Intelligent Climate and Eco-Environment 1 (2026) 100002

Contents lists available at ScienceDirect

Intelligent Climate and Eco-Environment

journal homepage: www.keaipublishing.com/en/journals/icee/

# AI-enhanced subseasonal forecasting of extreme temperature risks

Jia Xing[a], Siwei Li[b,c,*], Shuxin Zheng[d], Ge Song[b], Jiaxin Dong[b], Gonzalo A. Ferrada[a,e], Tie-Yan Liu[d], Joshua S. Fu[a,**]

[a] Department of Civil and Environmental Engineering, the University of Tennessee, Knoxville, TN 37996, USA
[b] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
[c] Perception and Effectiveness Assessment for Carbon-neutrality Efforts, Engineering Research Center of Ministry of Education, Institute for Carbon Neutrality, Wuhan University, Wuhan 430072, China
[d] Zhongguancun Academy, Zhongguancun Institute of Artificial Intelligence, Beijing 100080, China
[e] Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado, Boulder, CO 80309, USA

ARTICLE INFO

*Keywords:*
Machine learning
Subseasonal forecasting
Weather
Temperature

ABSTRACT

Sub-seasonal weather prediction remains a significant scientific challenge due to the chaotic nature of the atmosphere, with current numerical and AI-driven models exhibiting limited skill, particularly at the fine spatial scales for human exposure, agriculture, and infrastructure. Here, we introduce DeepMet, a high-resolution, AI-driven sub-seasonal forecasting system designed to improve the prediction of temperature extremes and their associated health risks, demonstrated successfully over the continental United States. Specifically, DeepMet substantially outperforms the benchmark of European Centre for Medium-Range Weather Forecasts, reducing the root mean square error by 20–60 % for key surface variables, including daily maximum and minimum 2-meter temperature, specific humidity, and 10-meter wind speed. The model also improves the detection of extreme heat and cold events by over 40 % across all evaluation metrics. By enhancing early warning capabilities, DeepMet enables more accurate identification of extreme weather conditions, potentially improving risk communication to prevent additional extreme-weather related deaths in the United States. Remarkably, such performance is achieved using only a single GPU for training, making the method highly accessible for local agencies to enhance early warning systems and protect public health. This underscores its strong potential to transform long-range forecasting and significantly enhance public health preparedness in a changing climate.

## Introduction

In the context of climate change, extreme weather events are becoming more frequent and increasingly threaten human health and living conditions [1,2]. Among all weather-related hazards, extreme temperatures associated with heatwaves and cold spells are the leading cause of mortality, contributing to over five million deaths globally each year [3–6]. Early warning systems, especially those extending to the sub-seasonal timescale, are essential for improving preparedness [7]. Numerous previous studies have been conducted to predict extreme temperatures at sub-seasonal scales [8–10]. In recognition of the importance of early warnings, the United Nations launched the Early Warnings for All initiative [11], aiming to ensure that every person on Earth is protected from hazardous weather, water, or climate events through life-saving early warning systems. Apparently, accurate and timely forecasts enable proactive healthcare planning, effective risk communication, and efficient resource allocation, particularly for vulnerable populations such as the elderly, children, and individuals with chronic illnesses.

Traditional numerical models face significant challenges in sub-seasonal forecasting due to error propagation across time and space, stemming from the inherently chaotic nature of the atmosphere [12]. While AI-driven approaches have shown promise, they are mostly constrained to short-term forecasts with limited skill beyond two weeks [13–16], due to the challenge of effectively balancing focus across the multi-dimensional atmospheric system, even with substantial computational resources. Moreover, many of these models do not focus on key surface-level variables, and forecasting typically global in scale with

* Corresponding author at: School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China.
** Corresponding author.
*E-mail addresses:* siwei.li@whu.edu.cn (S. Li), jsfu@utk.edu (J.S. Fu).

https://doi.org/10.1016/j.icee.2025.100002
Received 6 September 2025; Received in revised form 17 October 2025; Accepted 4 November 2025

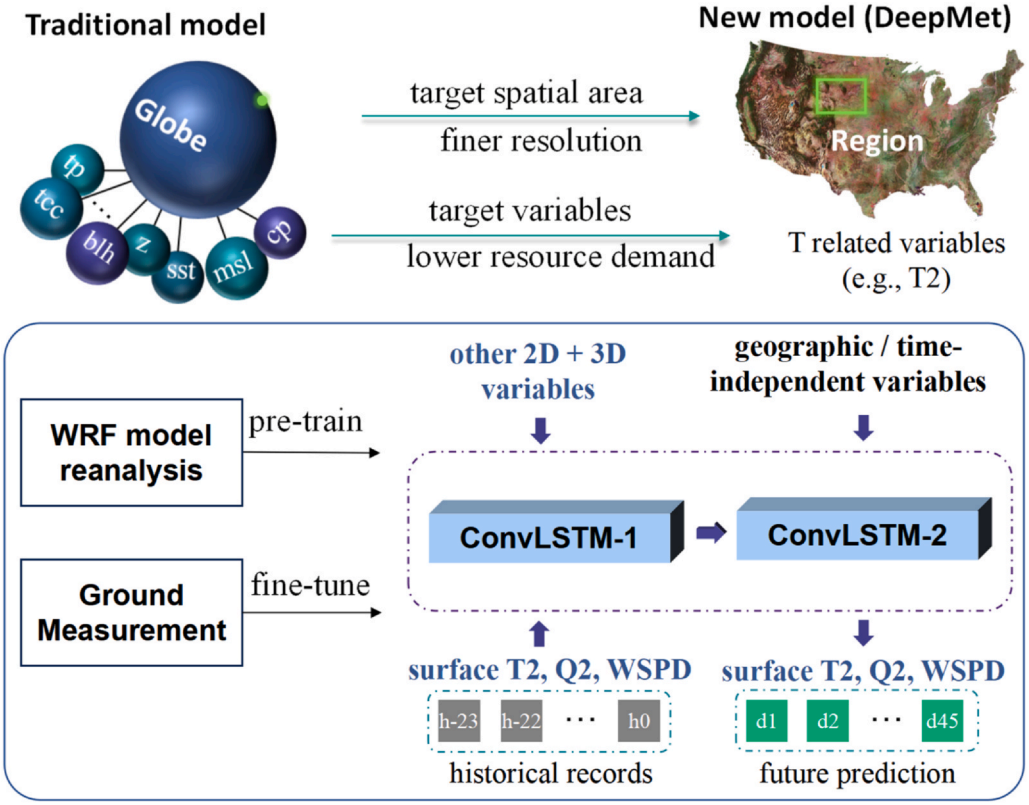


**Fig. 1.** Framework and advantages of the DeepMet model.

coarse spatial resolution [17], making it difficult to incorporate accurate ground-based observations due to the high spatial heterogeneity of surface conditions. To better support public health applications operating with minimal computational cost for local agency, there is a growing need for high-resolution, regionally focused weather forecasting models that emphasize surface variables relevant to human thermal stress over extended temporal horizons.

To address the limitations mentioned above, we extend AI-based forecasting to the high-resolution sub-seasonal scale (noted as DeepMet, see Fig. 1), with a focus on surface variables that are critical for assessing and managing the increasing risks associated with temperature extremes. The novelty of this study lies in three key aspects, which can be summarized as follows.

First, for the training dataset, unlike global-scale AI training models that primarily rely on global datasets which often limited by coarse spatial resolution, we leverage multi-year dynamical downscaling using a numerical weather model, incorporating abundant ground-based and upper-atmosphere historical observations through Four-Dimensional Data Assimilation [18]. This approach produces regional forecasts at a 12 km × 12 km resolution ten times finer than the widely used ERA5 dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF Reanalysis v5) [19] making it significantly more suitable for assessing human exposure. Additionally, it enables fine-tuning of the model with high-quality ground observations, resulting in improved forecasts of surface variables that are more consistent with ground measurements than those from reanalysis datasets.

Second, for the feature selection, we developed a deep learning architecture for meteorological forecasting that is both streamlined and computationally efficient, running on a single NVIDIA A100 GPU and reducing hardware demands by up to 60-fold compared to traditional multi-GPU systems. By avoiding unnecessary global-scale predictions of numerous unrelated factors for localized applications, DeepMet concentrates on key variables relevant to temperature extremes. Specifically, it targets daily maximum 2 m temperature (T2max) and specific humidity (Q2) for heatwaves, and daily minimum 2 m temperature (T2min) combined with wind speed for cold events, which are core components of widely used public health indices such as the Heat Index [20] and Wind Chill Index [21]. This low-cost design enables more efficient support for local agencies, allowing them to develop improved localized forecasting systems with limited resources.

Third, building on our previous findings regarding the importance of memory structures for long-term forecasting [22], we implement a ConvLSTM-based model [23] (Figure S1) that utilizes the past 24 h of multi-variable inputs to predict daily variations up to 45 days ahead aligning with the global sub-seasonal to seasonal (S2S) forecasting scale used by systems such as ECMWF. This architecture mitigates the chaotic nature of the atmosphere by leveraging temporal memory and optimizing the overall evolution of weather patterns through a statistical machine learning process, which helps constrain error propagation over time, which is an essential feature for achieving reliable long-term predictions in S2S systems.

Together, these innovations enable our approach to deliver high-resolution sub-seasonal forecasts with enhanced accuracy in predicting temperature extremes—all while operating at minimal computational cost, as detailed in the following section.

## Methods

### *Dynamical downscaling with numerical model*

To better represent the human exposure to the extreme temperatures, we leverage a mesoscale numerical weather prediction system which is the Weather Research and Forecasting (WRF) [24] simulations dynamically downscaled to a regional scale with a 12 km resolution, which is approximately 10 times finer than the original 1.5 by 1.5 ° resolution (about 110–160 km) of the global ECMWF S2S dataset.

The WRF model was configured following the setup used in our previous study [25], including the Morrison two-moment microphysics

scheme; the Rapid Radiative Transfer Model for Global Climate Models (RRTMG) for both longwave and shortwave radiation scheme; the Yonsei University (YSU) planetary boundary layer (PBL) scheme; the Pleim–Xiu land surface model; the revised MM5 (Jimenez) surface layer scheme; and Grell–Freitas (GF), with a radiative feedback cumulus parameterization option. The WRF model simulations were driven by the North American Mesoscale (NAM) model analyses from the National Centers for Environmental Prediction (NCEP), incorporating four-dimensional data assimilation (FDDA) with both surface and upper-air observations. Observation nudging was applied using NCEP's Automated Data Processing (ADP) Global Surface and Upper-Air Observational Weather Data.

For comparison, we obtained ECMWF sub-seasonal forecast data spanning six years from the open-access repository. The dataset includes daily control run forecasts of T2, T2max, T2min, U10, V10, and pressure-level specific humidity (q) with a lead time of up to 46 days. The original data, provided at a 1.5° × 1.5° resolution, were re-gridded to the target domain using bilinear interpolation (a commonly used and physically consistent approach for regridding meteorological fields) to facilitate comparison.

Ground-based observational data were obtained from the NOAA NCDC ISD-Lite archive, which provides hourly records of T2, Q2, and 10-meter wind speed (WSPD10) from approximately 6775 sites across the U.S. domain.

The downscaled WRF model presents better agreement with the ground-based measurements from NCDC (Figure S1), providing a reliable foundation for training DeepMet and enabling it to deliver improvement over global forecasting models such as ECMWF. We adopted the WRF simulation dataset as our reference because it provides fine-scale representation suitable for evaluating local and regional variations, particularly for applications related to human exposure assessment which is a key focus and novelty of this study.

*DeepMet model structure*

The DeepMet model is built upon a ConvLSTM-based architecture (Fig. 2), which inherently captures temporal dependencies through its recurrent structure. The model performs a direct multi-step forecast for the next 45 days of each predictive variable (i.e., T2max, T2min, Q2, and WSPD10) within a single forward pass, rather than a recursive autoregressive prediction. In other words, the model takes the historical inputs once and simultaneously predicts the entire 45-day sequence. Although the prediction is not recursively generated, the ConvLSTM cells internally propagate temporal information across multiple time steps during training, allowing the model to learn temporal evolution patterns. Moreover, because the loss function is computed over all 45 forecast days, the model jointly optimizes performance across the entire sequence, which helps mitigate error accumulation commonly observed in traditional autoregressive approaches.

More specifically, consistent with our previous applications in atmospheric chemistry forecasting, the DeepMet model architecture integrates two ConvLSTM modules, each consisting of three layers with varying channel sizes (256, 128, and 64) and a 3 × 3 kernel. The first module processes historical records from the past 24 h, incorporating multiple variables, such as 2D and 3D inputs from model reanalysis and ground-based measurements to extract key information relevant to forecasting future variations of the target variable. This processed historical data is then passed to a second ConvLSTM module to generate the forecast, along with time-independent variables (i.e., geographical features and climatological fields, listed in Table 1), but not the time-dependent 2D or 3D variables from future steps. During prediction, the hidden and cell states (h, c) are dynamically updated, enabling the model to retain and leverage long-term historical dependencies. This design is consistent with our prior work in atmospheric chemistry, where predictions at earlier time steps are recursively used as inputs for future steps. The methodology effectively captures the dual role of meteorological factors: their gradual modulation of baseline conditions over time (Role 1), and their direct interaction with other variables (Role 2), both of which are critical in shaping future atmospheric states. Only the target prediction variables are treated in an autoregressive manner, while the other meteorological predictors are not recursively predicted by the model. This design allows DeepMet to generate a direct 45-day forecast based on historical sequences without relying on unavailable future reanalysis inputs, thus maintaining a valid forecasting setup.

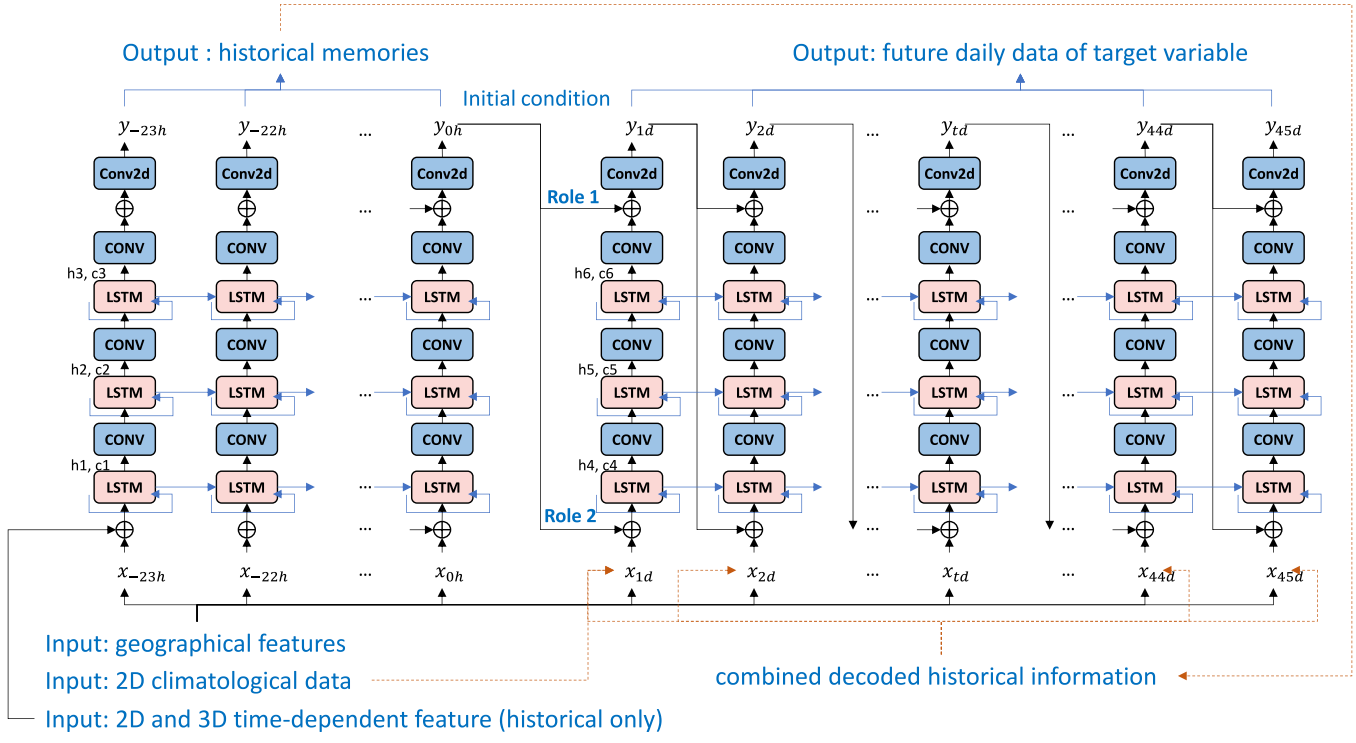


**Fig. 2.** Model architecture of DeepMet based on a dual-ConvLSTM structure for direct 45-day multi-step forecasting.

**Table 1**
Model feature for each prediction with DeepMet.

| Predict variable | Time independent | | Time-dependent | |
|---|---|---|---|---|
| | Geographic factors | 2D features (climatological data) | 2D Feature (historical only) | 3D Feature (historical only) |
| T2max | DLUSE, HT, LWMASK, MSFX2, LUFRAC, PURB, LAT, LON | LAI, VEG, ALBEDO, SWDNBC | UWIND, VWIND, CFRAC, PBL, prep, Q2, RGRND, T2, WSPD10, WSTAR, HFX, LH, USTAR, ZRUF, PRSFC, WBAR, WR, SNOCOV | TA |
| T2min | | | | TA |
| Q2 | | | | QC, QV, TA, CFRAC_3D |
| WSPD10 | | | | UW, VW |

A summary of all 2D and 3D variables used in the model is shown in Table 1. All variables are derived from the WRF reanalysis dataset. The geographical features used in this study consist of eight time-invariant variables: dominant land use category (DLUSE), terrain elevation (HT), land–water mask (LWMASK), map-scale factor squared (MSFX2), land use fraction (LUFRAC), percentage of urban area (PURB), latitude (LAT), and longitude (LON). These static features are consistently applied throughout the entire prediction period.

Four 2D surface characteristics, Leaf Area Index (LAI), vegetation cover (VEG), albedo, and clear-sky shortwave downwelling radiation (SWDNBC), are prescribed from climatological datasets, given their relatively stable annual cycles (Figure S2). We computed the mean of all historical years for each corresponding calendar day, enabling their use as inputs under both historical and future conditions. This approach ensures that the model's predictive performance is not influenced by potential uncertainties in forecasting these surface parameters. In contrast, physical based models like ECMWF model can dynamically predict these variables through its land surface and radiation schemes, where their accuracy depends on the performance of the underlying land-use and radiative transfer models. To maintain a fair comparison, we deliberately chose to use prescribed versions of these variables in DeepMet rather than predicted ones.

In addition, 18 time-dependent 2D meteorological variables are incorporated into the historical module. These include: U- and V-component winds (UV-wind), cloud fraction (CFRAC), planetary boundary layer height (PBL), precipitation (PREP), 2-meter specific humidity (Q2), shortwave radiation at the ground surface (RGRND), 2-meter temperature (T2), 10-meter wind speed (WSPD10), convective velocity scale (WSTAR), sensible heat flux (HFX), latent heat flux (LH), cell-averaged friction velocity (USTAR), surface roughness length (ZRUF), surface pressure (PRSFC), average liquid water content of clouds (WBAR), canopy moisture content (W*R*), and snow cover (SNOCOV).

Furthermore, six 3D meteorological variables are used, including air temperature (TA), cloud water mixing ratio (QC), water vapor mixing ratio (QV), three-dimensional cloud fraction (CFRAC_3D), and the U and V components of wind (UW, VW). These variables are resolved across 35 vertical levels, from the surface up to 100 mb, and incorporated into the historical module to enhance the model's ability to capture vertical atmospheric structure. All these variables are integrated within the dual-ConvLSTM structure of DeepMet, as illustrated in Fig. 2.

*DeepMet training and testing*

We leverage multiple WRF model simulations for pre-training and incorporate ground-based measurements from the NCDC dataset to enhance model performance at the surface level. The model is initially trained using data from five historical years (2008, 2012, 2014, 2019, and 2021, randomly selected from the historical period, due to the limited computational resources and memory), followed by fine-tuning with NCDC ground-based measurements. This strategy helps mitigate sample imbalance in the observational data by avoiding direct training solely on the ground measurements [26]. Model performance is evaluated using data from the year of 2023, emulating the strategy of leveraging historical data to forecast recent conditions. The selected years were chosen based on data availability from the CONUS 12 km WRF simulation dataset. Due to computational constraints, continuous simulations for all intermediate years were not available for this study. However, future work will incorporate additional historical years to further expand the training dataset and enhance model robustness.

Given the large dataset size and limited computational resources, particularly RAM constraints, we adopted a subset training strategy. Specifically, the model was trained on batches of 20 subsets of all approximately 1500 samples, over a total of 4000 epochs, preventing the need to load the entire dataset into memory simultaneously. For data augmentation, random cropping was applied to the feature maps, resizing them to 120 × 120 grid cells. The model was trained using the Mean Squared Error loss function. The learning rate was initialized at 0.0001 and linearly decayed to zero over the course of training. The Adam optimizer was employed to improve model convergence and stability [27]. The DeepMet model for each predictive variable (i.e., T2max, T2min, Q2, and WSPD10) using a single GPU with approximately 24 h of training time.

We evaluate the performance of the DeepMet model using several metrics: the Anomaly Correlation Coefficient (ACC), Structural Similarity Index Measure (SSIM), Root Mean Square Error (RMSE), and the Ranked Probability Skill Score (RPSS). Spatially averaged statistics are computed by comparing DeepMet predictions with both ECMWF forecasts and WRF downscaled meteorological reanalysis data. In addition, RPSS is used to evaluate forecast skill against surface-level observations from the NCDC ground-based measurement network.

To evaluate the heatwave and cold event prediction performance, we use four key metrics: F1 Score, Critical Success Index (CSI), Probability of Detection (POD), and False Alarm Rate (FAR). These metrics assess the models' ability to accurately detect impactful extreme temperature events, which are defined using T2max and Q2 for heatwaves, and T2min and 10-meter wind speed (WSPD10) for cold spells.

In this study, we focused on the model's performance for individual high-temperature or low-temperature days, rather than multi-day heatwave sequences. Specifically, a heatwave event is defined as T2max $>$ 32 °C and Q2 $>$ 0.014 kg/kg, while a cold event is defined as T2min $<$ 0 °C with WSPD10 $>$ 3 m s$^{-1}$. These thresholds are applied on a daily basis but can also be part of multi-day events or percentile-based definitions [20,21,28]. The criteria were adapted to represent extreme temperature conditions over the CONUS domain in our analysis.

*Health impact assessment*

To quantify the potential public health benefits of improved early warning provide by DeepMet, we estimated the number of individuals accurately identified as exposed to heatwaves by DeepMet compared to ECMWF. Population data were obtained from the Gridded Population of the World, Version 4 (GPWv4), for the most recent year available (2020) [29], at a spatial resolution of 2.5 arc-minutes, and regridded to match the target domain.

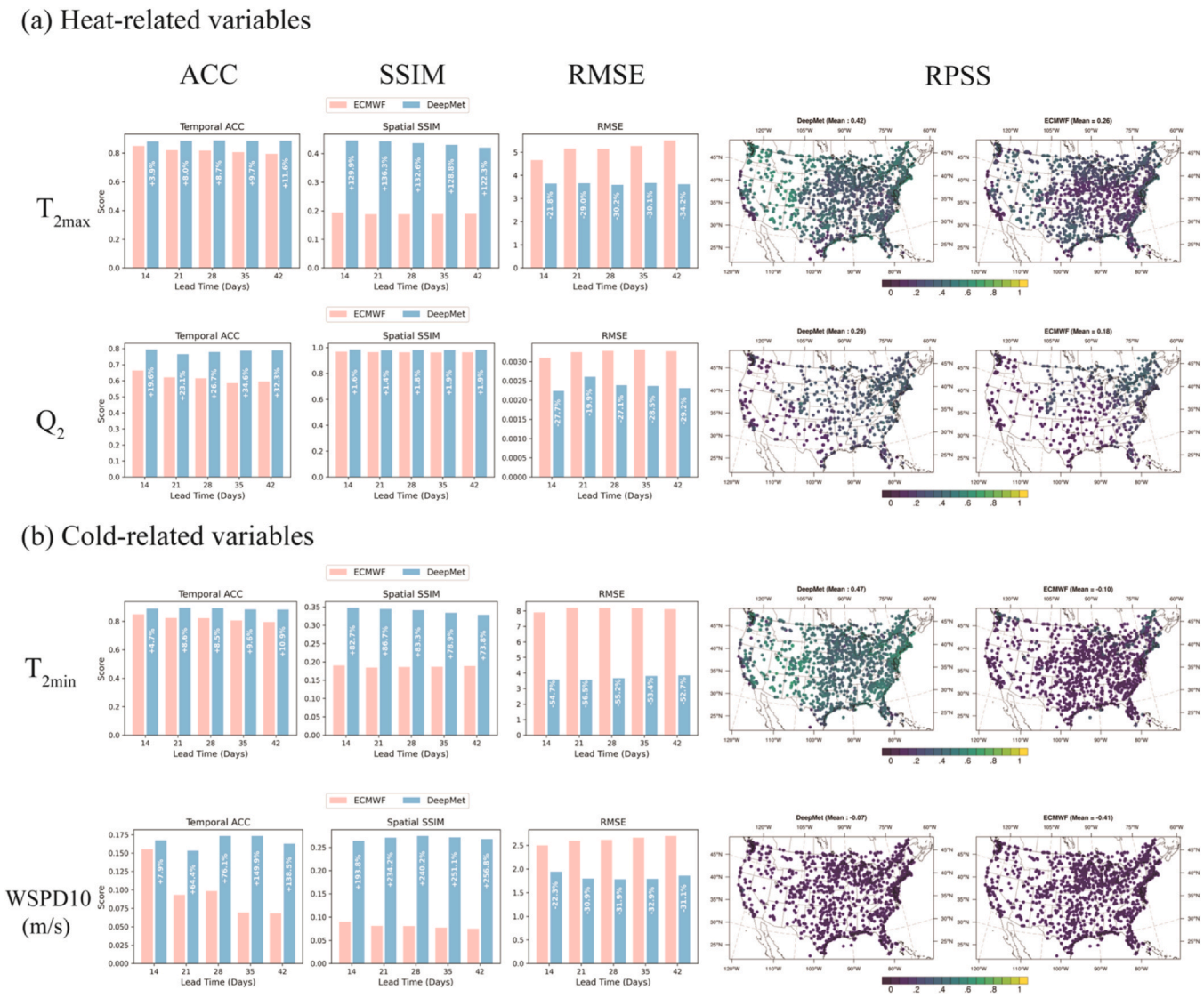


**Fig. 3.** Comparison of DeepMet and ECMWF predictions on crucial meteorological variables relate to extreme temperature at S2S scale (the percentage number shown in the blue bar represent the change in each metrics from ECMWF to DeepMet; the color bar is corresponding to each RPSS comparison).

## Results

### *Improved sub-seasonal prediction on crucial ground meteorological variables*

DeepMet demonstrates significant improvements in sub-seasonal forecasting of key surface meteorological variables, as illustrated in Fig. 3. Using dynamically downscaled reanalysis fields as ground truth, DeepMet outperforms a typical physics-based S2S forecast system (benchmarked against ECMWF) across the critical forecast range of 14–42 days. It not only captures the magnitude of meteorological variables more accurately but also more effectively reproduces their temporal evolution and spatial structure at high resolution.

Specifically, DeepMet increases the temporal anomaly correlation coefficient (ACC) by 4–11 % for T2max and T2min, 20–32 % for Q2, and up to 138 % for WSPD10 (indicted by the percentage number in each blue bar in Fig. 3), reflecting a stronger agreement with observed temporal variability, independent of mean bias, comparing to the ECMWF benchmark. Notably, these improvements become more pronounced increasing lead time, underscoring DeepMet's effectiveness in mitigating error propagation within the inherently chaotic S2S forecasting regime.

In terms of spatial accuracy, DeepMet substantially improves the Structural Similarity Index (SSIM) by 80–250 % for T2max, T2min and WSPD10 (indicted by the percentage number in each blue bar in Fig. 3), demonstrating enhanced spatial fidelity in structure, luminance, and contrast, largely attributable to its finer spatial resolution compared to global systems (as ECMWF). Additionally, DeepMet reduces RMSE by approximately 20–60 % across all four variables, indicating a substantial decrease in average prediction error and closer alignment with observed data. The differences in performances between the two models are also statistically significant, with p-values $< 0.001$ across all key metrics (Figure S3). These results confirm that the performance improvements achieved by DeepMet over ECMWF are robust and statistically significant.

We further validated DeepMet's performance using local ground-based observations from the NOAA National Climatic Data Center (NCDC) network. In addition to substantially reducing RMSE compared to ECMWF forecasts, DeepMet achieved significantly higher Ranked Probability Skill Scores (RPSS), with improvements of up to 0.5 for all four variables, demonstrating improved probabilistic forecasting skill relative to climatology (i.e., the historical average distribution). These results highlight DeepMet's effectiveness in delivering skillful and reliable S2S forecasts.

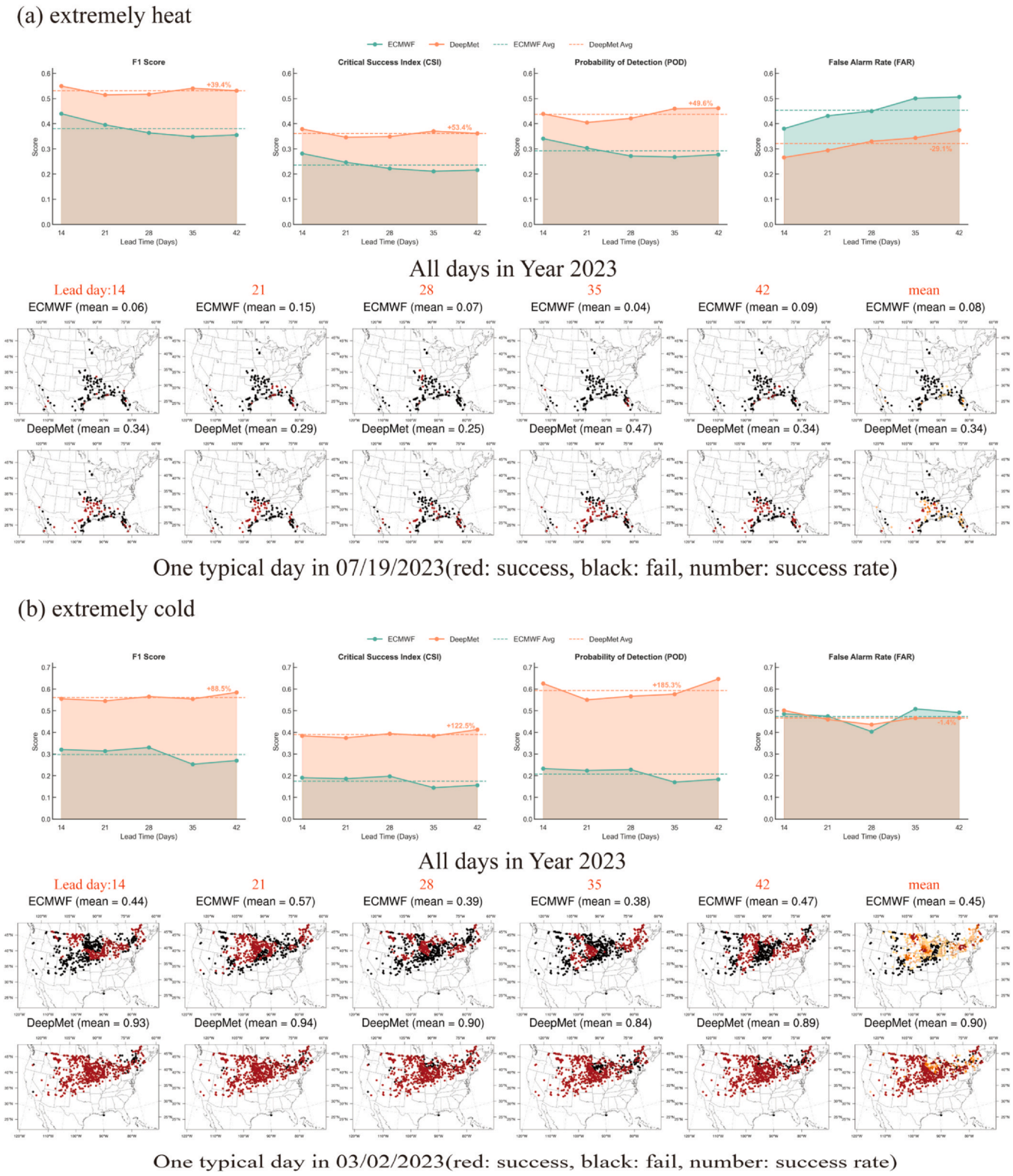


**Fig. 4.** Comparison of DeepMet and ECMWF predictions on extreme temperature events (the percentage number besides the line represent the changes from ECMWF to DeepMet).

*Enhanced predictive capability for heat-wave and cold event*

With enhanced predictability of key meteorological variables, DeepMet demonstrates superior capability in forecasting heatwaves and cold events, as shown in Fig. 4. Using NCDC observations as ground truth, DeepMet significantly outperforms ECMWF in long-range extreme temperature prediction.

Specifically, DeepMet achieves substantially higher F1 scores by 40 % for heatwaves and 90 % for cold events (indicted by the percentage number besides the line in Fig. 4), indicating a better balance between high recall (capturing true events) and high precision (reducing false alarms). The Critical Success Index (CSI) also shows marked improvement, increasing by 53 % for heatwaves and 123 % and cold events, confirming DeepMet's superior ability to detect extreme events

accurately while minimizing both misses and false positives. Moreover, the Probability of Detection (POD) increases by 50 % for heatwaves and 185 % for cold events, demonstrating that DeepMet captures significantly more true extreme events than ECMWF benchmark. Simultaneously, the False Alarm Ratio (FAR) decreases by 29 % and 1.4 % for heatwaves and cold events, respectively, resulting in fewer false alerts and more trustworthy forecasts.

We also compared the predictions of the two models for representative single-day heatwaves and cold events, using forecast lead times ranging from 14 to 42 days. The results show that DeepMet significantly outperforms ECMWF, consistently capturing extreme temperature events across the spatial domain, with 30–50 % more successful detections (indicted by the mean successful number above each spatial map in Fig. 4).

*Potential health benefits of early warning systems*

By accurately forecasting extreme heatwaves and cold events, DeepMet can significantly strengthen early warning systems, enabling communities, especially vulnerable populations and better preparing more effectively. We assessed the added value of DeepMet over ECMWF for both types of events throughout 2023. As shown in Fig. 5, substantial improvements in cold event prediction were observed across northern states, which are more frequently impacted by cold extremes. In contrast, the more significant improvements in heatwave prediction occurred in southern states, where extreme heat events are more common.

Compared to ECMWF forecasts, DeepMet enables more accurate identification of extreme weather events, potentially improving early warnings for an additional 600 million people-day during cold events and 3300 million people-day during heatwaves. This represents a 30 % increase in population coverage during cold spells and a 60 % increase during heatwaves. Given the annual estimates of 700–1300 heat-related deaths and 1200–2000 cold-related deaths in the U.S. [30–32], enhanced preparedness enabled by DeepMet such as timely healthcare interventions or emergency services could help prevent considerable deaths annually. These benefits become even more substantial as extreme heat and cold events intensify under a business-as-usual climate change scenario [32], potentially preventing additional premature deaths and avoiding associated economic losses each year, not to mention the additional risks posed to agriculture, infrastructure, and broader societal systems. These findings highlight the tangible value of more accurate and timely sub-seasonal extreme temperature forecasts.

## Discussion and conclusion

This study demonstrates a successful application of AI in extending the temporal scale of weather forecasts, an area that has traditionally posed significant challenges for numerical models. The key to DeepMet's success lies in its ability to incorporate time-series memory into the prediction process, enabling the model to effectively constrain error propagation over time. Although traditional numerical methods retain memory from the previous time step, their atmospheric initial conditions rarely extend beyond 1–2 weeks due to chaotic dynamics. They also rely heavily on step-by-step progression, making them prone to cumulative errors. DeepMet leverages information from multiple previous time steps, thereby reducing the risk of runaway inaccuracies. This finding is particularly important for forecasting highly turbulent systems such as the atmosphere, highlighting that overcoming the limitations of long-range weather prediction may require moving beyond purely mathematical approaches to models that incorporate historical memory. While DeepMet demonstrates strong performance over CONUS, its framework is generalizable and can be extended to other regions given adequate data support. Future applications should therefore consider not only global-scale modeling but also fine-scale, high-resolution implementations tailored to local climatic and observational contexts.

In the current version of DeepMet, we use the previous 24 h of data as input. That is mainly because unlike global-scale models (e.g., large foundation models) that capture spatially extensive information at a single time point and rely less on historical context, regional models such as DeepMet must account for external influences from surrounding areas that evolve over time. Incorporating multiple historical time steps allows the model to implicitly capture these broader dynamical effects. While we initially expected that incorporating longer historical records (e.g., extending to 10 days) might improve model performance, our experiments indicated otherwise. In fact, extending the input window to 10 days often resulted in reduced accuracy (Figure S4), likely because the most recent 24-hour data contains the most relevant predictive signals for S2S forecasting, while older data contributes diminishing value. Additionally, we did not include more than 24 h of high-resolution data due to the computational resource constraints (mostly due to RAM limitation), which would significantly reduce the effective size of the training dataset and compromise training efficiency. Therefore, the choice of window size represents a trade-off between predictive skill and computational efficiency. The optimal window size may also vary by region or season, depending on factors such as synoptic persistence or external forcing (e.g., teleconnections). Apparently, future work may explore methods to effectively incorporate longer historical time series into the model to systematically optimize window size based on regional characteristics and seasonality, particularly when computational resources become available.

The success of DeepMet also challenges traditional perspectives in S2S forecasting. Conventional wisdom holds that long-term predictions require global-scale models, such as ECMWF, to capture the influence of large-scale atmospheric circulation across regions. Particularly, previous studies have highlighted the importance of teleconnections and slow modes of variability such as the MJO and ENSO for predictability at S2S time scales. Though in regional scale model like DeepMet, the global effects may already be implicitly captured by the model through learning from historical information. Since large-scale impacts propagate into the target domain step by step, they are eventually reflected in related meteorological features within the domain. Thus, historical data (even at 24-hour intervals, but extendable to weekly scales) can help machine learning models capture and enhance the influence of remote drivers. Our results also demonstrate that a regional-scale S2S model can outperform global models even without explicitly representing global influences. This is probably because error propagation occurs not only temporally but also spatially. Errors originating outside the target domain can amplify and be transported into the region of interest, potentially degrading forecast accuracy. While global models are capable of capturing cross-boundary flows, they may also introduce additional uncertainty from distant regions, which can limit their benefit. Moreover, global models often face a trade-off between resolution and spatial coverage, often allocating computational resources to areas that are not directly relevant to the target region. That said, this does not imply that boundary conditions are unimportant. In fact, our experiments indicate that incorporating simplified boundary condition information improves forecast skill during the first two weeks (Figure S5). Additionally, for short-term forecasts (e.g., next-day prediction, one leading day), the global model ECMWF indeed performs better than DeepMet when compared against WRF (Figure S6-S7), which also supports the suitability of WRF as the reference dataset. While, our study focuses on the S2S forecast range, where ECMWF performance deteriorates more rapidly with lead time, while DeepMet maintains more stable performance over the full 45-day period. This trade-off between short-term accuracy and long-term stability explains why DeepMet eventually outperforms ECMWF at extended lead times. To ensure fairness, we also compared the results at the same coarse (1.5°) resolution as ECMWF forecasts and found consistent improvements by DeepMet across key metrics at S2S scales (Figure S8). This finding

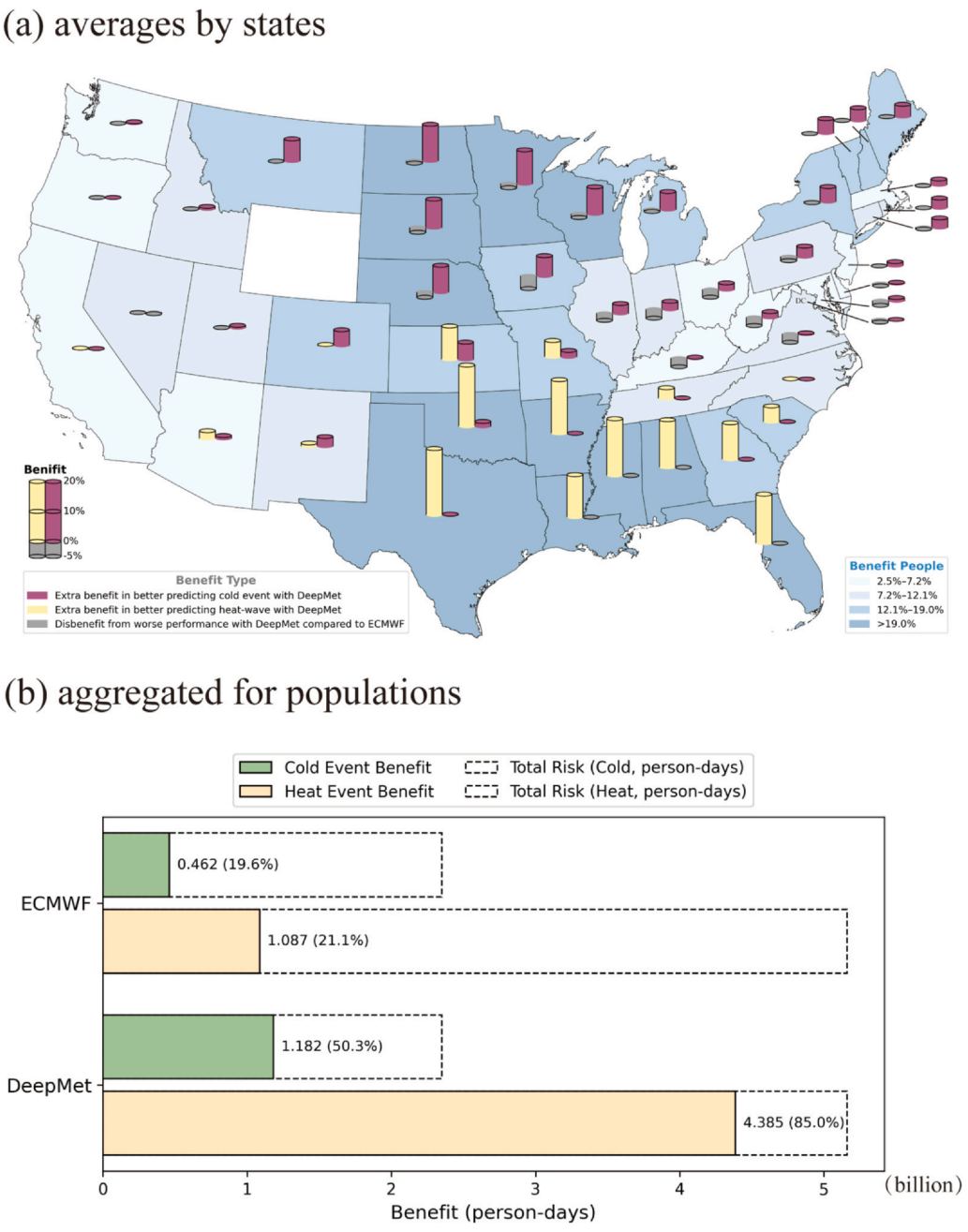


**Fig. 5.** Estimated benefits of DeepMet in identifying extreme heatwaves and cold events during 2023.

suggests that observed improvements from ECMWF to DeepMet are not driven by resolution effects alone, and that the interpolation and resolution differences do not substantially alter the main conclusions regarding DeepMet's forecast skill. Apparently, while global context is useful for short-term predictions, it may be less critical or even counterproductive for longer-term S2S forecasting.

Another challenge to traditional thinking lies in the assumption that meteorological variables are highly interdependent and should be predicted simultaneously. However, our findings suggest that this approach does not always yield better performance. Due to error propagation, inaccuracies in one variable can adversely affect the prediction of others. In contrast to other AI-based weather models that attempt to forecast multiple variables concurrently, DeepMet employs a single-variable prediction strategy. This allows the model to focus on learning the dynamics of each variable independently, without the need to balance competing priorities during optimization. It also enables the selection of more relevant 3D input features tailored to each specific variable (Table 1). Moreover, this modular approach enables parallel training across multiple GPUs, improving both efficiency and scalability while preserving task-specific accuracy.

We scaled the time series input to a daily resolution for S2S forecasting to reduce error propagation across time steps. Additionally, using hourly data would require significantly more memory, which limits the amount of training data that can be processed and ultimately hinders model performance. In our experiments, increasing the temporal resolution of predictions from daily to 3-hourly, or 6-hourly intervals which did not yield meaningful improvements (Figure S9). Therefore, daily resolution remains a practical and effective choice, especially for S2S forecasts where long lead times are the primary focus.

While large GPU resources can enhance model performance, our study demonstrates the feasibility of generating fine-scale weather forecasts using limited computational resources. This makes the

approach more realistic and accessible. Importantly, the method well-aligned with localized policy needs for protecting public health, agriculture, and infrastructure. Looking ahead, there is substantial potential to further enhance forecasting skill, both in accuracy and scope by expanding applications to other critical variables such as precipitation, wildfires, and floods, ultimately contributing to the protection of more lives.

Incorporating additional future-relevant information such as slowly varying or accurately predictable variables can further improve forecasting performance. For example, incorporating improved representations of time-series day for slowly varying geophysical inputs, such as leaf area index, vegetation cover, albedo, downward shortwave radiation, into DeepMet may lead to noticeable improvements in predictive accuracy (Figure S10). The current results can be viewed as a conservative (lower-bound) estimate of DeepMet's performance relative to ECMWF. There is no doubt that AI will play an increasingly vital role in extreme weather prediction, especially at the S2S scale, which continues to pose significant challenges for traditional numerical models.

## Inclusion & Ethics

The study does not involve human participants or sensitive data.

## Disclaimer

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

## Code availability

The DeepMet code will be made publicly available on Zenodo upon publication.

## CRediT authorship contribution statement

**Tie-Yan Liu:** Writing – review & editing, Supervision, Conceptualization. **Joshua S. Fu:** Writing – review & editing, Supervision, Conceptualization. **Jia Xing:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization. **Siwei Li:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Shuxin Zheng:** Writing – review & editing, Methodology, Conceptualization. **Ge Song:** Formal analysis. **Jiaxin Dong:** Formal analysis. **Gonzalo A. Ferrada:** Formal analysis.

## Data availability

Data will be made available on request.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The author Jia Xing is a Editor board member for Intelligent Climate and Eco-Environment and was not involved in the editorial review or the decision to publish this article.

## Acknowledgements

This work was supported by U.S. National Science Foundation [NO: 2100582], Fengyun Application Pioneering Project (FY-APP), National Natural Science Foundation of China [NO: 42375131], UT AI Tennessee Initiative Seed Funds, and MSRA collaborative research project. The author would like to acknowledge the support of the Bellagio Center Residency Program, funded by the Rockefeller Foundation. The author also acknowledges Dr. Kristen Foley and Dr. Christian Hogrefe from US EPA, Dr. Daniel Tong and Dr. Bok H. Baek from GMU for supporting the meteorological dataset.

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.icee.2025.100002.

## References

[1] L. Grant, I. Vanderkelen, L. Gudmundsson, E. Fischer, S.I. Seneviratne, W. Thiery, Global emergence of unprecedented lifetime exposure to climate extremes, Nature 641 (8062) (2025) 374–379.
[2] J. Patz, D. Campbell-Lendrum, T. Holloway, et al., Impact of regional climate change on human health, Nature 438 (2005) 310–317, https://doi.org/10.1038/nature04188.
[3] Q. Zhao, Y. Guo, T. Ye, A. Gasparrini, S. Tong, A. Overcenco, A. Urban, A. Schneider, A. Entezari, A.M. Vicedo-Cabrera, A. Zanobetti, Global, regional, and national burden of mortality associated with non-optimal ambient temperatures from 2000 to 2019: a three-stage modelling study, Lancet Planet. Health 5 (7) (2021) e415–e425.
[4] J. Berko, D.D. Ingram, S. Saha, J.D. Parker, Deaths Attributed to Heat, Cold, and other Weather Events in the United States, 2006–2010, US Department of Health and Human Services, 2014.
[5] P. Masselot, M.N. Mistry, S. Rao, et al., Estimating future heat-related and cold-related mortality under climate change, demographic and adaptation scenarios in 854 European cities, Nat. Med. 31 (2025) 1294–1302, https://doi.org/10.1038/s41591-024-03452-2.
[6] J. Yang, M. Zhou, Z. Ren, M. Li, B. Wang, D.L. Liu, C.Q. Ou, P. Yin, J. Sun, S. Tong, H. Wang, Projecting heat-related excess mortality under climate change scenarios in China, Nat. Commun. 12 (1) (2021) 1039.
[7] G. McGregor, Heatwave Responses: Early Warning Systems, Heatwaves: Causes, Springer International Publishing, Cham, 2024, pp. 549–599.
[8] F. Vitart, A.W. Robertson, The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events, npj Clim. Atmos. Sci. 1 (1) (2018) 3.
[9] H. Lin, R. Mo, F. Vitart, The 2021 western North American heatwave and its subseasonal predictions, Geophys. Res. Lett. 49 (6) (2022) e2021GL097036.
[10] J. Xie, P.C. Hsu, Y. Hu, H. Zhang, M. Ye, Advancing subseasonal surface air temperature and heat wave prediction skill in China by incorporating scale interaction in a deep learning model, Geophys. Res. Lett. 51 (20) (2024) e2024GL111076.
[11] World Meteorological Organization (WMO), EARLY WARNINGS FOR ALL: The UN Global Early Warning Initiative for the Implementation of Climate Adaptation (2022).
[12] P. Bauer, A. Thorpe, G. Brunet, The quiet revolution of numerical weather prediction, Nature 525 (7567) (2015) 47–55.
[13] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, Q. Tian, Accurate medium-range global weather forecasting with 3D neural networks, Nature 619 (7970) (2023) 533–538.
[14] Bodnar, C., Bruinsma, W.P., Lucic, A., Stanley, M., Brandstetter, J., Garvan, P., Riechert, M., Weyn, J., Dong, H., Vaughan, A. and Gupta, J.K., 2024. Aurora: A foundation model of the atmosphere. arXiv preprint arXiv:2405.13063.
[15] D. Kochkov, J. Yuval, I. Langmore, P. Norgaard, J. Smith, G. Mooers, M. Klöwer, J. Lottes, S. Rasp, P. Düben, S. Hatfield, Neural general circulation models for weather and climate, Nature 632 (8027) (2024) 1060–1066.
[16] I. Price, A. Sanchez-Gonzalez, F. Alet, T.R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia, R. Lam, Probabilistic weather forecasting with machine learning, Nature 637 (8044) (2025) 84–90.
[17] L. Chen, X. Zhong, H. Li, J. Wu, B. Lu, D. Chen, ... Y. Qi, A machine learning model that outperforms conventional global subseasonal forecast models, Nature Commun. 15 (1) (2024) 6425.
[18] D.R. Stauffer, N.L. Seaman, Use of four-dimensional data assimilation in a limited-area mesoscale model. Part I: experiments with synoptic-scale data, Mon. Weather Rev. 118 (6) (1990) 1250–1277.
[19] H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, The ERA5 global reanalysis, Q. J. R. Meteor. Soc. 146 (730) (2020) 1999–2049.
[20] G.B. Anderson, M.L. Bell, R.D. Peng, Methods to calculate the heat index as an exposure metric in environmental health research, Environ. Health Perspect. 121 (10) (2013) 1111–1119.
[21] R.J. Osczevski, The basis of wind chill, Arctic (1995) 372-382.
[22] J. Xing, S. Zheng, S. Li, L. Huang, X. Wang, J.T. Kelly, S. Wang, C. Liu, C. Jang, Y. Zhu, J. Zhang, Mimicking atmospheric photochemical modeling with a deep neural network, Atmos. Res. 265 (2022) 105919.
[23] X. Shi, Z. Chen, H. Wang, D.Y. Yeung, W.K. Wong, W.C. Woo, Convolutional LSTM network: A machine learning approach for precipitation nowcasting, Adv. Neural Inform. Proc. Syst. 28 (2015).
[24] W.C. Skamarock, J.B. Klemp, J. Dudhia, D.O. Gill, D.M. Barker, M.G. Duda, X.-Y. Huang, W. Wang, J.G. Powers, 2008, A Description of the Advanced Research WRF Version 3. NCAR Tech. Note NCAR/TN-475+STR, 113 pp. 2008..
[25] B.H. Baek, C. Coats, S. Ma, C.-T. Wang, Y. Li, J. Xing, D. Tong, S. Kim, J.-H. Woo, Dynamic Meteorology-induced Emissions Coupler (MetEmis) development in the Community Multiscale Air Quality (CMAQ): CMAQ-MetEmis, Geosci. Model Dev. 16 (2023) 4659–4676, https://doi.org/10.5194/gmd-16-4659-2023.

[26] S. Li, Y. Ding, J. Xing, J.S. Fu, Retrieving ground-level PM2.5 concentrations in China (2013–2021) with a numerical-model-informed testbed to mitigate sample-imbalance-induced biases, Earth Syst. Sci. Data 16 (2024) 3781–3793, https://doi.org/10.5194/essd-16-3781-2024.
[27] Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. arXiv preprint, 2014, arXiv:1412.6980.
[28] T.T. Smith, B.F. Zaitchik, J.M. Gohlke, Heat waves in the United States: definitions, patterns and trends, Clim. Change 118 (3) (2013) 811–825.
[29] Gridded Population of the World, Version 4 (GPWv4): Population Count Adjusted to Match 2015 Revision of UN WPP Country Totals, Revision 11. Center for International Earth Science Information Network - CIESIN - Columbia University. (2018). NASA Socioeconomic Data and Applications Center (SEDAC). DOI: ⟨https://doi.org/10.7927/H4PN93PB⟩.
[30] J.T. Howard, N. Androne, K.C. Alcover, A.R. Santos-Lozada, Trends of heat-related deaths in the US, 1999-2023, JAMA 332 (14) (2024) 1203–1204.
[31] J. Berko, Deaths attributed to heat, cold, and other weather events in the United States, 2006-2010 (No. 76), National Center for Health Statistics, 2014.
[32] USGCRP, D.R. Reidmiller, C.W. Avery, D.R. Easterling, K.E. Kunkel, K.L.M. Lewis, T.K. Maycock, and, B.C. Stewart (Eds.), Impacts, Risks, and Adaptation in the United States: Fourth National Climate Assessment, II U.S. Global Change Research Program, Washington, DC, USA, 2018, p. 1515, https://doi.org/10.7930/NCA4.2018.