Published in partnership with CECCR at King Abdulaziz University

https://doi.org/10.1038/s41612-024-00876-y

Skillful subseasonal ensemble predictions of heat wave onsets through better representation of land surface uncertainties

Check for updates

Qiyu Zhang[1,2,3], Mu Mu[4,5,6] & Guodong Sun[2,3,7] ✉

Uncertainties in land surface processes notably limit subseasonal heat wave (HW) onset predictions. A better representation of the uncertainties in land surface processes using ensemble prediction methods may be an important way to improve HW onset predictions. However, generating ensemble members that adequately represent land surface process uncertainties, particularly those related to land surface parameters, remains challenging. In this study, a conditional nonlinear optimal perturbation related to parameters (CNOP-P) approach was employed to generate ensemble members for representing the uncertainties in land surface processes resulting from parameters. Via six strong and long-lasting HW events over the middle and lower reaches of the Yangtze River (MLYR), HW onset ensemble forecast experiments were conducted with the Weather Research and Forecasting (WRF) model. The performance of the CNOP-P approach and the traditional random parameter perturbation ensemble prediction method was evaluated. The results demonstrate that the deterministic and probabilistic skills of HW onset predictions show greater excellence using the CNOP-P approach, leading to much better predictions of extreme air temperatures than those using the traditional method. This occurred because the ensemble members generated by the CNOP-P method better represented the uncertainties in important land physical processes determining HW onsets over the MLYR, notably vegetation process uncertainties, whereas the ensemble members generated by the random parameter perturbation method could not. This finding suggests that the CNOP-P method is suitable for producing ensemble members that more appropriately represent model uncertainties through more reasonable parameter error characterization.

As global temperatures increase, heat waves (HWs) are increasingly occurring in areas where they have not occurred before, and many researchers have demonstrated that the intensity, frequency, and duration of global and regional HW events are gradually increasing[1–4]. HW events are also referred to as silent killers owing to their negative impact on human health. The HW event in western Russia in 2010 caused more than 50,000 fatalities[5]. Moreover, the record-breaking European HW event in 2022 resulted in approximately 60,000 deaths[6]. Therefore, accurately predicting the onset of HW events in advance provides enormous social and economic benefits.

Skillful subseasonal predictions are extremely important for disaster preparedness, risk management, and agricultural planning[7,8]. Early warning or forecasting of high-impact weather events, particularly HW events, can considerably prevent the loss of life and property. With the advancement of

[1]Key Laboratory of Core Tech on Numerical Model-AI Integrated Forecast for Hazardous Precipitation, Chongqing Institute of Meteorological Sciences, Chongqing, China. [2]State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China. [3]University of Chinese Academy of Sciences, Beijing, China. [4]Department of Atmospheric and Oceanic Sciences and Institute of Atmospheric Sciences, Fudan University, Shanghai, China. [5]Shanghai Key Laboratory of Ocean-land-atmosphere Boundary Dynamics and Climate Change, Shanghai, China. [6]CMA-FDU Joint Laboratory of Marine Meteorology, Shanghai, China. [7]Key Laboratory of Earth System Numerical Modeling and Application, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China. ✉e-mail: sungd@mail.iap.ac.cn

numerical models and data assimilation technologies, as well as the widespread application of radar technology and satellites, forecasts at the typical timescale have substantially improved[9]. However, compared with those for weather forecasts and climate predictions, the forecast skills at the subseasonal to seasonal (S2S) timescales are much lower[10]. Many studies have demonstrated that the S2S model can hardly successfully predict HW onsets at the subseasonal timescale. Lin et al.[11] employed 10 subseasonal models to predict HW events that occurred in western North America in 2021 and noted that while most models could predict HWs two weeks in advance, their intensity was greatly underestimated. In China, the middle and lower reaches of the Yangtze River (MLYR), which is one of the most densely populated regions, also exhibit high uncertainty in subseasonal HW prediction. Xie et al.[12] demonstrated that the prediction skill for HWs over the MLYR decreased substantially after two weeks.

To improve the subseasonal prediction skills for HWs over the MLYR, the sources of HW predictability were investigated. Qi and Yang[13] demonstrated that the uncertainty in midlatitude intraseasonal oscillations was one of the primary reasons for underestimating the intensity of the HW event over the MLYR in 2012. Furthermore, Xie et al.[12] indicated that accurately describing the phase development and amplitude of high-pressure anomalies associated with intraseasonal oscillations may facilitate more accurate prediction of the intensity and duration of HWs over the Yangtze River Valley.

Moreover, the above studies have also revealed that the land surface is one of the main factors influencing subseasonal HW predictions, which is generally consistent with the findings of Koster et al.[14], Guo et al.[15] and Dirmeyer et al.[16]. Many researchers have demonstrated that soil moisture errors, soil temperature errors[17,18], snow cover errors[19], and vegetation status errors[20] are among the key factors leading to uncertainties in subseasonal HW predictions. To reduce the uncertainties due to the use of the deterministic prediction method, ensemble prediction is an effective strategy. Ensemble prediction is widely regarded as a useful tool for estimating forecast uncertainties[21–23].

Many numerical weather forecast centers worldwide have developed various ensemble forecast methods to represent model uncertainties related to the atmosphere. However, these methods have been applied to land surface models in fewer studies[24–26]. The accurate evaluation of land surface states and land surface physical parameters in numerical weather forecasts is crucial for enhancing forecast skills[27–29]. For example, MacLeod et al.[25] and Orth et al.[28] reported that randomly perturbing a small number of soil parameters increased the ensemble spread of the boundary layer while also enhancing the ensemble forecast skills.

However, the aforementioned studies have also indicated that ensemble forecast members produced exclusively using random perturbation (RP) methods are insufficient to adequately represent uncertainties in land surface processes, indicating that a small ensemble spread remains a key issue in existing ensemble forecasting systems. A suitable ensemble forecasting system should facilitate a more accurate evaluation of prediction uncertainties. Therefore, to overcome the limitations of random perturbation methods, ensemble members were generated by the conditional nonlinear optimal perturbation related to parameters (CNOP-P) approach[30] in this study. In the CNOP-P approach, parameter perturbations causing the highest forecast uncertainty are represented under a given constraint. Zhang et al.[31] demonstrated that land surface model parameter errors of the CNOP-P type could cause considerable uncertainties in subseasonal HW onset predictions, suggesting that the CNOP-P approach may be appropriate for representing model parameter uncertainties. Wang et al.[32] employed the CNOP-P approach to identify the most sensitive parameters causing the highest variation in precipitation, and the ensemble forecast experiments further confirmed the importance of the CNOP-P approach for resolving the underdispersive problem in a convection-allowing ensemble forecast system. This finding inspired us to apply the CNOP-P approach in ensemble HW predictions. Therefore, this study focused on the following questions: (1) Can the use of the CNOP-P method enhance the subseasonal ensemble prediction skills for HW onsets? (2) Compared with those generated by the traditional random parameter perturbation method, can the ensemble members generated by the CNOP-P approach better represent the uncertainties in land surface parameters?

Table 1 | Information and forecast periods of the six HW events

HWs	Initialization time	End time
HW1	24 June 1988	17 July 1988
HW2	3 July 2003	26 July 2003
HW3	20 July 2010	12 August 2010
HW4	15 July 2013	07 August 2013
HW5	02 July 2016	25 July 2016
HW6	12 July 2022	04 August 2022

Results

Ensemble forecast skills of the CNOP-P and RP experiments

Ensemble forecast experiments (Table 4) were conducted for the six selected HW events (Table 1), and the performance of the CNOP-P and RP experiments was assessed and compared. The ensemble forecast skills for Tm were investigated via a range of verification metrics. The deterministic skills were evaluated via the ensemble mean forecast error and the ensemble mean improvement. The Brier score (BS; Brier[33]), the continuously ranked probability score (CRPS; Matheson and Winklers[34]), relative operating characteristic (ROC) curves[35], the area under the ROC curve (ROCA[36],), and the reliability diagram (RD) were employed to evaluate the probabilistic skills. Additionally, the reliability of the ensembles generated by the CNOP-P approach and RP approach was assessed via the ratio of the ensemble spread to the ensemble mean forecast error. The details of all these verification metrics are described in “Methods” section.

Figure 1 shows the performance of the CNOP-P and RP ensemble forecasts and the control forecast for the area-averaged Tm of the six HW events over the MLYR. The ensemble mean of Tm of the CNOP-P experiment was closer to the observation than that of the control forecast, but the ensemble mean of Tm of the RP experiment almost approached the reference state. In addition, the ensemble members generated by the CNOP-P approach exhibited a greater spread, and the 95% confidence intervals of the ensemble means encompassed the observed trends, suggesting that the CNOP-P experiment successfully captured extremely high-temperature processes. In contrast, for the RP ensemble forecasts, the ensemble members with a limited spread were often concentrated around the control forecast, and the observed Tm trend exceeded the 95% confidence interval, preventing the forecasts from adequately capturing extremely high-temperature processes. Specifically, the ensemble mean errors of the CNOP-P and RP forecasts and the error of the control forecast for the six HW events were 1.49 °C, 1.83 °C, and 1.92 °C, respectively. This suggests that the ensemble members of the CNOP-P experiment positively influenced the prediction of subseasonal HW onsets.

Moreover, the ensemble members generated in the CNOP-P and RP experiments were investigated regarding the intensity and extent of the selected HW events. Figure 2 shows the spatial distribution of the observed Tm values exceeding 35 °C together with the forecast probabilities of the CNOP-P and RP experiments for the six HW events when Tm exceeds this threshold. The control forecast cannot accurately capture the spatial distribution and magnitude of the selected HW events. For the ensemble members generated on the RP experiment, the increase in the forecast skill for HW events is minimal, and it remains difficult to forecast HW onsets in most areas. Furthermore, because of the limited ensemble spread of the RP experiment, the probability predictions are mostly close to 1. Moreover, the 95% confidence interval of the RP method is narrower, with a minimal difference between the upper and lower limits (Fig. S2), which effectively indicates that decision-makers are not given additional probabilistic information. In contrast to the RP experiment, the ensemble predictions of the CNOP-P experiment for Tm exceed 35 °C at most grid points, which is

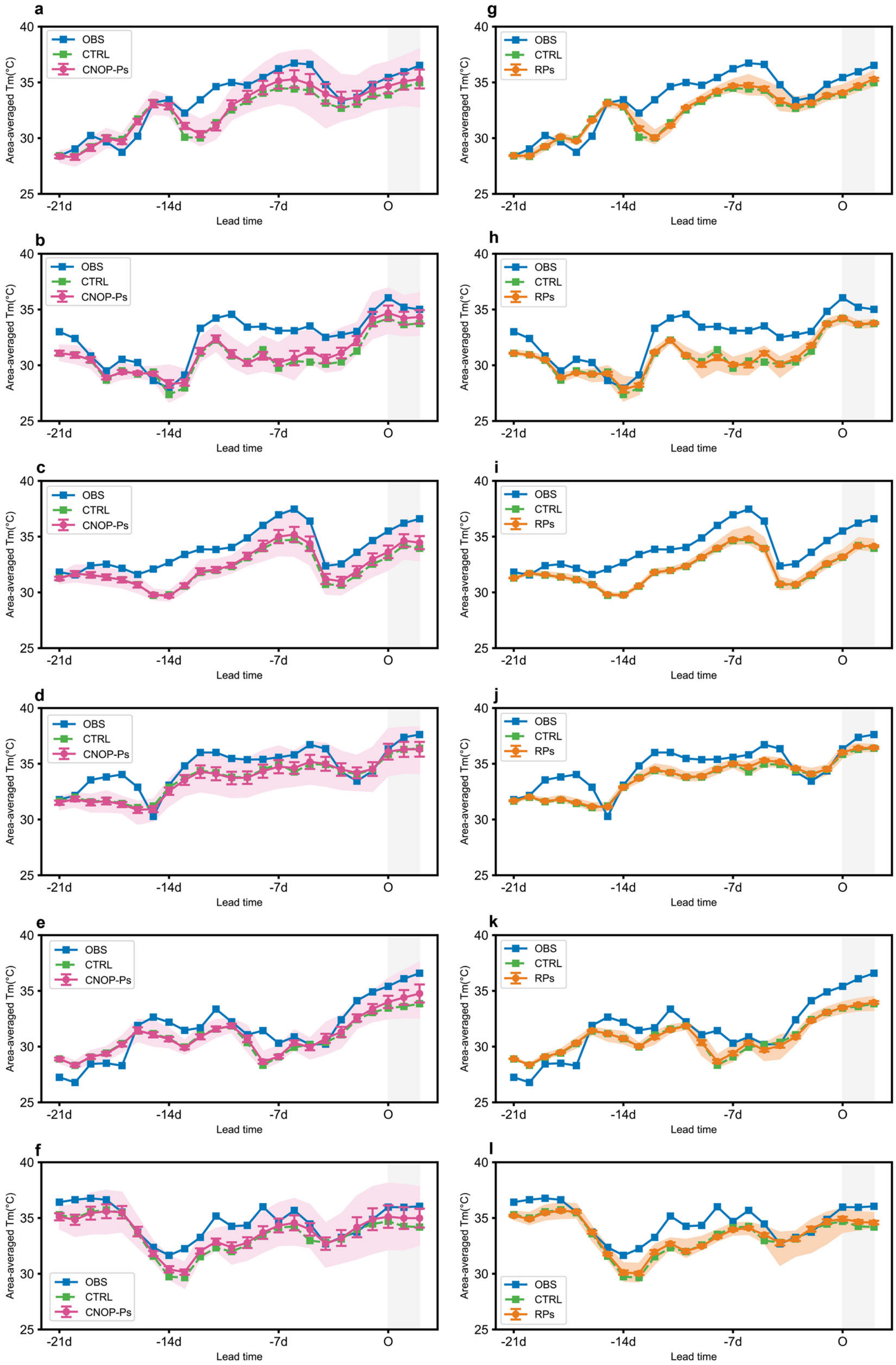


Fig. 1 | The temporal evolution of daily maximum temperature over the MLYR for six HW events. a-f The CNOP-P experiment for **a** HW1, **b** HW2, **c** HW3, **d** HW4, **e** HW5, and **f** HW6. The green line indicates the control forecast, the blue line indicates the observation, the pink line denotes the ensemble mean of the CNOP-P experiment, and the shadow denotes the ensemble members. The error bars indicate the 95% confidence intervals determined via the bootstrap method. **g–l** are similar to (**a–f**) but for the RP experiment.

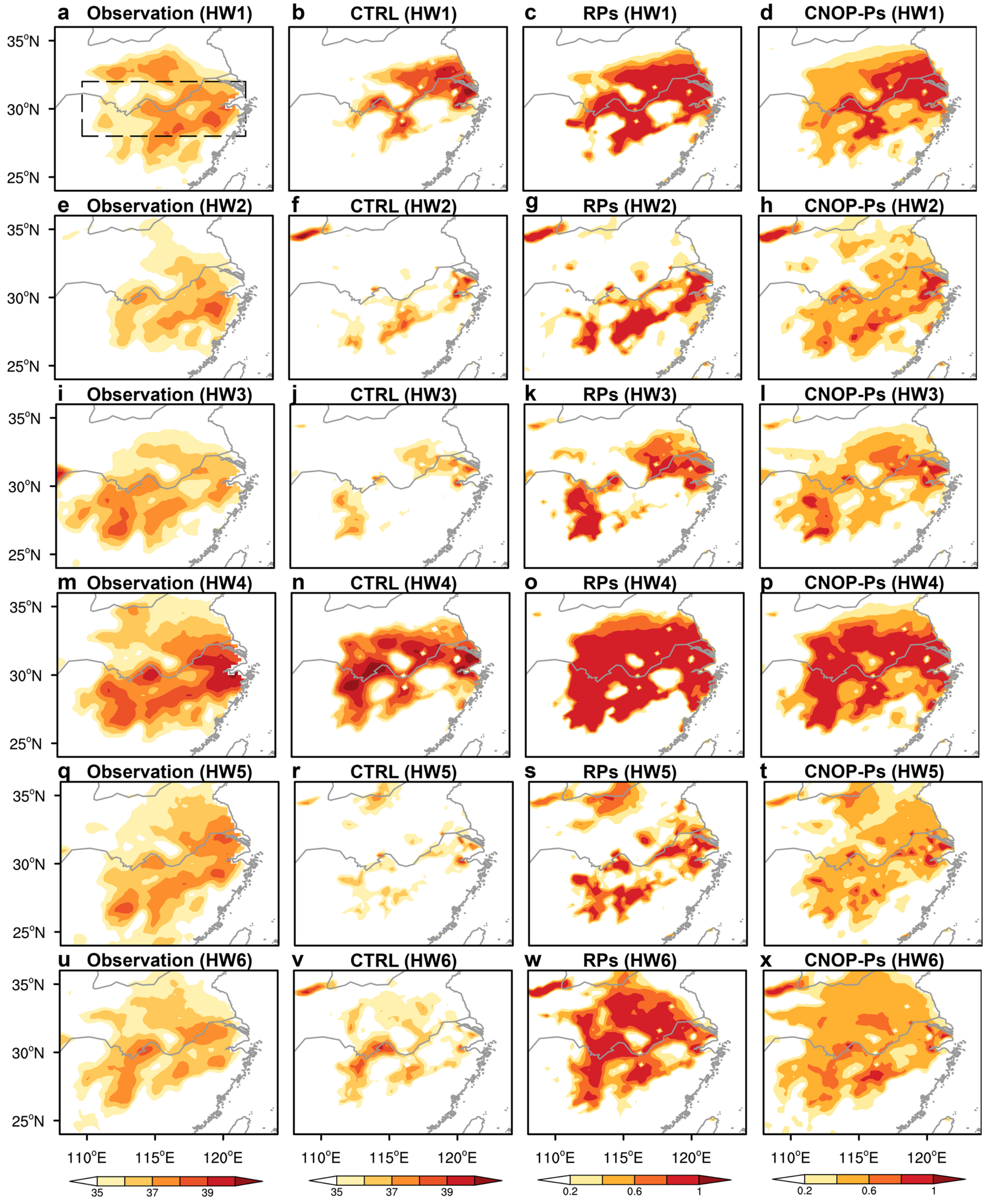


Fig. 2 | Spatial distributions and probabilistic analysis of six HW events. Spatial distributions of Tm for **a** observations and **b** control forecast during the heat wave periods and probability distributions of the HWs based on the ensembles generated in the **c** RP and **d** CNOP-P experiments for HW1. **e**–**h**, **i**–**l**, **m**–**p**, **q**–**t**, and **u**–**x** are similar to (**a**–**d**) but for HW2, HW3, HW4, HW5, and HW6, respectively. The black dashed box denotes the MLYR.

closer to the observations. The 95% confidence interval of the CNOP-P method is wide (Fig. S2), suggesting that the CNOP-P method provides an advantage in its ability to capture extreme weather events. The above results demonstrate that the CNOP-P approach performs better in HW forecasting, thereby providing users with more useful information and allowing individuals to determine whether to implement particular preventative steps to avoid losses.

To better evaluate the improvement achieved in the CNOP-P experiment, the improvements in the ensemble means obtained in the CNOP-P and RP forecasts relative to the control forecast during the HW period and

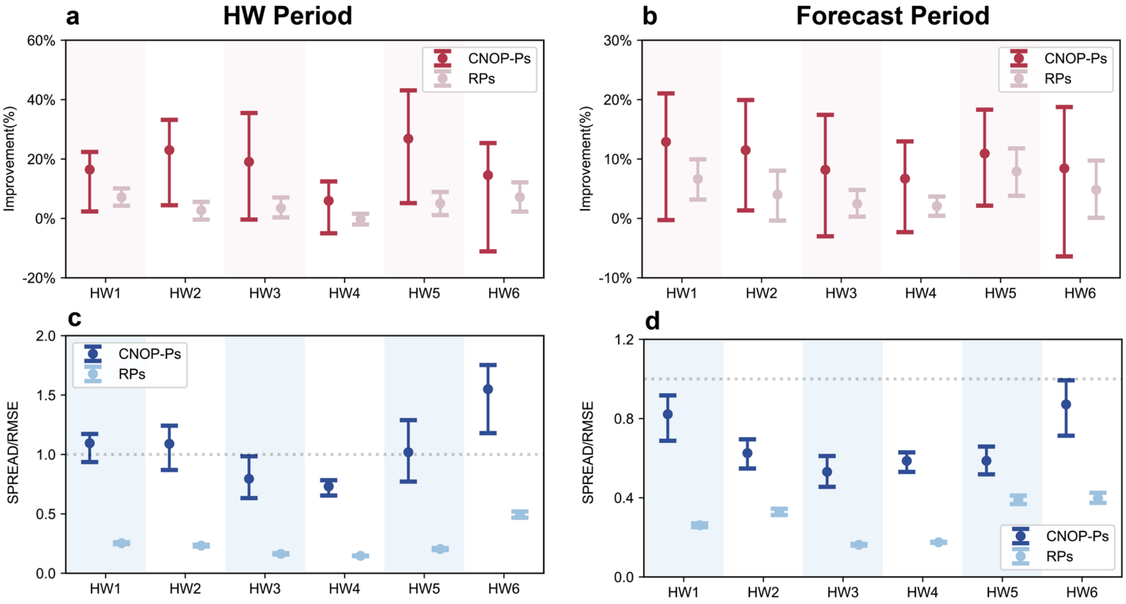


Fig. 3 | Improvements in ensemble mean forecasts and ensemble spread-error ratios of six HW events. Improvements in the ensemble means of the area-averaged Tm of the CNOP-P and RP experiments during **a** the HW period and **b** the entire forecast period. **c**, **d** are similar to (**a**, **b**) but for the ratio of the ensemble spread to the ensemble mean forecast error. The error bars indicate the 95% confidence intervals determined via the bootstrap method.

the entire forecast period are shown in Fig. 3a, b, respectively. In general, the improvement achieved by the CNOP-P forecasts is much greater than that achieved by the RP forecasts. In the CNOP-P forecasts, the improvement in the ensemble means of the six HW events during the HW period (22%) is much greater than that during the overall forecast period (11%). In the RP forecasts, the improvement is not as significant, as reflected by the average improvements of 5% and 4.5% for the six HW events during the HW period and the overall forecast period, respectively.

According to previous research, the ensemble spread and the root mean square error of a reliable ensemble forecasting system are approximately equivalent[37]. Therefore, the reliability of the ensemble members produced via the CNOP-P and RP methods is assessed using the ratio of the ensemble spread to the ensemble mean forecast error (Fig. 3c, d). Compared with that based on the RP-derived ensemble members, the difference between the ensemble spread and the ensemble mean forecast error based on the CNOP-P-derived ensemble members is notably small, with a value closer to 1, indicating that the use of the CNOP-P approach considerably mitigates the issue of insufficient dispersion and enhances the relationship between the ensemble spread and the ensemble mean error. This further suggests that the ensemble forecast system generated by the CNOP-P method offers more reliable ensemble predictions than those generated by the RP method.

Ensemble predictions can not only demonstrate deterministic forecast skills via ensemble means but also provide additional probability information for decision-makers. As a result, the probabilistic forecast skills of the CNOP-P and RP systems for the selected HWs are evaluated in terms of the BS, CRPS, ROC curve, ROCA, and reliability curve. Both the BS and CRPS indicate whether the forecast probability is consistent with the actual observations. However, the BS is a measure of the probabilistic forecast skill for binary events, whereas the CRPS primarily captures the probabilistic forecast performance of ensemble predictions via a comparison of the differences between the cumulative probability distributions of the predictions and observations. In this study, the probabilistic skill is evaluated via the average of the six HW events. In the comparison of the ensemble members produced via the CNOP-P approach with those produced via the RP method, the former results in lower BS and CRPS values over the whole forecast period (Fig. 4). In particular, over the HW period, the BS and CRPS values for the CNOP-P forecasts are 0.22 (0.16–0.30) and 1.29 (1.15–1.53), respectively, whereas the values for the RP forecasts are 0.34 (0.33–0.35) and 1.64 (1.61–1.67), respectively. This finding demonstrates that the ensembles generated via the CNOP-P approach are more reliable and that the corresponding forecast system provides a better probabilistic skill than the RP-based forecast system.

To better understand whether the forecast probability for the HW events is comparable to the actual occurrence frequency, reliability curves of the CNOP-P and RP forecasts under various high-temperature thresholds are shown in Fig. 5. Under a threshold of 35 °C, the reliability curve of the CNOP-P forecasts is closer to the diagonal line than that of the RP forecasts when the actual probability is low (below 0.4) or high (above 0.8), but the forecast probability is much lower than the actual probability for values between 0.4 and 0.8. To better demonstrate the reliability of the CNOP-P and RP forecasts, the distances between the two curves and the perfect curve are calculated, yielding values of 0.19 and 0.21, respectively. Consequently, compared with those generated via the random perturbation method, the ensemble members generated via the CNOP-P approach are still more reliable. A comparison of the reliability curves under temperature thresholds of 35 °C and 37 °C reveals that the reliability curves of the CNOP-P forecasts are closer to the diagonal line as the temperature threshold is increased, indicating that CNOP-P-derived ensemble members exhibit a higher probability of correctly forecasting extremely high temperatures.

Furthermore, it is important to consider not only the reliability of the ensemble predictions but also their capacity to discriminate between events and nonevents. The ROC curve provides the hit rate and the false alarm rate of the predictions for evaluating the discrimination capability of the ensemble predictions. The area under the ROC curve is often employed as a quantitative discrimination measure. Discrimination refers to the ability to distinguish between events and nonevents. Along with calibration, this aspect is one of the key attributes of probabilistic forecasts. ROCA values greater than 0.5 suggest that the ensemble forecast system provides a superior ability to distinguish between events and nonevents. Figure 6 shows the ROC curves of the CNOP-P and RP forecasts for various high-temperature thresholds. For the CNOP-P forecasts, the ROC curves under the different high-temperature thresholds closer to the left vertex indicate higher hit rates and lower false alarm rates. Additionally, the ROCA values of the CNOP-P forecasts are 0.863 (0.855–0.868) and 0.841 (0.827–0.848) under the 35 °C and 37 °C thresholds, respectively, whereas the ROCA values of the RP forecasts are 0.767 (0.760–0.772) and 0.711 (0.698–0.719), respectively. This finding demonstrates that the ensembles generated via the

Fig. 4 | The temporal evolution of BS and CRPS generated by the CNOP-P and RP methods. **a** BS and **b** CRPS for Tm. These values are computed across the MLYR and averaged over each period. The error bars indicate the 95% confidence intervals determined via the bootstrap method.

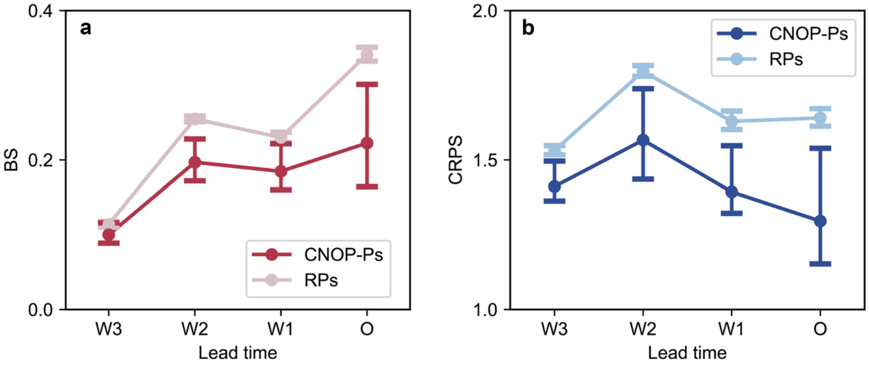


Fig. 5 | Reliability curves generated by the CNOP-P and RP methods for different high-temperature thresholds. **a** 35 °C high-temperature threshold, **b** 37 °C high-temperature threshold. Each point is computed across the MLYR and averaged over the whole forecast period. The error bars denote the 95% confidence intervals determined via the bootstrap method.

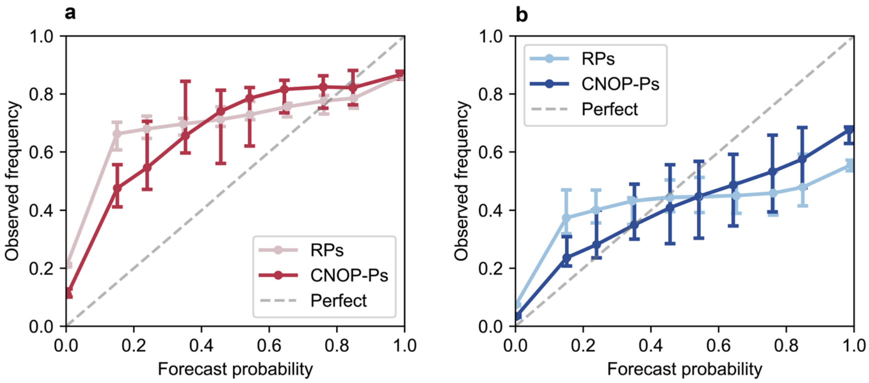
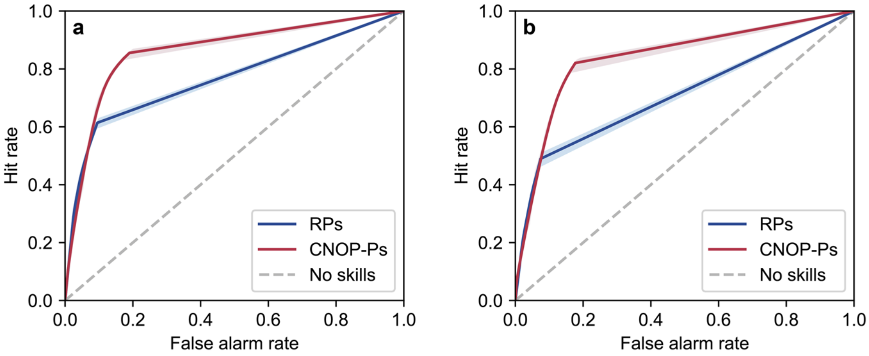


Fig. 6 | ROC curves generated by the CNOP-P and RP methods for different high-temperature thresholds. **a** 35 °C high-temperature threshold, **b** 37 °C high-temperature threshold. The ROC curves are computed across the MLYR and averaged over the whole forecast period. The shadow denotes the 95% confidence interval determined via the bootstrap method.



CNOP-P approach yield more skillful predictions of the HW events than those generated via the RP method.

Why does the CNOP-P method provide higher prediction skills?

What are the reasons that the ensemble prediction system based on the CNOP-P method achieves a higher subseasonal HW onset prediction skill than does the ensemble prediction system based on the RP method? Previous studies have demonstrated that land surface conditions (soil moisture and vegetation state) may play an important role in the onset and development of HW events by regulating water and energy transfer and distribution on the land surface and influencing high-pressure structures in the mid-troposphere. Seneviratne[38–40]. To better understand how the use of the CNOP-P approach enhances the subseasonal HW onset prediction skill, we examined the spread of surface energy, moisture, and atmospheric circulation in HW1 as an example (Fig. 7). Compared with those generated via the RP method, the CNOP-P method generates ensemble members with a larger spread for all the variables in Fig. 7. This suggests that one of the primary reasons for the better performance of the CNOP-P method may be its ability to assess the uncertainties in surface energy, water vapor, and atmospheric processes more accurately.

For surface energy processes, the difference in the spread of the sensible and latent heat fluxes between the two methods is particularly notable (Fig. 8). As a result, the impacts of the CNOP-P-derived and RP-derived ensemble members on the various components of the sensible and latent heat fluxes were examined in terms of ensemble uncertainty. Compared with those of soil evaporation, vegetation evaporation, and bare soil evaporation, the CNOP-P-derived ensemble members provide a greater spread of transpiration. Previous research has demonstrated that uncertainty in vegetation processes is one of the primary reasons for bias in HW predictions[41–43]. The ensemble members generated via the CNOP-P method can capture this uncertainty, suggesting that the CNOP-P method can characterize the uncertainty in vegetation processes more adequately, thus better describing and reducing the uncertainty in energy and moisture that

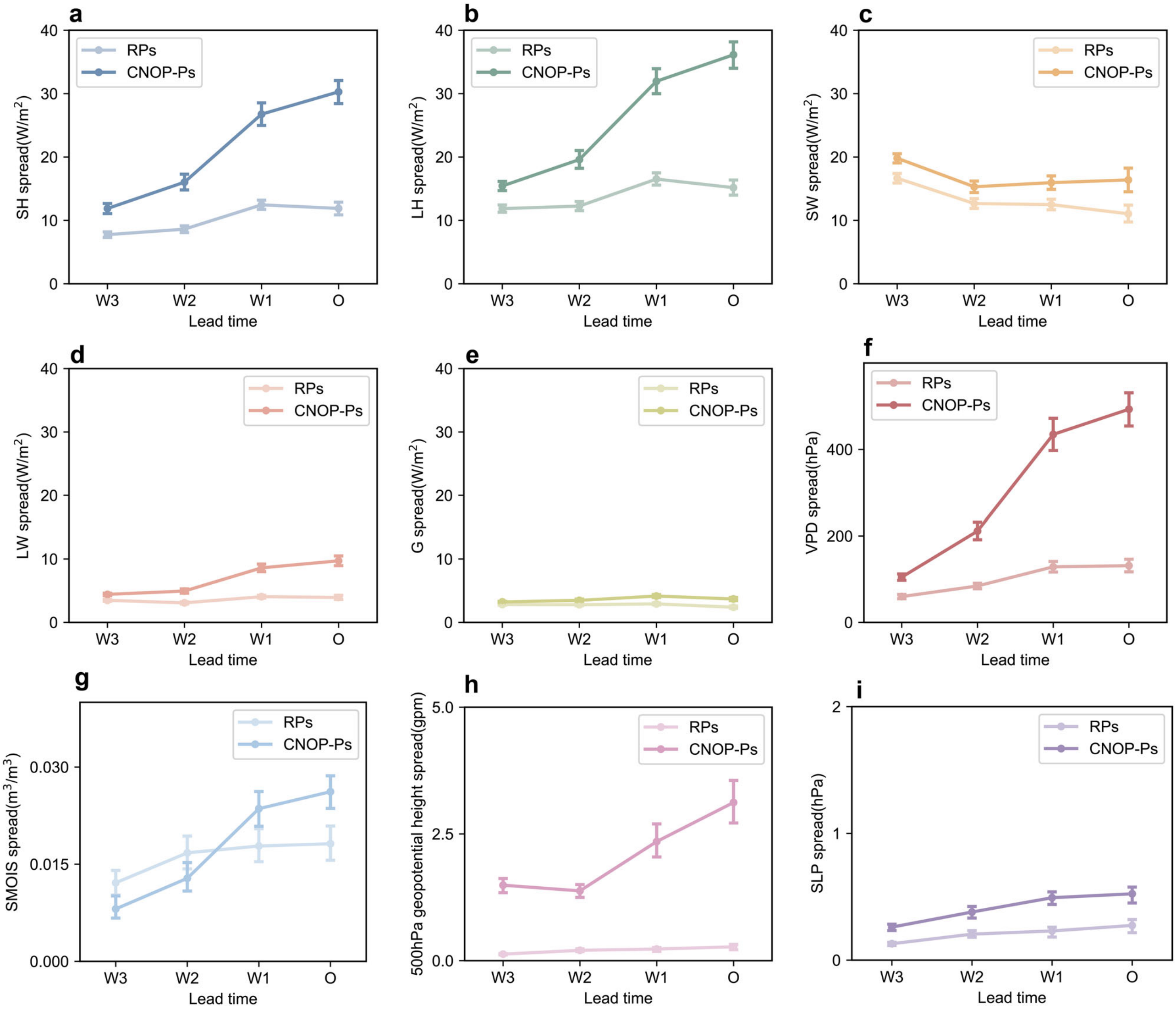


Fig. 7 | The temporal evolution of ensemble spread for key variables. **a** the sensible heat flux (SH, unit: W/m^2), **b** latent heat flux (LH, unit: W/m^2), **c** net shortwave radiation flux (SW, unit: W/m^2), **d** net longwave radiation flux (LW, unit: W/m^2), **e** soil heat flux (G, unit: W/m^2), **f** vapor pressure deficit (VPD, unit: Pa), **g** soil moisture (SMOIS, unit: m^3/m^3), **h** 500-hPa geopotential height (unit: gpm) and **i** sea level pressure (unit: hPa) of the CNOP-P and RP forecasts. The ensemble spread is computed across the MLYR and averaged over each period. The error bars denote the 95% confidence intervals determined via the bootstrap method.

dominate the HW process and providing greater ensemble forecast skills for subseasonal HW onset predictions.

Experiment with CNOP-P ensemble forecasts for non-HW events

As described above, the CNOP-P-derived ensemble members provide a better representation of the uncertainties in the model parameters, which decreases the missing alarm rate and renders the HW predictions more accurate. Therefore, do the CNOP-P ensemble predictions still yield low false alarm rates for the non-HW events? Although the analysis of the ROC curve has demonstrated that the CNOP-P ensemble predictions yielded a lower false alarm rate than did the RP ensemble predictions, it was further evaluated whether the CNOP-P ensemble predictions yielded false alarms for non-HW events. Therefore, six non-HW events (Table 2) are analyzed in this section, and the CNOP-P and RP methods were employed to perform ensemble forecast experiments. From the perspectives of the ensemble mean and ensemble members, we explored whether the CNOP-P ensemble members would predict non-HW events as HW events.

In this section, ensemble forecast experiments consistent are conducted for the above six non-HW events via the WRF model to determine whether the CNOP-P ensemble members yield false alarm rates for subseasonal HW onset predictions. Figure 9 shows the spatial distributions of the differences between Tm and the high-temperature threshold of 35 °C for the observations, the control forecast, and the ensemble means of the RP and CNOP-P forecasts over the validation period for the six non-HW events.

According to the observations, although the area-averaged Tm over the MLYR of the six events is lower than 35 °C overall, Tm may exceed 35 °C at some grid points, particularly NHW1, NHW2, and NHW3. Regarding the Tm predictions, the control forecast skill is low. The ensemble means of the RP forecasts are similar to those of the control forecast. Regarding the CNOP-P forecasts, the ensemble means of Tm indicates that the non-HW events are not predicted as HW events. Furthermore, in NHW3, the use of the CNOP-P ensemble members reduces the notable overestimation of Tm at some grid points in the control forecast, suggesting that the ensembles generated via the CNOP-P approach provide forecasts that are closer to the observations than the control forecast is.

Furthermore, the predictions of Tm obtained with ensemble members generated via the CNOP-P and RP methods (Fig. 10) were analyzed. The predictions obtained with the ensemble members generated via the RP method were consistently distributed around the control forecasts. In

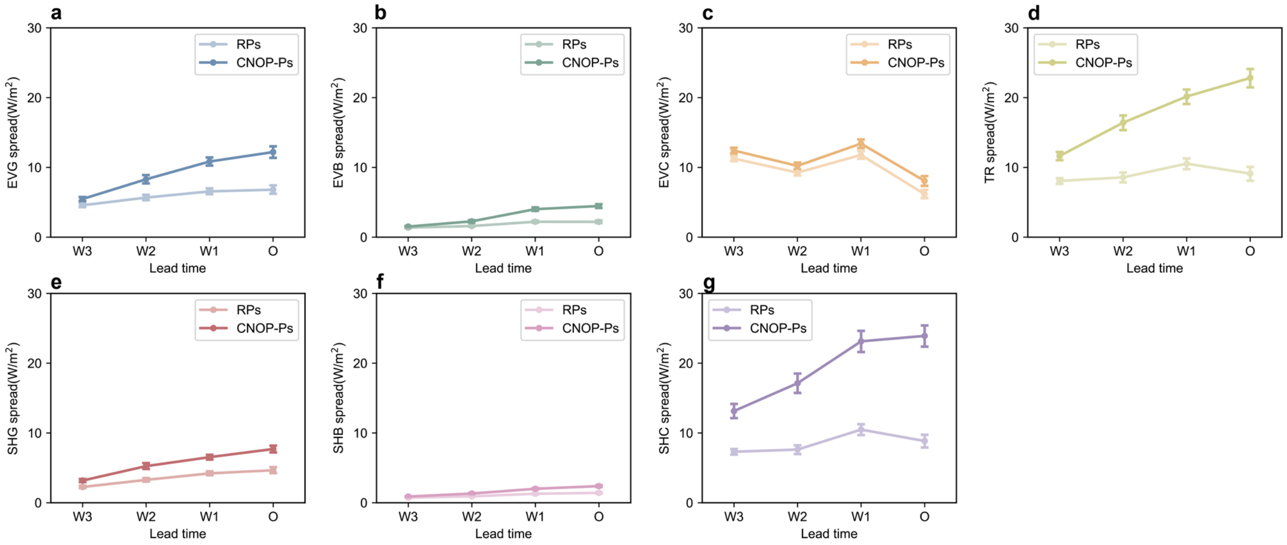


Fig. 8 | The temporal evolution of ensemble spread for different energy fluxes. **a** vegetated soil evaporation (EVG, unit: W/m^2), **b** bare soil evaporation (EVB, unit: W/m^2), **c** canopy water evaporation (EVC, unit: W/m^2), **d** plant transpiration (TR, unit: W/m^2), **e** vegetated ground sensible heat flux (SHG, unit: W/m^2), **f** bare ground sensible heat flux (SHB, unit: W/m^2) and **g** canopy sensible heat flux (SHC, unit: W/m^2) of the CNOP-P and RP forecasts. The ensemble spread is computed across the MLYR and averaged over each period. The error bars denote the 95% confidence intervals determined via the bootstrap method.

contrast, the ensemble members generated via the CNOP-P approach yielded predictions with a greater spread and provided a better representation of the uncertainty in estimates. Nevertheless, for the six non-HW events, none of the ensemble members generated via the CNOP-P approach produced Tm values greater than 35 °C, which indicates that the probability of the CNOP-P ensemble members predicting a non-HW event as an HW event during the target period is 0. In summary, the CNOP-P ensemble forecast experiments for the selected non-HW events demonstrated that the ensemble members generated via the CNOP-P approach exhibit low false alarm rates, thus verifying the usefulness of this approach.

Discussion

In this study, the conditional nonlinear optimal perturbation related to parameters (CNOP-P) approach was employed to represent the uncertainties in land surface model parameters. To investigate whether CNOP-P ensemble members with nonlinear perturbation characteristics could better represent the uncertainties in model parameters than could those generated via the traditional random parameter perturbation (RP) method, CNOP-P, and RP ensemble forecast experiments were conducted for six HW events over the middle and lower reaches of the Yangtze River (MLYR) via the WRF model. The performance of the CNOP-P and RP forecasts was assessed via different metrics, including an evaluation of the deterministic and probabilistic forecast skills, as well as the reliability of the ensemble members.

An examination of the temporal evolution and horizontal distribution of the daily maximum temperature (Tm) revealed that the CNOP-P approach provides ensemble members that can capture extremely high-temperature processes, whereas the ensembles obtained via the RP method primarily yielded predictions that occur near the control forecast, thus making it difficult to capture high-temperature processes. Furthermore, the ensemble mean predictions of Tm generated by the CNOP-P ensemble members exhibited substantially smaller forecast errors than those of the predictions generated by the RP ensemble members, which greatly enhanced the forecast skill. Moreover, the predictions properly captured the range of HWs, which led to a reduction in HW underreporting and provided more probabilistic information for subseasonal HW onset predictions.

Table 2 | Information and forecast periods of the six non-HW events

Non-HW events	Year	Target period
NHW1	2011	07.03–07.05
NHW2	2012	07.27–07.29
NHW3	2017	08.23–08.25
NHW4	2011	08.05–08.07
NHW5	2012	06.23–06.25
NHW6	2017	07.01–07.03

From the perspective of ensemble prediction uncertainties, we explored why the CNOP-P method provided a higher ensemble prediction skill than the RP method did. Compared with the RP ensemble members, the CNOP-P ensemble members provided more varied forecasts of physical processes related to HWs and could more accurately characterize the uncertainties in predicting important physical processes such as surface energy, moisture, and atmosphere. Moreover, the CNOP-P ensemble members could better capture the uncertainties in the vegetation transpiration process, which determines surface energy partitioning. This provided a better characterization of the uncertainty in subseasonal HW onset predictions.

Additionally, compared with the ensembles produced via the RP method, those generated via the CNOP-P approach usually produced predictions with greater spreads and smaller ensemble mean forecast errors. Therefore, there was a smaller difference between the ensemble mean forecast error and the ensemble spread, indicating that the ensembles produced via the CNOP-P approach were more reliable for predicting HWs. The probabilistic forecasting performance of the CNOP-P ensemble members for HW prediction was much greater than that of the RP ensemble members, as indicated by the superior ROC curves, smaller BS values, and lower CRPS values of the former ensemble members. As a consequence, the CNOP-P ensemble members could more accurately represent the uncertainties in land surface model parameters and outperformed the RP ensemble members in terms of HW forecast skill.

To further explore whether the CNOP-P ensemble members could yield false alarms for non-HW events, similar CNOP-P ensemble forecast experiments were conducted for six non-HW events. The ensemble mean predictions of Tm generated by the CNOP-P ensemble members were more similar to the observations than the control forecasts were. Furthermore, an analysis of the CNOP-P ensemble members revealed that no one member produced Tm values above the high-temperature threshold during the target period. Therefore, the CNOP-P ensemble members did not predict non-HW events as HW events, indicating low false alarm rates for HW

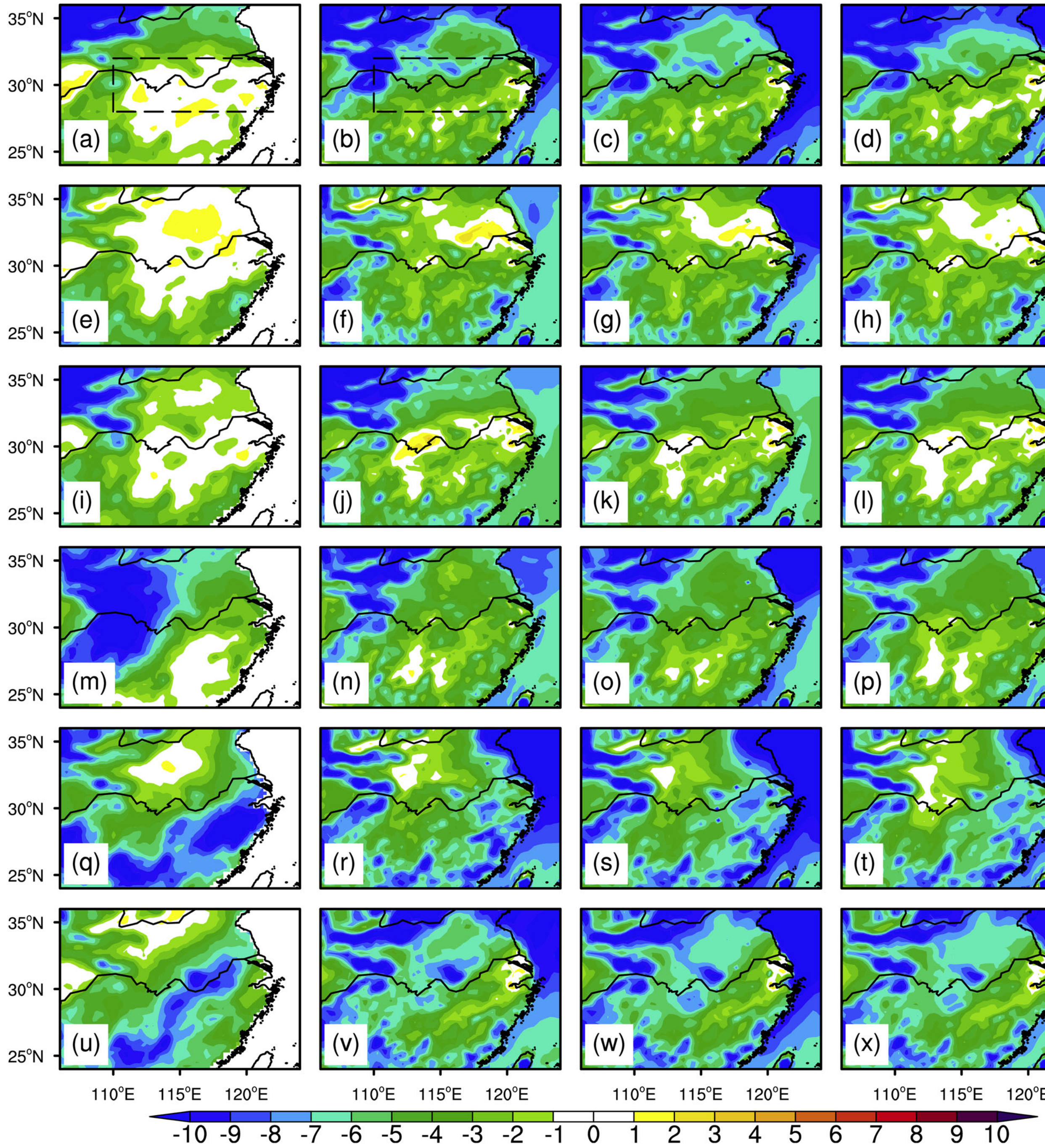


Fig. 9 | The differences between daily maximum temperature and 35 °C high-temperature threshold. **a** observations, **b** control forecast, and ensemble means based on the ensembles generated via the **c** RP and **d** CNOP-P methods for NHW1 during the heat wave period. The black dashed box denotes the MLYR. **e**–**h**, **i**–**l**, **m**–**p**, **q**–**t**, and **u**–**x** are similar to **a**–**d** but for NHW2, NHW3, NHW4, NHW5 and NHW6, respectively.

events. The obtained findings further support the reliability and effectiveness of the CNOP-P ensemble forecasting approach in subseasonal HW onset predictions.

As demonstrated by our findings, fully accounting for the uncertainties in land surface parameters could greatly enhance the forecast skill and reliability of subseasonal HW onset predictions. The results reported by MacLeod et al.[25], Orth et al.[28], and Gehne et al.[29] also support this phenomenon. However, these studies, along with ours, show that random parameter perturbation approaches often produce ensemble members with a modest spread. To explore whether increasing the ensemble size of random perturbations could improve forecast uncertainty, we doubled the sample size of random perturbations from the original to observe any significant increase in spread. Specifically, for six HW events, the average spread over the entire forecast period and the HW period with the increased random perturbation ensembles (46 ensemble members) was 0.58 and 0.44, respectively, which is slightly lower than that of the original ensemble (23 ensemble members). This suggests that despite the increase in random perturbation samples, their unstructured nature limits their effectiveness in capturing uncertainties, particularly the nonlinear processes that are crucial in HW predictions. This further highlights the importance of structured perturbation samples, such as CNOP-Ps, which are particularly valuable for accurately representing the complex and nonlinear interactions present in land surface parameters, contributing to the improved prediction of HW events.

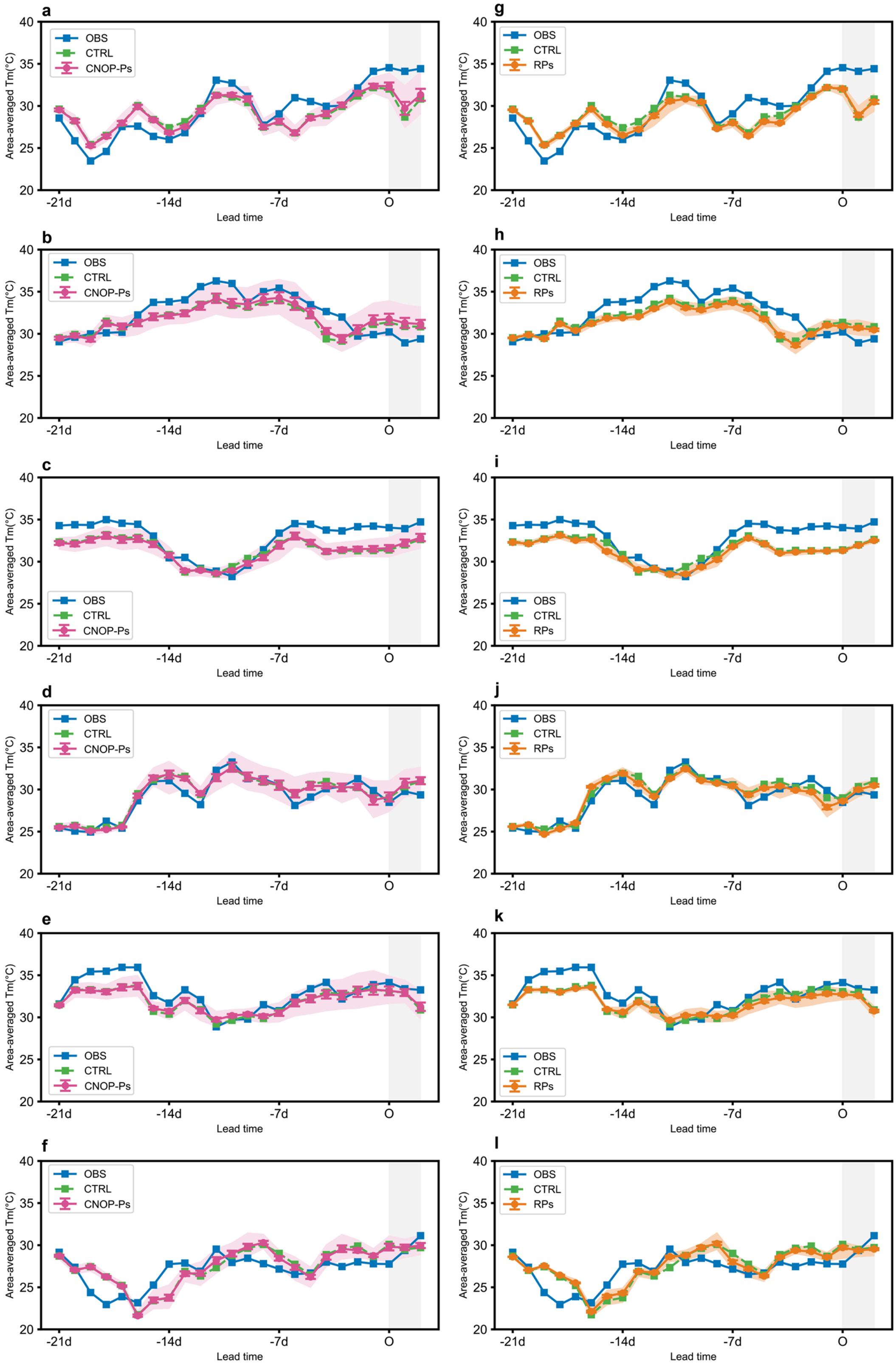


Fig. 10 | The temporal evolution of daily maximum temperature over the MLYR for six non-HW events. **a-f** The CNOP-P experiment for **a** NHW1, **b** NHW2, **c** NHW3, **d** NHW4, **e** NHW5 and **f** NHW6. The green line denotes the control forecast, the blue line denotes the observation, the pink line denotes the ensemble mean of the CNOP-P forecasts, and the shadow denotes the ensemble members. The error bars indicate the 95% confidence intervals determined via the bootstrap method. **g–l** are similar to (**a–f**) but for the RP experiment.

The potential of the CNOP-P method in improving ensemble prediction skills for subseasonal HW onsets has been demonstrated in this study. We know that for extreme HW events, the impact of missing alarms is much greater than that of false alarms. The ensemble members generated by the CNOP-P method can effectively capture HW events due to large spread and high reliability, which greatly reduces the risk of missing alarms and is particularly important in practical applications. However, the process of calculating nonlinear perturbations requires substantial computational resources. Typically, calculating a CNOP-P ensemble member requires at least 200 iterations. Despite this process being time-consuming, its advantages are significant, especially in the forecasting of extreme HW events. Therefore, future research should focus on developing more effective optimization methods as well as improving the integration of numerical models while taking into account both computational costs and efficiency. This will help to enhance the practicality of the ensemble forecasts related to the CNOP method, allowing it to maximize its forecast performance with higher timeliness. It is worth mentioning that recent studies have highlighted the potential of artificial intelligence (AI) models in exploring predictability research for various weather and climate events via the CNOP method[44,45]. With the help of higher efficiency and self-contained optimization modules of AI models, the solving of CNOP in AI models can be implemented conveniently. In addition, the obtained results in AI models can also be verified in numerical models, which also indicates AI models can learn physical mechanisms to some extent, enhancing the reliability and interpretability of findings. This also inspires us to perform predictability studies for HW events in a skillful AI model for targeted observation and ensemble forecasts in the future.

While this study focused on the potential effects of the uncertainties in land surface parameters on subseasonal HW onset predictions, it is crucial to note that initial land surface errors are also important[16,39,46,47]. Therefore, the ensemble forecast system should consider both the initial land surface errors and land surface model parameter perturbations to represent greater uncertainties. Furthermore, with the increasing complexity of many model components, the increase in model resolution, and the increasing number of parameters in the future, it is necessary to consider the function of each parameter in-depth and to determine the most sensitive parameters via the CNOP-P approach to reduce the computational cost before performing ensemble forecast experiments.

Methods

CNOP-P approach

The CNOP-P method is briefly described below. Assuming that there is a state vector U, it can be predicted via Eq. (1):

$$\begin{cases} \frac{\partial \boldsymbol{U}}{\partial t} + F(\boldsymbol{U}, \boldsymbol{P}) = \boldsymbol{0} \\ \boldsymbol{U}|_{t=0} = \boldsymbol{U}_0, \end{cases} \tag{1}$$

where $\boldsymbol{U_0}$ is the initial state, and $\boldsymbol{F}$ is a nonlinear operator. Assuming that the initial and boundary conditions are perfect, $\boldsymbol{M}$ denotes the nonlinear propagation operator from the initial time to time t. Then, the state variable $\boldsymbol{U}$ with the parameter vector $\boldsymbol{P}$ at time t can be expressed as $\boldsymbol{U(t)}=\boldsymbol{M_t}(\boldsymbol{U_0},\boldsymbol{P})$. When there is a parameter error $\boldsymbol{p}$ in the model parameters, we can obtain $\boldsymbol{U(t)} + \boldsymbol{u(t)}=\boldsymbol{M}_t(\boldsymbol{U}_0, \boldsymbol{P}+\boldsymbol{p})$, where $\boldsymbol{u(t)}$ denotes the prediction error caused by the parameter error at time t.

To address the parameter errors that impact the forecast results at time t the most, the following nonlinear constrained optimization problem was defined:

$$J(p_\delta) = \max_{\boldsymbol{p}\in\Omega} \|\boldsymbol{u}(t)\| = \max_{\boldsymbol{p}\in\Omega} \|\boldsymbol{M}_t(\boldsymbol{U}_0, \boldsymbol{P}+\boldsymbol{p}) - \boldsymbol{M}_t(\boldsymbol{U}_0, \boldsymbol{P})\|_b \tag{2}$$

where $\boldsymbol{J}$ is an objective function that aims to measure the maximum deviation from the reference state with the parameter constraint condition Ω. Moreover, $\| \|_b$ is a measure of the magnitude of the prediction errors caused by parameter perturbations, p_δ denotes the CNOP-P, which represents the parameter perturbations that cause the maximum prediction errors are represented under certain constraints.

WRF model, HW events, and non-HW events

The WRF model version 4.2.1 was employed in this study. The model simulation area covers the domain of China (15.8°N–45.5°N, 64.6°E–131.4°E), and it exhibits a horizontal grid spacing of 30 km, a grid cell number of 100 × 180, and a total of 34 vertical levels from the surface to 50 hPa. The parameterization schemes adopted in this study include the Thompson microphysics scheme[48], the Rapid Radiative Transfer Model (RRTM) longwave radiation scheme[49], the Dudhia shortwave[50], the Yonsei University (YSU) boundary layer scheme[51], the Kain-Frisch cumulus parameterization scheme[52], and the Noah-multiparameterization (Noah-MP) land surface parameterization scheme[53].

For the initial and lateral boundary conditions, the fifth-generation European Centre for Medium-Range Weather Forecasts (ECMWF) atmospheric reanalysis (ERA5) data with a horizontal resolution of 0.5° and a temporal resolution of 6 h were used. Notably, this study focused on the performance of the CNOP-P experiments and primarily aimed to investigate the influence of model parameter uncertainty on subseasonal HW onset predictions. Therefore, more precise initial and boundary conditions could highlight the importance of model parameters in subseasonal HW onset predictions. Via the use of observation data[54,55], which were interpolated from the original observation data of 2416 surface meteorological stations in China, six HW events from 1979 to 2022 over the MLYR (28–32°N, 110–122°E) were selected for the ensemble forecast experiments. The selection criteria for HW events have been provided by Zhang et al.[31] Table 1 provides brief information on the six HW events over the MLYR and the initialization and end times of the corresponding forecast periods.

Considering the interannual temperature variability, we focused on non-HW periods between June and August over the past decade (2010–2021). Referring to the definition of an HW event, an event was classified as a non-HW event if the following criteria were satisfied concurrently: (1) the area-averaged Tm does not exceed 35 °C for three consecutive days during the target period as well as the week before or after the target period; (2) the area-averaged Tm remains between the 5th and 95th percentile values of the climate state (1979–2016) throughout the target period and the week prior and following the target period. These criteria were designed to ensure a generally stable temperature trend, with no HW events occurring before or after the target period. In addition, considering that anticyclonic circulation is conducive to HW onset, we selected three non-HW events with anticyclonic anomalies during the HW period, thereby aiming to investigate whether the CNOP-P ensemble members would yield false alarms for non-HW events under background conditions favoring HW onset. For comparison, three non-HW events with no anticyclonic anomalies were selected during different periods of the same year. Table 2 provides detailed information on the six non-HW events selected.

Experimental design

To implement the CNOP-P approach, in which the uncertainties in the model parameters are represented, the degree of deviation of the daily maximum temperature (Tm) from the control forecast was adopted as the objective function, and 24 key physical parameters (Table 3) affecting surface energy and moisture changes were perturbed. Under the given parameter constraint conditions, one CNOP-P type of land surface parameter perturbation could be obtained. Considering the high uncertainty in the parameter perturbation amplitude, multiple CNOP-P-type land surface parameter perturbations were generated as ensemble members via different perturbation constraints. Therefore, the constraint optimization problem in Eq. (2) can be rewritten as follows:

$$J_i(p_\delta) = \max_{\boldsymbol{p}\in\Omega_i} \|\boldsymbol{u(t)}\| = \max_{\boldsymbol{p}\in\Omega_i} |\frac{\boldsymbol{1}}{\boldsymbol{3}}(\int_{\boldsymbol{day\ 0}}^{\boldsymbol{day\ 2}} \overline{\boldsymbol{T_m}}(\boldsymbol{P}+\boldsymbol{p})\boldsymbol{dt} - \int_{\boldsymbol{day\ 0}}^{\boldsymbol{day\ 2}} \overline{\boldsymbol{T_m}}(\boldsymbol{P})\boldsymbol{dt})| \tag{3}$$

Table 3 | Selected parameters of the Noah-MP land surface model

Number	Parameter	Description	Default value	Units	Min.	Max.
P1	RHOL-vis	Leaf reflectance (visible wavelengths)	0.10	-	0.09	0.11
P2	RHOL-nir	Leaf reflectance (near-infrared wavelengths)	0.45	-	0.405	0.495
P3	RHOS-vis	Stem reflectance (visible wavelengths)	0.16	-	0.144	0.176
P4	RHOS-nir	Stem reflectance (near-infrared wavelengths)	0.39	-	0.351	0.429
P5	TAUL-vis	Leaf transmittance (visible wavelengths)	0.05	-	0.045	0.055
P6	TAUL-nir	Leaf transmittance (near-infrared wavelengths)	0.25	-	0.225	0.275
P7	TAUS-vis	Stem transmittance (visible wavelengths)	0.001	-	0.0009	0.0011
P8	TAUS-nir	Stem transmittance (near-infrared wavelengths)	0.001	-	0.0009	0.0011
P9	XL	Leaf/stem orientation index	0.250	-	0.225	0.275
P10	Z0MVT	Momentum roughness length	0.80	m	0.72	0.88
P11	HVT	Top of the canopy	16.0	m	14.4	17.6
P12	HVB	Bottom of the canopy	10.0	m	9	11
P13	VCMX25	Maximum rate of carboxylation	55.0	$\mu mol\, m^{-2}s^{-1}$	49.5	60.5
P14	BP	Minimum leaf conductance	2000	$\mu mol\, m^{-2}s^{-1}$	1800	2200
P15	MP	Slope of the conductance–photosynthesis relationship	9.	-	8.1	9.9
P16	QE25	Quantum efficiency	0.06	$\mu mol\, m^{-2}s^{-1}$	0.054	0.066
P17	FOLNMX	Foliage nitrogen concentration before limitation	1.5	%	1.35	1.65
P18	BEXP	Pore size distribution index	8.17	-	7.353	8.987
P19	SMCWLT	Wilting point soil moisture	0.103	m^3m^{-3}	0.0927	0.1133
P20	SMCMAX	Saturated value of soil moisture	0.465	m^3m^{-3}	0.4185	0.5115
P21	SMCREF	Reference soil moisture	0.382	m^3m^{-3}	0.3438	0.4202
P22	PSISAT	Saturated soil matric potential	0.263	$m\, m^{-1}$	0.2367	0.2893
P23	DKSAT	Saturated soil hydraulic conductivity	2.45E-6	$m\, s^{-1}$	2.21E-06	2.7E-06
P24	DWSAT	Saturated soil hydraulic diffusivity	1.13E-5	m^2s^{-1}	1.02E-05	1.24E-05

where $\boldsymbol{u(t)}$ is the prediction error of the area-averaged daily maximum temperature over the MLYR during the HW period. The L1 norm, which is expressed in terms of absolute values, was employed to measure the magnitude of the prediction errors. Moreover, $\overline{T_m}$ denotes the area-averaged Tm, i denotes the number of ensemble members, and Ω is the amplitude of parameter perturbation. The target period, which ranges from days 0 to 2, was defined as occurring within three days following the onset of each HW event. Twenty-four land surface parameters were represented by the parameter vector $\boldsymbol{P}$.

Since the different parameters exhibited widely varying default values, normalization was performed to assess the effect of parameter perturbation. The normalization approach[56,57] can be expressed as follows:

$$\begin{cases} y = \frac{x-Defvalue}{Max\ value-Defvalue}, & when\ x \geq Defvalue \\ y = \frac{x-Defvalue}{Defvalue-\ Min\ value}, & when\ x<Defvalue \end{cases} \tag{4}$$

where *Defvalue*, *Maxvalue*, and *Minvalue* denote the default, maximum, and minimum values, respectively, of the parameters, and x and y denote the values before and after normalization, respectively. Hence, after normalization, the value of the perturbed parameter ranges from -1 to 1.

Given that a small perturbation amplitude leads to modest forecast uncertainties and that an excessive perturbation amplitude easily causes model instability, employs a perturbation amplitude ranging from 5%–10% was employed to produce 11 sets of ensemble members, and the value of Ω_i is $\{10\%, 9.5\%, 9\%, \ldots\ldots 6\%, 5.5\%, 5\%\}$. Furthermore, to ensure that the ensemble members were dispersed on both sides of the reference state, two optimization processes were adopted to obtain two types of CNOP-P ensemble members with positive and negative biases.

If a type of positive parameter error with the highest degree of deviation from the reference state is needed $(\overline{T_m}(\boldsymbol{P}+\boldsymbol{p_\delta}) \geq \overline{T_m}(\boldsymbol{P}))$, the first optimization process is as follows:

$$J1_i(p_\delta) = \max_{p\in\Omega_i}\{\frac{1}{3}(\int_{day\,0}^{day\,2}\overline{T_m}(\boldsymbol{P}+\boldsymbol{p})dt - \int_{day\,0}^{day\,2}\overline{T_m}(\boldsymbol{P})dt)\} \tag{5}$$

where p_δ denotes the type of parameter error that could cause a maximum positive error from the reference state.

If a type of negative parameter error with the highest degree of deviation from the reference state is needed $(\overline{T_m}(\boldsymbol{P}+\boldsymbol{p_\delta}) \leq \overline{T_m}(\boldsymbol{P}))$, the second optimization process is as follows:

$$J2_i(p_\delta) = \max_{p\in\Omega_i}\{\frac{1}{3}(\int_{day\,0}^{day\,2}\overline{T_m}(\boldsymbol{P})dt - \int_{day\,0}^{day\,2}\overline{T_m}(\boldsymbol{P}+\boldsymbol{p})dt)\} \tag{6}$$

where p_δ denotes the type of negative parameter error with the highest degree of deviation from the reference state.

Therefore, 23 ensemble members were generated by the CNOP-P approach, including 22 perturbed members and one control forecast member. To examine the usefulness of the CNOP-P approach in representing the uncertainties in model parameters, the ensemble predictions generated by the CNOP-P approach were compared with those obtained by the traditional random parameter perturbation method. In this study, two sets of ensemble prediction experiments (the CNOP-P and RP experiments) and a control experiment based on the WRF model were conducted (Table 4).

In the RP experiment, the 24 land surface parameters listed in Table 3 were disturbed via the random parameter perturbation method. The perturbation amplitude ranged from 5% to 10% of the default value of each parameter, which is consistent with the CNOP-P experiment to ensure experimental comparability and consistency. Specifically, a total of 11 sets of perturbation amplitudes were defined within the parameter perturbation

Table 4 | Experimental design of this study

Experiment	Parameter perturbation	Members	Integration period
CTRL	No parameter perturbation	1	24 days
CNOP-P	CNOP-P-type parameter perturbation	23	24 days
RP	Random parameter perturbation	23	24 days

range of 5% to 10%, and a series of positive and negative perturbation pairs were generated from each perturbation amplitude set. A set of 23 ensemble members was generated for the random parameter perturbation experiment, including 22 perturbation samples (both positive and negative perturbations) and one control forecast. Notably, this approach is not the standard way to generate random perturbations (i.e., multiple samples should be obtained from the 5%–10% range), but this method was adopted to maintain consistency and comparability with the CNOP-P experiment. In addition, an experiment was conducted via the traditional random perturbation method (Fig. S1), and the results were similar to those of the RP experiment.

In this study, the CNOP-P experiment was conducted via the differential evolution (DE) algorithm[58]. This study aimed to enhance the forecast skill for HWs at the subseasonal timescale. Thus, model integration was initiated three weeks before HW onset and lasted until three days thereafter. Day 0 marks the beginning of the HW event, and the mean values from days -21 to -15, days -14 to -8, days -7 to -1, and days 0 to 2 are denoted as W3, W2, W1, and HW, respectively.

Statistical evaluation methods for the prediction skills

The ensemble mean refers to the average of the forecast values of different ensemble members. The ensemble mean forecast error can be measured as the difference between the ensemble mean and the observations. The ensemble mean forecast error can be calculated as follows:

$$Forecast\ errors = \bar{X} - O = \frac{1}{N}\sum_{i=1}^{i} X_i - O \tag{7}$$

where $\bar{X}$ denotes the ensemble mean forecast, O denotes the observations, and N denotes the number of ensemble members.

The ensemble spread indicates the degree of forecast uncertainty and is a measure of the variation in the ensemble members with respect to the ensemble mean. The ensemble spread can be calculated as follows:

$$Spread = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |F_i - \bar{F}|^2} \tag{8}$$

where F_i and $\bar{F}$ are the forecast values of the i-th ensemble member and the ensemble mean, respectively. A reliable ensemble forecast system exhibits approximately equal ensemble spread and ensemble mean forecast error values[59].

The Brier score (BS) represents the mean square error of the probabilistic forecast[33], which is mainly used to assess the probabilistic forecasting skill of the ensemble forecast for dichotomous events and can be expressed as follows:

$$BS = \frac{1}{N}\sum_{i=1}^{N} (P_i - O_i)^2 \tag{9}$$

where N denotes the number of forecast trials, P_i is the probability that the event occurs in the i-th forecast, and O_i is the probability that the event actually occurs in the i-th forecast. In this study, the BS is employed to assess the probabilistic forecast ability for HWs. An HW event occurs when the daily maximum temperature increases beyond 35°C. In this case, $O_i = 1$, while otherwise, $O_i = 0$. A lower BS value suggests a closer forecast probability to the real probability, indicating a better probabilistic forecast skill for the predicted events.

The difference between the observed and predicted cumulative probability distributions can be statistically compared via the continuous ranked probability score (CRPS; Matheson and Winklers, 1976), which is defined as:

$$CRPS(P, x_a) = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2\, dx \tag{10}$$

where $P(x)$ and $P_a(x)$ denote the cumulative distributions of the forecast and observation data, respectively. Moreover, $P_a(x) = H(x - x_a)$, and $H(x)$ is the well-known Heaviside function, also referred to as the switching function, which can be expressed as follows:

$$H(x) = \begin{cases} 0\ ,\ x < 0 \\ 1\ ,\ x \geq 0 \end{cases} \tag{11}$$

A lower CRPS value suggests that the forecast probability is closer to the observed probability.

The relative operating characteristic (ROC) curve can be used to describe the probabilistic forecast skill of an ensemble forecast by computing the hit rate and the false alarm rate[35], which can be employed to evaluate the ability of forecasts to discriminate between events and nonevents. The horizontal coordinate of the ROC curve represents the false alarm rate, whereas the vertical coordinate represents the hit rate. Therefore, ROC curves occurring toward the top left corner of the graph, or a larger area under the curve (ROCA), suggest greater forecast skills.

The reliability diagram (RD) curve can be used to evaluate the probabilistic forecast skill by determining if the forecast probability corresponds to the observed probability. The horizontal axis of the RD curve denotes the forecast probability, and the vertical axis denotes the observed probability. As the RD curve approaches the diagonal line, the forecast and observed probabilities become more similar, indicating a better probabilistic forecast ability. An RD curve occurring above the diagonal indicates that the forecast probability is lower than the observed frequency. Conversely, the forecast probability is greater than the observed probability.

Bootstrap method

To quantify the uncertainty in the ensemble predictions, we adopted the bootstrap method. Each ensemble member was resampled randomly to achieve 10,000 realizations. In this random resampling process, any member was allowed to be selected again. The 95% confidence interval of the 10,000 realizations was calculated to quantify the uncertainty range.

Data availability

All the data used in this study are openly available. The ERA5 reanalysis dataset is available at https://doi.org/10.24381/cds.bd0915c6. The CN05.1 gridded temperature observation data is from https://ccrc.iap.ac.cn/resource.

Code availability

All source codes can be obtained upon request to the corresponding author.

Received: 5 July 2024; Accepted: 6 December 2024;
Published online: 14 January 2025

References

1. Christidis, N., Jones, G. S. & Stott, P. A. Dramatically increasing chance of extremely hot summers since the 2003 European heatwave. *Nat. Clim. Change* **5**, 46–50 (2015).
2. King, A. D. et al. Emergence of heat extremes attributable to anthropogenic influences. *Geophys. Res. Lett.* **43**, 3438–3443 (2016).

3. Perkins-Kirkpatrick, S. E. & Lewis, S. C. Increasing trends in regional heatwaves. *Nat. Commun.* **11**, 3357 (2020). 2020.
4. IPCC. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (Cambridge Univ. Press 2021).
5. Barriopedro, D. et al. The hot summer of 2010: redrawing the temperature record map of Europe. *Science* **332**, 220–224 (2011).
6. Ballester, J. et al. Heat-related mortality in Europe during the summer of 2022. *Nat. Med.* **29**, 1857–1866 (2023).
7. White, C. J. et al. Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteorol. Appl.* **24**, 315–325 (2017).
8. White, C. J. et al. Advances in the application and utility of subseasonal-to-seasonal predictions. *Bull. Am. Meteorol. Soc.* **103**, E1448–E1472 (2022).
9. Bauer, P., Thorpe, A. & Brunet, G. The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55 (2015).
10. Vitart, F. & Robertson, A. W. The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *npj Clim. Atmos. Sci.* **1**, 3 (2018).
11. Lin, H., Mo, R. & Vitart, F. The 2021 Western North American heatwave and its subseasonal predictions. *Geophys. Res. Lett.* **49**, e2021GL097036 (2022).
12. Xie, J. H., Yu, J. H., Chen, H. S. & Hsu, P.-C. Sources of subseasonal prediction skill for heatwaves over the Yangtze River basin revealed from three S2S models. *Adv. Atmos. Sci.* **37**, 1435–1450 (2020).
13. Qi, X., Yang, J., Gao, M., Yang, H. & Liu, H. Roles of the tropical/extratropical intraseasonal oscillations on generating the heat wave over Yangtze River Valley: a numerical study. *J. Geophys. Res. Atmos.* **124**, 3110–3123 (2019).
14. Koster, R. D. et al. Realistic initialization of land surface states: impacts on subseasonal forecast skill. *J. Hydrometeorol.* **5**, 1049–1063 (2004).
15. Guo, Z., Dirmeyer, P. A. & DelSole, T. Land surface impacts on subseasonal and seasonal predictability. *Geophys. Res. Lett.* **38**, L17808 (2011).
16. Dirmeyer, P. A., Halder, S. & Bombardi, R. On the harvest of predictability from land states in a global forecast model. *J. Geophys. Res. Atmos.* **123**, 13111–13127 (2018).
17. Fischer, E. M. et al. Soil moisture-atmosphere interactions during the 2003 European summer heat wave. *J. Clim.* **20**, 5081–5099 (2007).
18. Xue, Y. et al. Impact of Initialized Land Surface Temperature and Snowpack on Subseasonal to Seasonal Prediction Project, Phase I (LS4P-I): organization and experimental design. *Geosci. Model Dev.* **14**, 4465–4494 (2021).
19. Orsolini, Y. J. et al. Impact of snow initialization on sub-seasonal forecasts. *Clim. Dyn.* **41**, 1969–1982 (2013).
20. Williams, I. N. et al. Land-atmosphere coupling and climate prediction over the US Southern Great Plains. *J. Geophys. Res. Atmos.* **121**, 12125–12144 (2016).
21. Epstein, E. S. Stochastic dynamic prediction. *Tellus* **21**, 739–759 (1969).
22. Leith, C. E. Theoretical skill of Monte Carlo forecasts. *Mon. Weather Rev.* **102**, 409–418 (1974).
23. Buizza, R. Introduction to the special issue on "25 years of ensemble forecasting. *Q. J. Roy. Meteor. Soc.* **145**, 1–11 (2019).
24. Lavaysse, C. et al. Impact of surface parameter uncertainties within the Canadian regional ensemble prediction system. *Mon. Weather Rev.* **141**, 1506–1526 (2013).
25. MacLeod, D. A. et al. Improved seasonal prediction of the hot summer of 2003 over Europe through better representation of uncertainty in the land surface. *Q. J. Roy. Meteor. Soc.* **142**, 79–90 (2016). 2016.
26. Draper, C. S. Accounting for land model uncertainty in numerical weather prediction ensemble systems: toward ensemble-based coupled land-atmosphere data assimilation. *J. Hydrometeorol.* **22**, 2089–2104 (2021).
27. Bouttier, F. et al. Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX. *Q. J. Roy. Meteor. Soc.* **142**, 390–403 (2016).
28. Orth, R., Dutra, E. & Pappenberger, F. Improving weather predictability by including land surface model parameter uncertainty. *Mon. Weather Rev.* **144**, 1551–1569 (2016).
29. Gehne, M. et al. Land surface parameter and state perturbations in the global ensemble forecast system. *Mon. Weather Rev.* **147**, 1319–1340 (2019).
30. Mu, M., Duan, W. S., Wang, Q. & Zhang, R. H. An extension of conditional nonlinear optimal perturbation approach and its applications. *Nonlinear Process. Geophys.* **17**, 211–220 (2010).
31. Zhang, Q. Y., Sun, G. D., Dai, G. K. & Mu, M. Impact of uncertainties in land surface processes on subseasonal predictability of heat waves onset over the Yangtze River valley. *J. Geophys. Res. Atmos.* **129**, e2023JD038674 (2024). 2024.
32. Wang, L. et al. Model uncertainty representation for a convection-allowing ensemble prediction system based on CNOP-P. *Adv. Atmos. Sci.* **37**, 817–831 (2020).
33. Brier, G. W. Verification of forecasts expresses in terms of probability. *Mon. Weather Rev.* **78**, 1–3 (1950).
34. Matheson, J. E. & Winkler, R. L. Admissible scoring systems for continuous distributions. *Manag. Sci.* **22**, 1087–1096 (1974).
35. Mason, S. J. & Graham, N. E. Conditional probabilities, relative operating characteristics, and relative operating levels. *Weather Forecast.* **14**, 713–725 (1999).
36. Mason, S. J. & Graham, N. E. Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Q. J. Roy. Meteor. Soc.* **128**, 2145–2166 (2002).
37. Bowler, N. E. Comparison of error breeding, singular vectors, random perturbations and ensemble Kalman filter perturbation strategies on a simple model. *Tellus A.* **58**, 538–548 (2006).
38. Seneviratne, S. I. et al. Investigating soil moisture-climate interactions in a changing climate: a review. *Earth Sci. Rev.* **99**, 125–161 (2010).
39. Seo, E. et al. Impact of soil moisture initialization on boreal summer subseasonal forecasts: mid-latitude surface air temperature and heat wave events. *Clim. Dyn.* **52**, 1695–1709 (2019).
40. Miralles, D. G. et al. Land–atmospheric feedbacks during droughts and heatwaves: state of the science and current challenges. *Ann. N. Y. Acad. Sci.* **1436**, 19–35 (2019).
41. Bastos, A. et al. Direct and seasonal legacy effects of the 2018 heat wave and drought on European ecosystem productivity. *Sci. Adv.* **6**, eaba2724 (2020).
42. Nogueira, M. et al. Role of vegetation in representing land surface temperature in the CHTESSEL (CY45R1) and SURFEX-ISBA (v8.1) land surface models: a case study over Iberia. *Geosci. Model Dev.* **13**, 3975–3993 (2020).
43. Ruiz-Vasquez, M. et al. Exploring the relationship between temperature forecast errors and Earth system variables. *Earth Syst. Dynam.* **13**, 1451–1471 (2022).
44. Qin, B. et al. The first kind of predictability problem of El Ni¤o predictions in a multivariate coupled data-driven model. *Q. J. R. Meteorol. Soc* **150**, 5452–5471 (2024).
45. Mu, M. et al. The predictability study of weather and climate events related to artificial intelligence models. *Adv. Atmos. Sci*. https://doi.org/10.1007/s00376-024-4372-7 (2024).
46. Koster, R. D. et al. Contribution of land surface initialization to subseasonal forecast skill: first results from a multi-model experiment. *Geophys. Res. Lett.* **37**, L02402 (2010).
47. Yoon, D. et al. Role of land-atmosphere interaction in the 2016 Northeast Asia heat wave: impact of soil moisture initialization. *J. Geophys. Res. Atmos*. **128**, e2022JD037718 (2023).

48. Thompson, G. et al. Explicit forecasts of winter precipitation using an improved bulk microphysics scheme. Part II: implementation of a new snow parameterization. *Mon. Weather Rev.* **136**, 5095–5115 (2008).
49. Mlawer, E. J. et al. Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *J. Geophys. Res. Atmos.* **102**, 16663–16682 (1997).
50. Dudhia, J. Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. *J. Atmos. Sci.* **46**, 3077–3107 (1989).
51. Hong, S.-Y., Noh, Y. & Dudhia, J. A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon. Weather Rev.* **134**, 2318–2341 (2006).
52. Kain, J. S. The Kain-Fritsch convective parameterization: an update. *J. Appl. Meteor* **43**, 170–181 (2004).
53. Niu, G.-Y. et al. The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res. Atmos.* **116**, D12109 (2011).
54. Wu, J. & Gao, X. A gridded daily observation dataset over China region and comparison with the other datasets (in Chinese with English abstract). *Chin. J. Geophys.* **56**, 1102–1111 (2013). 2013.
55. Xu, Y. et al. A daily temperature dataset over China and its application in validating a RCM simulation. *Adv. Atmos. Sci.* **26**, 763–772 (2009).
56. Sun, G. & Mu, M. A new approach to identify the sensitivity and importance of physical parameters combination within numerical models using the Lund-Potsdam-Jena (LPJ) model as an example. *Theor. Appl. Climatol.* **128**, 587–601 (2017).
57. Sun, G., Mu, M., Zhang, Q., Ren, Q. & You, Q. Application of the CNOP-P ensemble prediction (CNOP-PEP) method in evapotranspiration forecasting over the Tibetan Plateau to model parameter uncertainties. *J. Adv. Models* **15**, e2022MS003110 (2023).
58. Storn, R. & Price, K. Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.* **11**, 341–359 (1997).
59. Hopson, T. M. Assessing the ensemble spread-error relationship. *Mon. Weather Rev.* **142**, 1125–1142 (2014).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grants 42288101, 42175077 and 42475063).

Author contributions

Q.Y.Z. designed the research under the supervision of M.M. and G.D.S. Q.Y.Z. performed the data analysis and led the writing of this paper. Q.Y.Z. prepared all the figures. All the authors contributed to the interpretation of the findings and paper revision.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41612-024-00876-y.

Correspondence and requests for materials should be addressed to Guodong Sun.

Reprints and permissions information is available at http://www.nature.com/reprints

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