atmosphere coupling experiment: soil moisture contributions to subseasonal forecast skill. *J. Hydrometeor.* **12**, 805–822 (2011).
19. Perkins, S. E. A review on the scientific understanding of heatwaves—their measurement, driving mechanisms, and changes at the global scale. *Atmos. Res.* **164-165**, 242–267 (2015).
20. Saha, S. et al. The NCEP climate forecast system version 2. *J. Clim.* **27**, 2185–2208 (2014).
21. Anderson, G. B. & Bell, M. L. Heat waves in the United States: mortality risk during heat waves and effect modification by heat wave characteristics in 43 U.S. communities. *Environ. Health Perspect.* **119**, 210–218 (2011).
22. Langlois, N., Herbst, J., Mason, K. & Nairn, J. Using the Excess Heat Factor (EHF) to predict the risk of heat related deaths. *J. Forensic. Leg. Med.* **20**, 408–411 (2013).
23. Miralles, D. G., Teuling, A. J., van Heerwaarden, C. C. & Vilà-Guerau, J. Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation. *Nat. Geosci.* **7**, 345–349 (2014).

24. Dirmeyer, P. A., Peters-Lidard, C. & Balsamo, G. in *Seamless Prediction of the Earth System: From Minutes to* Months (eds Brunet, G. et al.) Ch. 8 (WMO, Geneva, 2015).
25. Miralles, D. G., van den Berg, M. J., Teuling, A. J. & de Jeu, R. A. M. Soil moisture-temperature coupling: a multiscale observational analysis. *Geophys. Res. Lett*. **39**: L21707 (2012).
26. Sippel, S. Refining multi-model projections of temperature extremes by evaluation against land-atmosphere coupling diagnostics. *Earth Syst. Dynam*. **8**, 387–403 (2017).
27. Steadman, R. G. A universal scale of apparent temperature. *J. Clim. Appl. Meteorol.* **23**, 1674–1687 (1984).
28. Climate Prediction Center. CPC Global Daily Temperature Analysis. ftp://ftp.cpc.ncep.noaa.gov/precip/wd52ws/global_temp/ (2017).
29. Bosilovich, M. G. et al. *MERRA-2: Initial Evaluation of the Climate. Series on Global Modeling and Data Assimilation*.
30. Reichle, R. H. et al. Land surface precipitation in MERRA-2. *J. Clim.* **30**, 1643–1664 (2017a).
31. Reichle, R. H. et al. Assessment of MERRA-2 land surface hydrology estimates. *J. Clim.* **30**, 2937–2960 (2017b).
32. Griffies, S. M., Harrison, M. J., Pacanowski, R. C. & Rosati, A. *A Technical Guide to MOM4*. Technical Report No. **5**, 371 (GFDL Ocean Group, 2004).
33. Ek, M. B. et al. Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J. Geophys. Res.* **108**, D22 (2003).
34. Zobler, L. A. *World Soil File for Global Climate Modeling*. Vol. **87802** of NASA technical memorandum, 1–32 (National Aeronautics and Space Administration, Goddard Space Flight Center, Institute for Space Studies, Greenbelt, MD, 1986).
35. Vitart, F. et al. The Subseasonal to Seasonal (S2S) prediction project database. *Bull. Am. Meteor. Soc.* **98**, 163–176 (2017).
36. Nairn, J., Fawcett, R. & Ray, D. Defining and predicting excessive heat events, a national system. in *Understanding High Impact Weather: Extended Abstract of Third CAWCR Modelling Workshop* Vol. **30**, 83–86 (Australian Bureau of Meteorology, Melbourne, Australia, 2009).
37. Perkins, S. E., Alexander, L. V. & Nairn, J. R. Increasing frequency, intensity, and duration of observed global heatwaves and warm spells. *Geophys. Res. Lett.* **39**, L20714 (2012).
38. Wilks, D. S. A skill score based on economic value for probability forecasts. *Meteorol. Appl.* **8**, 209–219 (2001).
39. Weisheimer, A., Schaller, N., O'Reilly, C., MacLeod, D. A. & Palmer, T. 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. J. R. Meteorol. Soc.* **143**, 917–926 (2017).
40. DelSole, T. & Tippett, M. K. Predictability: recent insights from information theory. *Rev. Geophys.* **45**, RG4002 (2007).