The Soil Moisture–Surface Flux Relationship as a Factor for Extreme Heat Predictability in Subseasonal to Seasonal Forecasts

DAVID O. BENSON[a] AND PAUL A. DIRMEYER[a,b]

[a] George Mason University, Fairfax, Virginia
[b] Center for Ocean–Land–Atmosphere Studies, George Mason University, Fairfax, Virginia

(Manuscript received 15 June 2022, in final form 30 May 2023, accepted 31 May 2023)

ABSTRACT: Thresholds of soil moisture exist below which the atmosphere becomes hypersensitive to land surface drying, inducing thermal feedbacks that can exacerbate heatwaves. Realistic representation of threshold transitions in forecast models could improve extreme heat predictability and understanding of the role of land–atmosphere coupling. This study evaluates the performance of several forecast models from the Subseasonal Experiment (SubX) and several prototype versions of the Unified Forecast System (UFS) in their representation of threshold transitions by validation against reanalysis data. A metric of skill (true skill score) is applied to soil moisture breakpoint values, which mark the transition to heatwave hypersensitivity for drying soils. Forecast models have poor skill at being initialized on the correct side of the breakpoint, but show improvement when normalized to account for deficiencies in their soil moisture climatologies. Regionally, models performed best in the U.S. Northwest and worst in the Southwest. They effectively capture the tendency of western regions to spend more summer days in the hypersensitive regime than the eastern United States. Models represent well extreme heat as a consequence of atmospheric initial state for the first week of the forecast, but struggle to represent the soil moisture feedback regime. Forecast models generally perform better at extreme heat prediction when they are already dry and in the hypersensitive regime, even when erroneously so, implying that errors or biases exist in model parameterizations. Nevertheless, composite analysis shows encouraging model performance of the "hit" category, suggesting that an improvement in soil moisture initialization could further improve extreme heat forecast skill.

KEYWORDS: Land surface; Forecast verification/skill; Model initialization; Atmosphere-land interaction; Soil moisture

1. Introduction

In recent years, efforts have been made to improve our understanding of the characteristics of heatwaves in a changing climate. Numerous studies have been undertaken to gain knowledge on how the frequency, duration, and magnitudes of heatwaves evolve in light of the present increase in global mean temperatures, due largely to anthropogenic causes (Peterson et al. 2013; Perkins-Kirkpatrick and Lewis 2020; Schoof et al. 2019; Hirsch et al. 2021).

Models have made strides in simulating heatwave climatology and frequency but struggle to accurately depict the persistence and magnitudes of these events (Vautard et al. 2013; Lhotka et al. 2018). This inadequacy in our current abilities to represent extreme conditions has inspired the need for further investigation into the mechanisms that drive heatwaves, informing our understanding of predictability. Research on the connection between soil moisture and extreme hot temperatures has shown that low soil moisture is a potential precursor for intensive heatwave episodes (Herold et al. 2016) and that soil moisture memory is linked to heatwave persistence (Lorenz et al. 2010). Hence, proper initialization of soil moisture in forecast models could prove to be a crucial step in the advancement of heatwave predictability (Sillmann et al. 2017).

Specifically, soil moisture–temperature coupling relationships have been investigated in both observations and models using a variety of techniques (e.g., Ford and Quiring 2014; Gevaert et al. 2018), and these studies reveal that the precise representation of land–atmosphere feedbacks via a detailed expression of the soil moisture–temperature coupling relationship is an essential component for the predictability of hot extremes. Soil moisture is connected to daytime maximum temperature through its control on the partitioning of surface energy between sensible and latent heat fluxes (Miralles et al. 2012). When soil moisture over a region drops beyond a critical threshold (what we call here the breakpoint), the latent heat of evaporation substantially shuts down and the atmosphere responds to the land surface entirely through transfer of energy through sensible heat flux. The loss of evaporation as a negative feedback to heating moves the region into a hypersensitive regime where the atmosphere is primed for more intense heatwaves, provided the necessary atmospheric circulation conditions are in place (Benson and Dirmeyer 2021; Dirmeyer et al. 2021).

The idea of such soil moisture–climate regimes has been previously introduced in studies of the role of soil moisture in the variation of evapotranspiration (Koster et al. 2009a; Seneviratne et al. 2010; Schwingshackl et al. 2017; Haghighi et al. 2018; Denissen et al. 2020; Wu and Dirmeyer 2020; Sehgal et al. 2021). These studies have shown that there exist three distinct regimes of soil moisture–evapotranspiration relationships. There is a dry regime where soil moisture is too depleted to contribute to

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JCLI-D-22-0447.s1.

Corresponding author: David O. Benson, dbenson3@gmu.edu

DOI: 10.1175/JCLI-D-22-0447.1

evapotranspiration, a wet regime where evaporation is at the potential rate and soil moisture fluctuations have no impact, and a central transitional regime where soil moisture variations directly modulate evapotranspiration and has some feedback on the atmosphere.

Based on this framework, the study of Benson and Dirmeyer (2021) demonstrated the existence of comparable regimes based on the relationship between soil moisture and maximum daytime temperature. Benson and Dirmeyer (2021) showed that there is a wet soil, weakly coupled regime where there is sufficient availability of soil moisture and changes in near-surface air temperature are unaffected by soil moisture variation. In the transitional regime, decreasing soil moisture leads to diminishing latent heat flux, and increasing sensible heat flux, resulting in a net soil moisture feedback on maximum temperature similar to the transitional regime of evapotranspiration, but with temperature increasing as soil moisture declines. The third regime described as the hypersensitive regime is the state when soil moisture is sufficiently low that latent heat of evaporation shuts down, resulting in sensible heat flux being the dominant, unopposed driver of energy transfer from the land to the atmosphere. The distribution of soil moisture breakpoint values across the United States exposes regions that are more likely to experience the hypersensitive regime as soil moisture is depleted. These are regions where heatwaves can potentially persist or intensify as a result of subsequent thermal land–atmosphere feedback.

This understanding provides a way of evaluating subseasonal to seasonal (S2S) forecast models, to determine their proficiency in predicting heatwaves by observing their representation of breakpoints and soil moisture initialization relative to the breakpoints. Improving prediction of extremes within S2S time scales is valuable to decision makers and stakeholders who rely on this information to help mitigate the impact of weather and climate extremes on society.

In this study, an evaluation of S2S forecast models from the Subseasonal Experiment (SubX) forecast model suite—a multimodel project aimed at developing better research to operation models for S2S prediction (Pegion et al. 2019) and the Unified Forecast Systems S2S model prototypes (Xue et al. 2021) is carried out, with the intention of detailing the role of soil moisture and breakpoint representation in the skill of extreme heat day forecasts and heatwave predictability. The focus of this study is on whether the skill of breakpoint representation reflects on the skill of heatwave prediction. Our hypothesis is that accurately capturing the amount of time any given region spends in the hypersensitive regime as seen in observations corresponds to better skill in extreme heat day forecasts.

Section 2 details the data and methodology used in this paper, while section 3 lays out the results found in the study showing the models' distribution of breakpoint estimates across the United States and their corresponding skill scores. Section 4 contains a discussion of findings and the conclusions reached.

2. Data and methods

The region of focus is the contiguous United States (CONUS) and immediately adjacent regions of Mexico and Canada. Analyses are carried out over the summertime [June–August (JJA)] across the years available in both the reanalysis and forecast models' products. Forecasts are initialized from May through August but only forecasts that validate within JJA are used in the study. With the forecast models initialized at varying dates and frequencies, careful attention has been put into guaranteeing that validation dates line up accurately and that, as closely as possible, comparisons have been made between comparable forecast leads.

a. *ERA5*

The ERA5 (Hersbach et al. 2020) is used in this study as "verification" to assess the performance of the forecast models. Daily maximum 2-m air temperature and total column soil moisture (0–2.89 m) from the years that overlap within the model verification period (1999–2017) are obtained from a 31-km-resolution grid over the contiguous United States. The ERA5 is the first to employ remotely sensed soil moisture observations for assimilation—namely, soil moisture correlated backscatter data from the Advanced Scatterometer (ASCAT) onboard the *MetOp-A*/*-B* satellites. While this is beneficial, it is to be noted that one of the drawbacks to using satellite observations for data assimilation in the ERA5 is the simultaneous assimilation of other variables such as screen temperature and relative humidity can negatively affect the soil moisture estimates (Muñoz-Sabater et al. 2019). The availability of observations from ground stations also helps to constrain the 2-m temperature values, leading to a more reliable temperature product output, particularly over regions where there is a high density of in situ measurements such as CONUS.

The ERA5 dataset is selected for comparison with the forecast models for this analysis because of its reliable soil moisture breakpoint representation and spatial coverage. Recent studies have shown that ERA5 soil moisture and surface fluxes validate well against in situ observations over CONUS (Benson and Dirmeyer 2021) and Europe (Dirmeyer et al. 2021) in scenarios of extreme heat.

b. *Forecast models*

1) The Subseasonal Experiment (SubX) project

The SubX project is a multimodel endeavor to provide retrospective and real-time forecasts to understand the predictive skill of climate events at lead times out to 4 weeks (Pegion et al. 2019). Participating SubX models provide forecasts for at least 32 days after initialization and provide daily output interpolated to a common $1^\circ \times 1^\circ$ global grid. Only hindcasts are used for this experiment. Interpolation was performed without regard for underlying ocean versus land surface, so grid cells around coastlines mix the two surfaces and may be questionable. Soil moisture and surface fluxes are not available from all models involved in the SubX project. Furthermore, the SubX specifications for volumetric soil moisture data output is for the entire column and surface soil moisture data are unavailable for all but one model. Therefore, column soil moisture is used in this study only from models that have a constant column depth. Two models from the SubX suite are used in this study as listed in Table 1. They are the NCEP Environmental Modeling Center, Global Ensemble Forecast

TABLE 1. Specifications for the SubX models used in this study, showing the number of ensembles, length of forecast, time span, soil moisture availability, and model grid resolution.

SubX model	Ensemble members	Forecast length (days)	Years	Initialization frequency (days)
EMC-GEFS	11	35	1999–2016	7
ESRL-FIM	4	32	1999–2017	7

System (EMC-GEFS) and the National Oceanic and Atmospheric Administration, Earth System Research Laboratory, Flow-Following Icosahedral Model (ESRL-FIM). The EMC-GEFS model has T574 horizontal resolution (approximately 30 km), and the Noah land surface model (LSM; described below) is initialized with initial conditions from the Global Land Assimilation System (GLDAS). The ESRL-FIM model at ~60-km resolution also runs a version of the Noah LSM but with initial conditions from the Climate Forecast System Reanalysis (CFSR).

2) THE UNIFIED FORECAST SYSTEM (UFS) SUBSEASONAL TO SEASONAL (S2S) PROTOTYPES

The UFS S2S prototypes are part of NOAA's research to operations (R2O) effort to produce the next generation of weather and climate forecast models for the National Weather Service. Each prototype global model is a step toward a final operational version and sets of individual (no ensembles) subseasonal retrospective forecasts have been produced as part of the development effort to assess model behavior, biases and skill. The prototypes consist of a four-way coupled atmosphere–ocean–ice–wave model with ~25-km horizontal resolution and 64 vertical levels for the atmospheric component. The ocean and ice model components operate at a 1/4° resolution (Xue et al. 2021). Prototypes 5 and 6 (P5 and P6) employ a version of the Noah LSM similar but not identical to those in GEFS or FIM. All versions of the Noah LSM are a vertically one dimensional four-layer model that represents surface and subsurface layer processes including energy processes at the surface (Ek et al. 2003). Beginning with prototype 7 (P7), UFS uses the candidate GFSv17 physics package that significantly includes a change of LSM to Noah with multiple parameterizations (Noah-MP), which is designed to improve the simulation of water and heat exchanges over the land surface (Barlage et al. 2015; Salamanca et al. 2018). Among other changes, this physics package also has an improved boundary layer parameterization. JJA maximum temperature and soil moisture from the retrospective forecasts for years 2011–17 from P5, P6, and P7 are used in this study. The forecasts are initialized on the 1st and 15th of every month. The frequency and years covered are both substantially smaller than for the SubX models. The UFS prototypes are included in this analysis to improve sample size, due to the limited number of SubX models with output datasets that are relevant to this analysis, but also as an opportunity to provide evaluation of the performance of the UFS model while still under development.

c. *Breakpoint estimation*

To characterize the relationship between extreme dry conditions and elevated temperature, soil moisture breakpoint analyses have been carried out on the ERA5, the S2S forecast models, and the UFS prototypes using the methodology described in Benson and Dirmeyer (2021). The calculated breakpoint indicates a threshold for soil moisture below which the land–atmosphere coupling feedback shifts into a hypersensitive regime with regard to daytime maximum temperature (Fig. 1), exhibiting a significantly steeper slope of temperature increase with declining soil moisture.

Breakpoint estimation involves fitting of a regression model of daily values of maximum temperature against soil moisture with piecewise linear relationships by estimating optimal changepoints (minimizing fitted temperature mean square error) in the slope of time series at each grid cell over the study area. This is done across all dates in JJA using segmented regression (Muggeo 2008). For this study, as in Benson and Dirmeyer (2021), the Python SciPy "optimize" package's function "curve_fit" is employed. Four quantities are estimated: the values of the breakpoint along the abscissa and ordinate and the slopes of the two lines intersecting at the breakpoint.

The piecewise linear regression is applied separately at each grid cell of ERA5 and each forecast model for the independent variable daily mean soil moisture and the dependent variable daily maximum 2-m temperature. The breakpoints of the forecast models are investigated as a function of the forecast lead times based on dates of validation falling within JJA, not the date of forecast initialization.

d. *Extreme heat days*

An extreme heat day (EHD) is defined in this study as any day that exceeds the 90th percentile for daily maximum 2-m air temperature during JJA in the period of data record, defined at each grid point relative to its own climatology, and without consideration for persistence over some number of consecutive days. This definition only takes into consideration

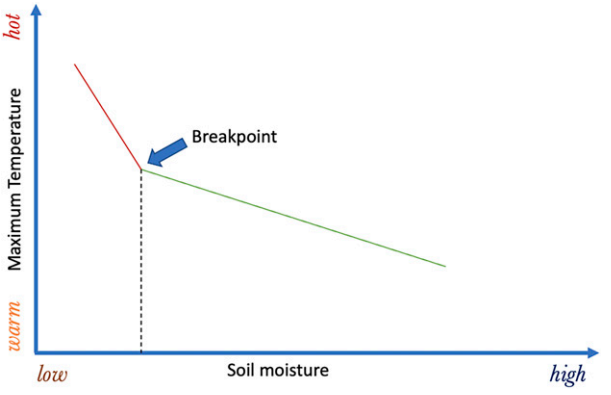

FIG. 1. Schematic illustrating breakpoints of soil moisture for maximum temperature.

the physical response of air temperature to surface level heating on a day-by-day basis, potentially dictated by soil moisture, without accounting for duration, thermal comfort, and other variables that would typically constitute a classical heatwave definition. In this way, the impact of land–atmosphere coupling on increasing temperatures can be approached from a process-based standpoint: Is soil moisture a contributing factor to extreme heat? For the forecast models, the extreme heat day is also calculated as a function of lead time.

e. Skill scores

A set of dichotomous skill scores are employed for the verification of model skill. A simple contingency table (Table 2) is created from the distribution of the frequency of a combination of yes and no forecasts and validations for a predetermined threshold. There are four categories: "hits" and "correct negatives" indicate correct forecast of the occurrence or nonoccurrence of an event respectively. "Misses" are events that were not forecasted, and a "false alarm" is a model forecast of an event that did not occur. The event in this study is defined as a day when soil moisture lies below the breakpoint, or an EHD for temperature. The ensemble members for the GEFS and ESRL models were aggregated during the calculation of skill scores to increase their sample sizes.

The true skill score (TSS; also called the Hanssen and Kuipers discriminant) is one of two dichotomous skills scores used in this analysis:

$$\mathrm{TSS} = \frac{\mathrm{hits}}{\mathrm{hits} + \mathrm{misses}} - \frac{\mathrm{false\ alarms}}{\mathrm{false\ alarms} + \mathrm{correct\ negatives}}. \tag{1}$$

TSS ranges from −1 to 1. A score of 0 indicates the threshold for no skill (Woodcock 1976). TSS does not differentiate between the ability to predict events and nonevents. It is therefore less sensitive to event frequency and useful in verifying forecasts with varying climatology (Tartaglione 2010).

The equitable threat score (ETS; also called the Gilbert skill score) is also utilized:

$$\mathrm{ETS} = \frac{\mathrm{hits} - \mathrm{hits}_{\mathrm{random}}}{\mathrm{hits} + \mathrm{misses} + \mathrm{false\ alarms} - \mathrm{hits}_{\mathrm{random}}}, \tag{2}$$

where

$$\mathrm{hits}_{\mathrm{random}} = \frac{(\mathrm{hits} + \mathrm{misses})(\mathrm{hits} + \mathrm{false\ alarms})}{\mathrm{total}}. \tag{3}$$

ETS ranges from −1/3 to 1. A score of 0 indicates no better than a random forecast (Woodcock 1976). ETS penalizes the model for missing events. Hence, in the forecast of rare events like the tail of a distribution such as extreme heat, ETS produces lower forecast skill scores that might be misleading (Stephenson et al. 2008; Tartaglione 2010). However, when there are larger forecast errors, the uncertainty of ETS is reduced when compared to TSS uncertainties. Stanski et al. (1989) provides greater detail on these verification methods. The skill scores are evaluated as a function of forecast lead times just like with the breakpoint estimates.

TABLE 2. Contingency table for categorizing features of forecast performance.

		Validation	
		Yes	No
Forecast	Yes	Hits	False alarms
	No	Misses	Correct negatives

In this study, contingency tables are constructed based on two different dichotomous categorizations. First, the soil moisture values for each model forecast at each grid cell are examined to validate whether they fall below or above its breakpoint. On any given day at a particular grid point, if the model soil moisture is on the same side of its breakpoint as ERA5, it is considered to be skillful. Since our interest is in extreme heat and we assume that correlates with dry soils, we define a "hit" as a match on the dry side of the breakpoint, and a "correct miss" as a match on the wet side. The values of model breakpoint values are not directly compared to those from the reanalysis because soil moisture is derived in the models through a variety of methods, and different LSM volumetric soil moisture ranges vary even for the same location due to differences in the parameterizations (Koster and Milly 1997; Koster et al. 2009a) and can also be seen in Fig. 1. A good skill score indicates that a model is able to effectively capture the soil moisture regime and the shifts as they occur as part of the land–atmosphere coupling relationship.

Second, the skill of the forecast models' extreme temperatures is also assessed by comparing the extreme heat day of the models to the reanalysis. Whereas the number of days on one or the other side of soil moisture breakpoint can vary greatly with location or across models, our definition of extreme heat days ensures a fixed ratio of 1:9 or 10% between columns or rows. The skill of each model is investigated as a function of the forecast lead time. For both soil moisture and temperature, skill at lead 0 is largely a reflection of the quality of model initialization, which is also crucial for the model's forecast skill.

Composites are made combining the contingency tables to determine if the skill of breakpoint estimation has any influence on the skill of extreme heat forecasts. Extreme heat day skill scores are calculated in four subsets using the results of the contingency table (Table 2) for breakpoint skill. The extreme heat days skill score for when the soil moisture relative to its breakpoint forecast was a hit, miss, false alarm, and correct negative is determined. This means that for each model, the extreme heat day skill is assessed for when the model is accurately versus inaccurately in the hypersensitive regime. Composite time series of the skill scores are generated for each element of the contingency table relative to the forecast lead time. Figure S1 in the online supplemental material presents a flowchart of the methods described in section 2c through section 2e.

3. Results

a. Soil moisture breakpoint climatology

Figure 2 displays the JJA mean volumetric soil moisture values from the initial states of the S2S models (Figs. 2a,b) and UFS prototypes (Figs. 2c–e), with the climatology of the

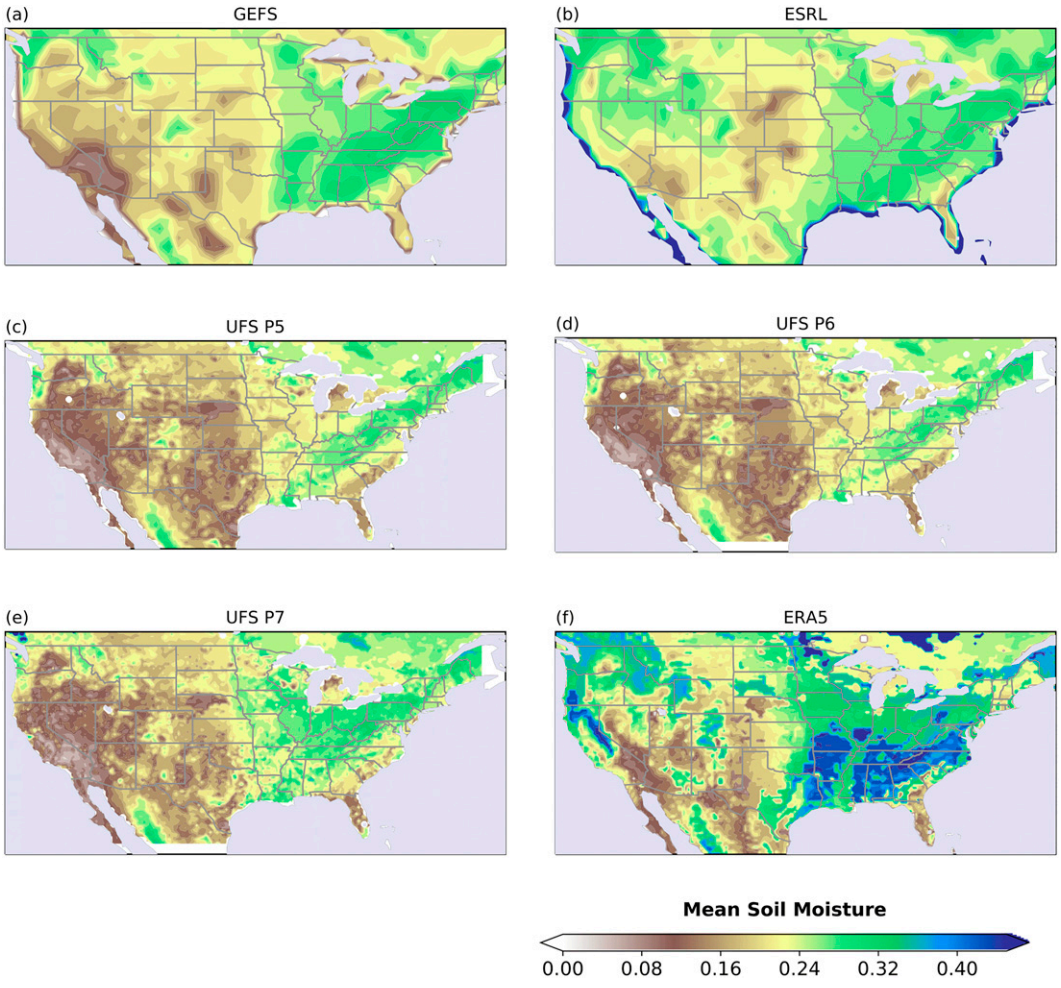

FIG. 2. Maps of mean volumetric soil moisture ($m^3\ m^{-3}$) June–August (JJA) climatology for (a),(b) GEFS and ESRL models; (c)–(e) UFS P5, P6, and P7 prototypes; and (f) ERA5, used as verification.

ERA5 (Fig. 2f). The soil moisture climatology from the reanalysis reveals dry soils in the southwestern region of the United States and the Great Plains. The eastern portion of the study area is wetter with higher values of soil moisture in the Mississippi Basin stretching out to the U.S. East Coast. The U.S. West and Northwest regions also have wetter soils due to the influence of the northern Pacific Ocean storm tracks on precipitation over that region. Mean soil moisture values from all forecast models and prototypes exhibit varying magnitudes and spatial structures for their soil moisture climatologies. However, all models display a west–east gradient of increasing wetness, but appear to have mean values that are significantly drier than ERA5.

The differences among the UFS P5 and P6 and the UFS P7 prototypes are largely a result of the changes in the LSM from P6 to P7. The summer soil moisture climatology in Fig. 2 highlights the variations in the models' soil moisture estimates and could potentially dictate the sensitivity and regime shifts in the models. Total column volumetric soil moisture is not ideal as subsurface soil moisture can decouple from the surface in dry conditions (Qiu et al. 2016) and previous results have shown that breakpoint values are mainly driven by the influence of surface soil moisture changes on sensible heat flux (Benson and Dirmeyer 2021). However, latent heat flux and the wilting point also play an important role in breakpoint estimation, involving water content across the column. The shutting down of latent heat flux, which is essential for the breakpoint process, involves the accessibility of root zone soil moisture for evapotranspiration and is a function of the soil type. Because of this, the breakpoints of surface soil moisture based on sensible heat flux and latent heat flux are not identical (Benson and Dirmeyer 2021). Yet they are similar, hence using the entire column soil moisture should provide useful information.

Figure S2 in the online supplemental material shows the mean soil moisture climatology minus the breakpoint estimates relative to maximum temperature. The breakpoint values in ERA5 seem to be near or slightly lower than the summertime climatology of soil moisture, particularly along the Great Plains and western Mexico (Fig. S2f). Otherwise, the mean summer

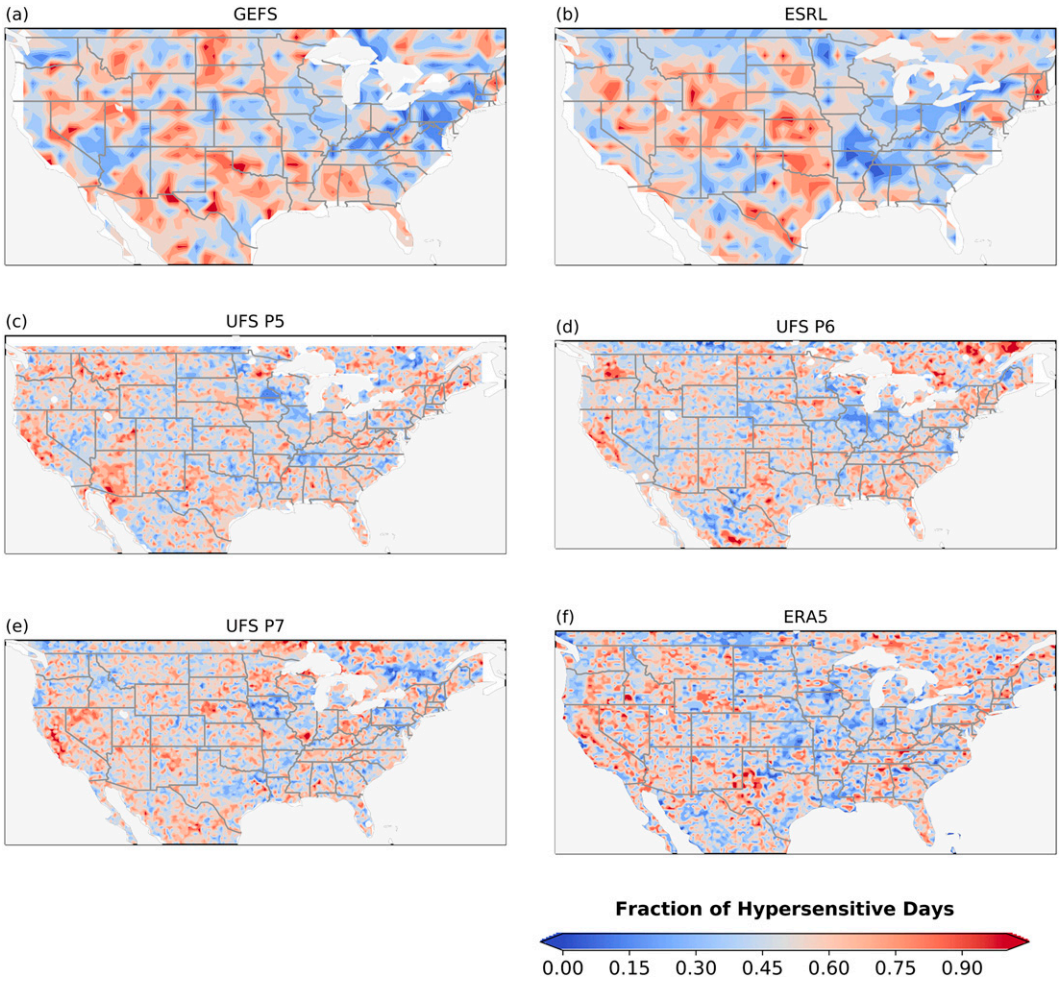

FIG. 3. Maps of the fraction of JJA days in the hypersensitive regime (relating daily maximum 2-m air temperature to soil moisture) for (a),(b) GEFS and ESRL models at lead 0; (c)–(e) UFS P5, P6, and P7 prototypes at lead 0; and (f) ERA5.

climatology of soil moisture for ERA5 (Fig. 2f) is very close to the breakpoint values. Regions in the West and some areas of the southeastern CONUS particularly show breakpoint values that are higher than the summer climatological soil moisture values, which would imply that such regions spend the majority of the summer in the hypersensitive regime.

At initialization, the GEFS model shows breakpoint estimates that are mostly equal to or drier than their average soil moisture value except in the South (Fig. S2a). The ESRL model (Fig. S2b) is similar to the reanalysis but the breakpoint values over the Great Plains appear to be about equal to the average soil moisture values. The lower values of breakpoints tend to occur in the Midwest between the Mississippi River and the Appalachians. The UFS prototypes (Figs. S2c–e) show breakpoint values that are generally higher than the summertime average soil moisture except in the southern Great Plains, south of the Great Lakes, and Southeast for UFS P5 and P6. The change in LSM in UFS P7 is evident as it has lower breakpoint values than its previous iterations but does not show the lower breakpoint values in the Great Plains. The pattern of breakpoint distribution in the ERA5 is most closely mirrored by the GEFS model in regard to the lower breakpoint values in the Great Plains.

Regions in the central United States away from the coasts are often regions where breakpoint values are below the climatological mean soil moisture value. There are studies that have identified the Great Plains as a transition zone between the dry and moisture-limited regimes that is a hotspot of land–atmosphere coupling (e.g., Koster et al. 2009b; Dirmeyer 2011).

The fraction of all summer days that are in the hypersensitive regime (Fig. 3) quantifies how often a location is preconditioned to experience potential soil moisture feedback-driven intensification or persistence of extreme heat days. Most of the United States has a large fraction of the summer days in the hypersensitive regime in ERA5 (Fig. 3f). Though not uniformly true, much of the Mississippi basin tends to have relatively few days that are in the hypersensitive regime. The overall zonal gradient, which

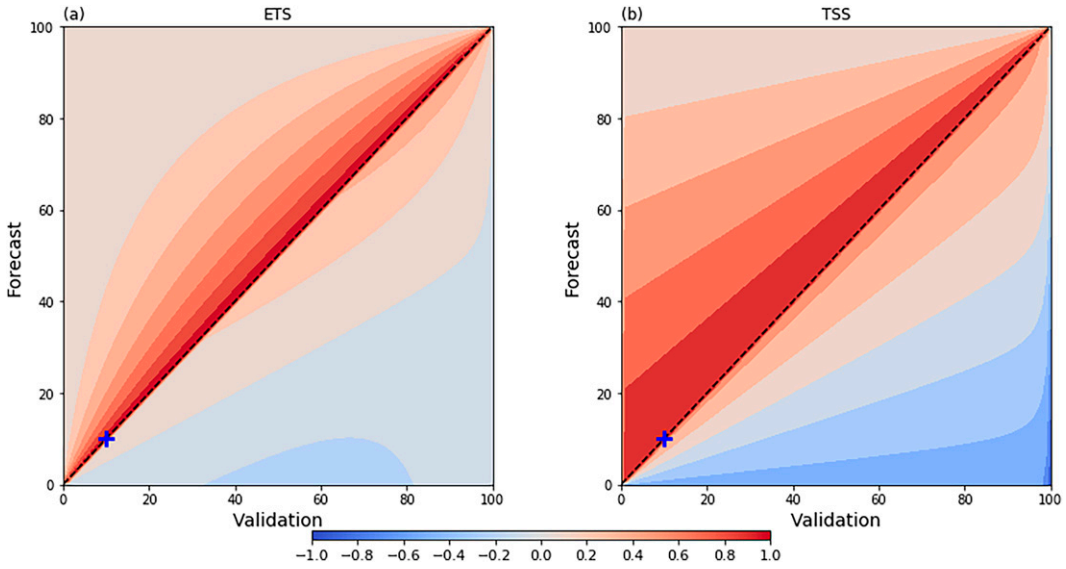

FIG. 4. Maps of best attainable skill score for (a) ETS and (b) TSS based on the percentage of events that count as a "hit" in the validation data (x axis) and the forecasts (y axis). Blue crosses indicate the EHD scenario in this study where there is an equal 10% of the days in both the validation and forecast sets of events.

ranges westward into regions that spend more than half the summer days in the hypersensitive regime, has also been seen clearly in surface layer soil moisture breakpoint estimates (Benson and Dirmeyer 2021). Regardless, the regions west of the Great Plains as well as parts of the southeastern United States appear to have many grid cells that spend more than half of summer days on the dry side of the local breakpoint in ERA5.

This pattern is generally replicated in the forecast models, albeit with differing strengths. GEFS (Fig. 3a) does not produce some of the low values in the central United States seen in ERA5. The ESRL model (Fig. 3b) tends to miss the low fractions in the Mississippi Basin apparent in other models and ERA5, placing them instead along the Gulf Coast. The UFS models (Figs. 3c–e) do better at matching the pattern of the fraction of hypersensitive days from the reanalysis (Fig. 3f). They all have high values in the West, low values in the central United States, and high values in the eastern portion of the United States. The UFS forecasts also seem to match ERA5 better across each iteration from P5 to P7 in this metric.

The low values over the Mississippi basin suggest that one factor that determines whether a region is in the hypersensitive regime is the amount of surface soil water content that is readily available for evapotranspiration. This is consistent with the idea that having higher fractions of hypersensitive days is driven by the lack of available soil water for evapotranspiration.

b. *Breakpoint skill*

Typically, dichotomous skill scores are implemented to compare forecasts to validation over a fixed threshold that allows for the same fraction in both datasets to be compared, and the possibility for a perfect score to be achieved. For example, the extreme heat validation compares the top 10% of maximum temperature in the forecast to the same percentile in validation data, and a perfect model would mean that in both datasets the thresholds are exceeded on exactly the same days. In a case where the percentage of validating events does not match the percentage of forecast events, the forecast cannot achieve a perfect score, hence the interpretation of model skillfulness needs to be reconsidered. As illustrated in Fig. 3, this is usually the case for the number of hypersensitive days. Figure 4 illustrates how different counts for "hits" and "correct misses"—the diagonal of the contingency table that indicates the number of events and nonevents accurately predicted by the model (Table 2)—affect the maximum skill score values.

The ETS score (Fig. 4a) is less forgiving, especially if there are more days defined as a "hit" in the validation dataset than for the forecast model. The TSS score (Fig. 4b) is more asymmetric about the diagonal because it does not punish the models as strongly for false alarms. In both skill scores, a model is not penalized as harshly if it overcounts hypersensitive days but is punished strongly for undercounting them, more so in the case of the TSS score.

The number of days with soil moisture content below the breakpoint is dependent on the range of soil moisture of the grid cell and on its breakpoint value. For any location to achieve a perfect skill ETS or TSS value, not only does it have to correctly have soil moisture below the breakpoint value on exactly the same days as the validation dataset, but it would also have to have the same number of days below the breakpoint. Figure 4 provides a way to provide point-by-point scaling for skill scores that account for the differing soil moisture climatologies relative to their estimated breakpoints. This puts the skill scores for breakpoints into perspective when interpreting the results from the skill assessment.

Figure 5 shows models' skill to forecast soil moisture on the dry side of the breakpoint as a function of forecast lead using

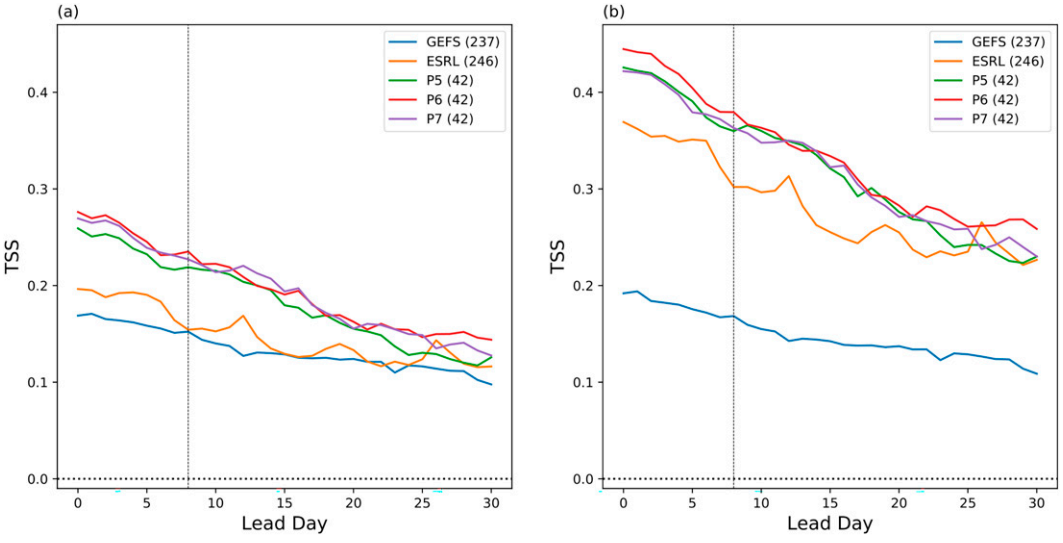

FIG. 5. Time series of TSS for estimating model soil moisture relative to its breakpoint showing the evolution of skill with respect to lead time for (a) the nonnormalized score and (b) the normalized score. Each lead is the spatially averaged skill score over the study region.

TSS. Results are shown before and after renormalization of skill scores using information in Fig. 4. We validate against ERA5 because previous work has shown ERA5 compares well to in situ measurements in representing the soil moisture–temperature relationship and breakpoint values (Benson and Dirmeyer 2021; Dirmeyer et al. 2021). A direct comparison of the models' soil moisture breakpoint values to ERA5 would be an unfair measure of the model skill due to the differences in the model derivations of soil moisture (Koster et al. 2009a). Hence, each model is judged relative to its own breakpoints.

The grid cell skill scores over the continental United States have been spatially averaged for each forecast lead time to generate Fig. 5. Ideally, any value above 0 for either ETS or TSS is considered skillful based on both measures of skill applied in this study. The UFS prototypes, however, have a very limited sample, available only for single forecasts initialized on the 1st and 15th of the JJA months over 7 years. The results obtained from this skill assessment only provide a glimpse into their current performance with the expectation that the final product will improve on the findings in this study.

Focusing first on initialization (lead 0), the forecast models start out with imperfect skill even at initialization, demonstrating potential problem in model initialization techniques. The UFS prototypes on average appear to be more skillful than the two SubX models at having soil moisture on the correct side of the breakpoint. ESRL and GEFS SubX models are also skillful at capturing this relationship, albeit less so than the UFS prototypes. There is noticeable decay in skill as lead time increases for all models. The skill of the UFS prototypes decays at a faster rate than the ESRL and GEFS models, becoming comparable by the end of the forecast. The low skill scores in Fig. 5a arise by a combination of low hit counts from the models on the dry side of the breakpoint, and the significant disparities in the number of hypersensitive days over the summer period between the reanalysis and the models.

The models' skill in representing soil moisture relative to the breakpoint is normalized by the best attainable skill in an effort to isolate the impact of poor initialization and predictive skill from the inaccurate soil moisture climatology (Fig. 5b). When normalized, the overall TSS scores improve, but a clear separation emerges between GEFS and the other forecast models. GEFS has the highest potential skill scores compared to the other models, indicating its climatological distributions of soil moisture relative to the breakpoint agree well with ERA5. GEFS is thus underperforming in Fig. 5b either because of a problem with the initialization of soil moisture in the model, or a general inability to predict extreme heat regardless of land surface conditions. The significantly lower skill of GEFS at 0-day lead in Fig. 5b argues for poor soil moisture initialization.

Meanwhile, the ESRL model and UFS prototypes show much more of their potential forecast skill is being realized. Nevertheless, these models would likely benefit from an improved representation of model soil moisture distribution relative to the breakpoint and also soil moisture initialization. Getting the climatology right would likely improve the skill in these models.

The vertical lines in Fig. 5 are drawn across a lead that shows maximum separation in the skill scores between the UFS prototypes and other models: lead day 8. Maps of TSS for each model at this lead are shown in Fig. 6. In general, the models do better in the eastern part of the United States where breakpoint values are found to be relatively higher, but soil moisture is also high. The most skillful regions for most models are found in the areas from the Mississippi basin eastward and in the Pacific Northwest. The models likely perform better in these regions because the soil is usually on the wetter side of the breakpoint. We find the high skill scores in this region are mostly skewed toward

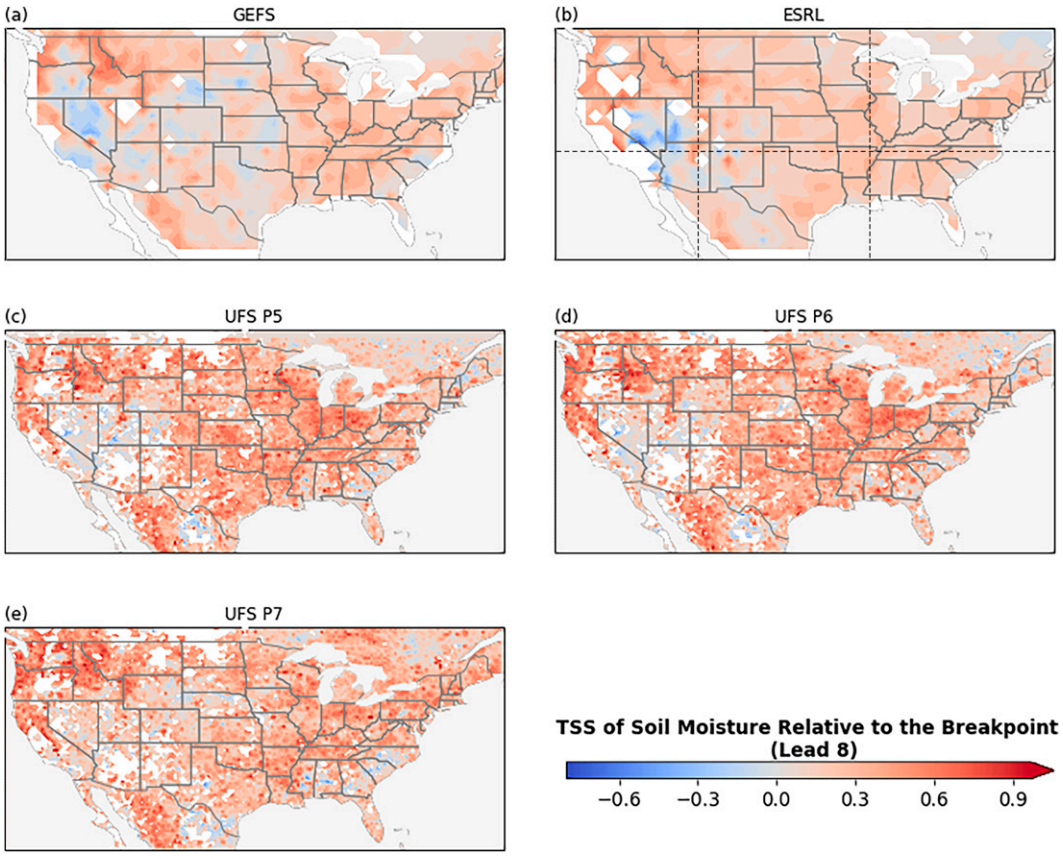

FIG. 6. Maps of the TSS of soil moisture relative to the breakpoint for lead day 8. The dashed lines across Fig. 6b indicate the division of the study area into regions in subsequent analyses. The white colored areas are grid cells that failed the threshold for the determination of breakpoint significance.

correct negatives (not shown)—that is, the correct prediction of soil moisture above the hypersensitive regime. The models appear to have difficulty in areas where breakpoint values are low in ERA5, especially in the Southwest. Accurate simulation of land–atmosphere interactions in semiarid regions with low-resolution models is a demonstrated problem, where evaporation over small riparian zones can have major effects on regional temperature (Barlage et al. 2021).

GEFS and ESRL are very similar in their spatial structure of skill. The UFS prototypes appear noisier partly because of smaller sample sizes and partly due to their higher spatial resolution, leading to more small-scale features. The P7 prototype (Fig. 6e) appears to be less skillful than the other prototypes over the Mississippi basin. This finding raises questions because P7 employs the newer Noah-MP land surface model and has an upgraded boundary layer physics package that is expected to improve skill.

Figure S3 shows the difference between the normalized and original TSS scores from Fig. 6, highlighting regions that are most affected when soil moisture climatology differences are properly accounted for. The GEFS model shows only slight improvement (Fig. S3a). The ESRL model and UFS prototypes seem to have the majority of their adjustments in the eastern and central United States, with the ESRL model showing the greatest adjustment over the southern Great Plains (Fig. S3b).

The ETS scores for soil moisture breakpoint as a function of forecast lead in Fig. S4 reveal that when the models are being penalized for missing forecasts; that is, not correctly being on the dry side of the breakpoint, they perform worse than the TSS scores. For the nonnormalized ETS scores (Fig. S4a), GEFS slightly outperforms ESRL. A distinction in skill of the ESRL and UFS models from those of the GEFS model is evident after normalizing for soil moisture climatology relative to the breakpoint (Fig. S4b). When normalized by their best attainable skill scores (Fig. S4b), the models all depict vast improvement in skill, especially the ESRL model, which becomes the most skillful after the first few days of the forecast. The skill values are mostly dictated by the low count of hits, and the counts of false alarms and misses. This skill is skewed toward the forecast of nonevents or correctly placing soil moisture in the regime of less sensitivity.

The regional variations of the skill scores, particularly the TSS scores, are difficult to discern unequivocally from the maps and requires further examination. By dividing the study area into six regions (land areas marked in Fig. 6b)—the Northwest (NW), North Central (N), Northeast (NE), Southwest (SW), South Central (S), and Southeast (SE)—the regional characteristics of

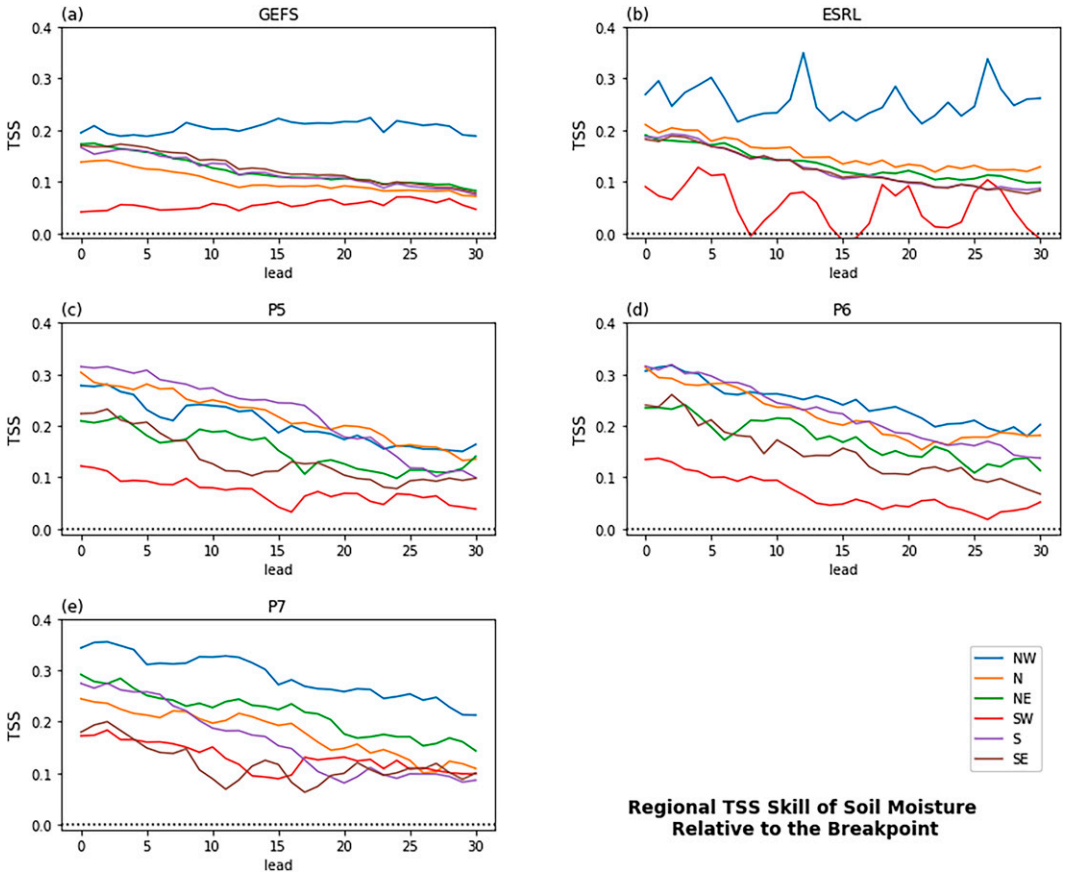

FIG. 7. Regional dependency of the TSS for soil moisture relative to the breakpoint across forecast lead days.

skill are more clearly revealed (Fig. 7). In the GEFS and ESRL models the Northwest shows the highest skill, while both models show poor skill in the dry Southwest region. However, both regions seem to have similar characteristics as lead increases, indicating a west–east element in soil moisture values relative to breakpoints. These models' performance is very consistent through the 30-day forecast period for both regions as well, showing little or no decline. This persistence in skill might be an indication of high soil moisture memory during summer in the West (e.g., Dirmeyer et al. 2009), highlighting the importance of proper model initialization in both regions. The UFS prototypes also perform poorly in the Southwest, but do not show as much separation in skill from the other regions, especially for P7, which also has the greatest skill in the Northwest. This is another result that may be affected by the smaller sample size for UFS. However, the evidently weaker soil moisture memory for the UFS prototypes over the western regions is an interesting feature.

c. *Extreme heat day skill*

The root-mean-square error (RMSE) of maximum temperature for day 1 (lead 0) is shown in Fig. 8. All models and prototypes indicate a roughly south–north gradient of decreasing fidelity in representing maximum temperature on the first afternoon of the forecast across the United States. While extreme heat is defined as a quantile relative to each model's climatology, the RMSE is an unqualified error. Although the emphasis of this study is on extreme heat, this result is nonetheless a reflection of the models' initialization quality. Understanding the source of the forecast errors may provide relevant information in the determination of model processes that impede their ability to predict extremes.

Consequently, the model mean bias is also calculated to look for systemic temperature biases within the models and to provide context for the succeeding analysis of extreme heat days. All models have a cold bias in maximum temperature over the Rockies, and all except UFS P7 have a warm bias centered in the Great Plains (Fig. S5). The SubX models tend to experience stronger and broader warm biases, while UFS prototypes tend toward a notable cold bias that is strongest in P7. ESRL has most of its warm bias in the Great Plains and both ESRL and GEFS are too warm over coastal California, where their low model resolutions may struggle to represent the effects of coastal mountains. The UFS models and GEFS also have a warm bias over the Southeast, while all models show a cool bias over the Northwest. Figure S5 suggests that the RMSE patterns of the models at initialization are a result of a combination of systematic biases in the models and issues with initialization.

Figure S6 depicts the value of the 90th percentile of maximum temperature over JJA for the 40-yr period from 1979 to

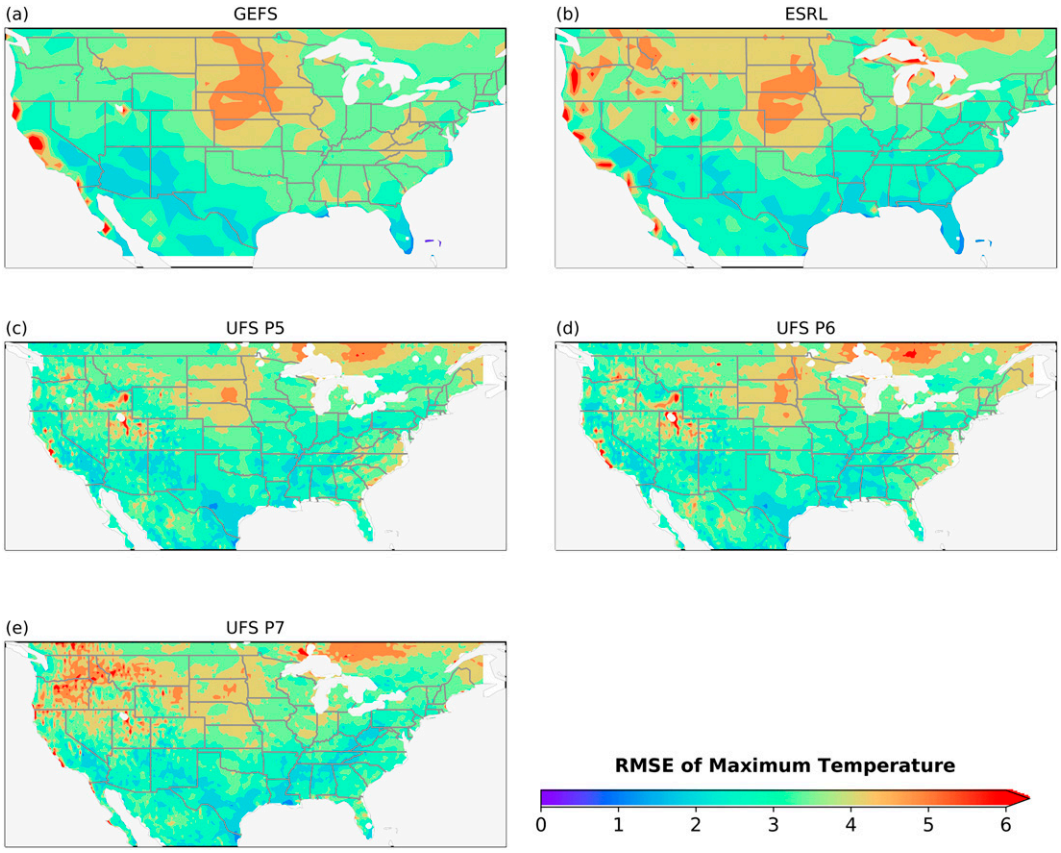

FIG. 8. RMSE of maximum temperature (°C) at lead 0 (within the first 24 h of the start of the forecasts) for SubX models and UFS prototypes.

2018 for ERA5—the reference threshold for EHD in the reanalysis. The Southwest experiences the highest 90th-percentile maximum temperature threshold of up to 45°C. A southwest–northeast gradient can be seen with the U.S. Northeast having average EHD thresholds around 25°C. High terrain also has lower threshold temperatures. Admittedly, it is understandable why there might be a hesitation to categorize seemingly mild temperature values as an extreme without taking into account other factors that go into the classical definition of a heatwave. However, a definition based on the statistical threshold of quantiles allows for an assessment of the physical relationship based squarely on the available data.

Each models' ability to forecast EHD relative to its own climatology has been determined. Figure 9 shows the domain mean TSS for each model's EHD forecasts as a function of forecast lead. All grid cells for a fixed validation time period have the same number of EHD days, so the best possible skill score at any location is 1.0 (Fig. 4) and no adjustments are necessary. The models are assessed to determine whether there is an EHD on the same date for the same location as in the ERA5. The ESRL model outperforms the other models in this study during the first two weeks of the forecast. The P7 prototype outperforms the other UFS prototypes for at least the first 13 days. Overall, the models perform better at representing extreme heat for the first 7–8 days than they do at having soil moisture at the correct side of the breakpoint (Fig. 5a), but afterward their TSS skill for EHD decays faster as forecasts lose contributions to skill from initial conditions.

Lead day 7 shows distinct spread across models, so it is selected to observe the nature of the skill distribution across the United States (Fig. 10). Here, the small sample size for the UFS forecasts is evident in the spatial noisiness of the results. GEFS appears to have no major regional dependency on EHD skill. Most models show the highest skill over the southern and southeastern United States and northern Mexico, while the UFS prototypes have generally high skill along the East Coast. UFS does better at representing hot days in regions that typically have relatively lower heat anomalies as opposed to regions in the West that are inclined to have very high temperature during hot anomalies, particularly in the Southwest. UFS P7 shows improvement over P6 and P5, but it still struggles with skillful representation over the northern Rockies and the Northwest.

The time series of ETS scores show the models being skillful at initialization due to a significant hit count in EHD, but they become indifferentiable after the first week of the forecast (Fig. S7). UFS P7 performs best in this skill score while UFS P5 and P6 are no longer the worst performers. Considering that ETS scores are typically lower for rare events such as extreme heat, the results show that the models are skillful at initialization but lack persistently skillful forecasts beyond the first week.

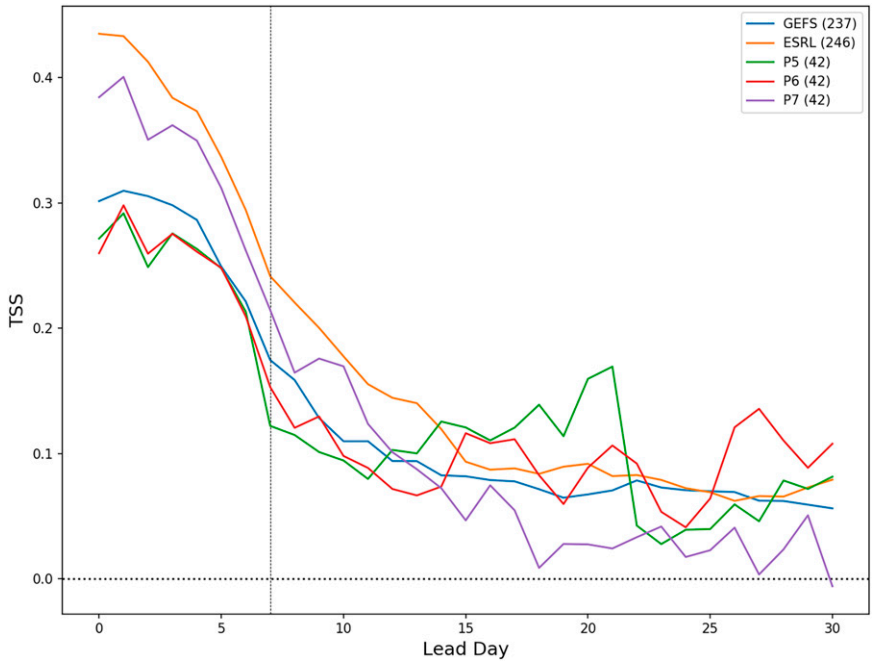

FIG. 9. TSS of extreme heat days spatially averaged across CONUS at each lead time for JJA.

A further look into the spatial distribution of the TSS skill scores shows that the models do not show much discernable regional dependency (Fig. 11), in stark contrast to the soil moisture results (Fig. 7). The limited sample size for the UFS prototypes leads to the extra noisiness seen in the time series (Figs. 11d–f) and could be responsible for the apparent increase in skill in the Northwest and Southwest regions in week 4. UFS P7 also has clear regional separation at day 0 with the Northwest performing best and the Southeast worst, but separation is lost after a few days. A larger set of forecasts with ensembles would help determine if this separation is significant and would shed light on the behavior of the UFS prototypes.

The skill of the SubX models and UFS prototypes in regime shifts (i.e., being on the correct side of the breakpoint) does seem to be tied to their skill in predicting extreme heat. The better model performance over the Mississippi basin and Northwest CONUS implies a connection between soil moisture and extreme heat skill that needs to be improved upon in model physics. Similarly, the subpar performance of the models, particularly the UFS prototypes in southwest CONUS also seems to support this assessment. The UFS models seem to do better overall in representing breakpoints and soil moisture estimates than the SubX models, while the ESRL model does better in capturing heat extremes but not as well in determining the right soil moisture–maximum temperature relationship. It remains to make a direct connection between soil moisture and EHD forecast skill.

d. Relationship between breakpoint and extreme heat day skill

To determine whether proper breakpoint representation has any effect on the skill of EHD prediction, we scrutinize the relationship between soil moisture estimates and EHD forecasts in the models. Composites have been made to quantify the influence of breakpoint estimation skill on EHD prediction. The composites obtain the skill of the EHD forecasts when the skill of the soil moisture breakpoint representation is divided into four categories: hits, misses, false alarms, and correct negatives. The composites are calculated at each grid point across the United States and spatially averaged at each lead day.

Figure 12 and Fig. S8 show the influence of skillful soil moisture relative to its breakpoint on extreme heat day prediction. The blue lines denote the model skill of EHD prediction for days when the model correctly had soil moisture drier than the breakpoint (hits), and the false positive (orange) lines are for when a model inaccurately indicated soil moisture was below the breakpoint. In both instances, a model is simulating soil moisture to be drier than the breakpoint. On the other hand, the green lines denote the model skill of EHD forecasts when the model incorrectly simulates soil moisture to be above the breakpoint (misses), while the correct negatives (red) signify the model EHD forecast skill when it accurately represents soil moisture wetter than the breakpoint. In these two cases, the soil moisture in the model is above the breakpoint threshold. To determine whether the lines are significantly different, uncertainty analysis is carried out by calculating the standard error of each forecast applying a bootstrap sample-with-replacement method to each model's set of forecasts, repeating the process 50 times to determine confidence levels. The error bars indicate one standard error of each forecast; overlap between a pair of models suggests skill for that lead is not significantly different.

The results for TSS (Fig. 12) show that the models unanimously appear to perform better at extreme heat prediction when they are dry (i.e., in the hypersensitive regime), suggesting a key role for soil moisture. Extreme heat forecast skill when there are

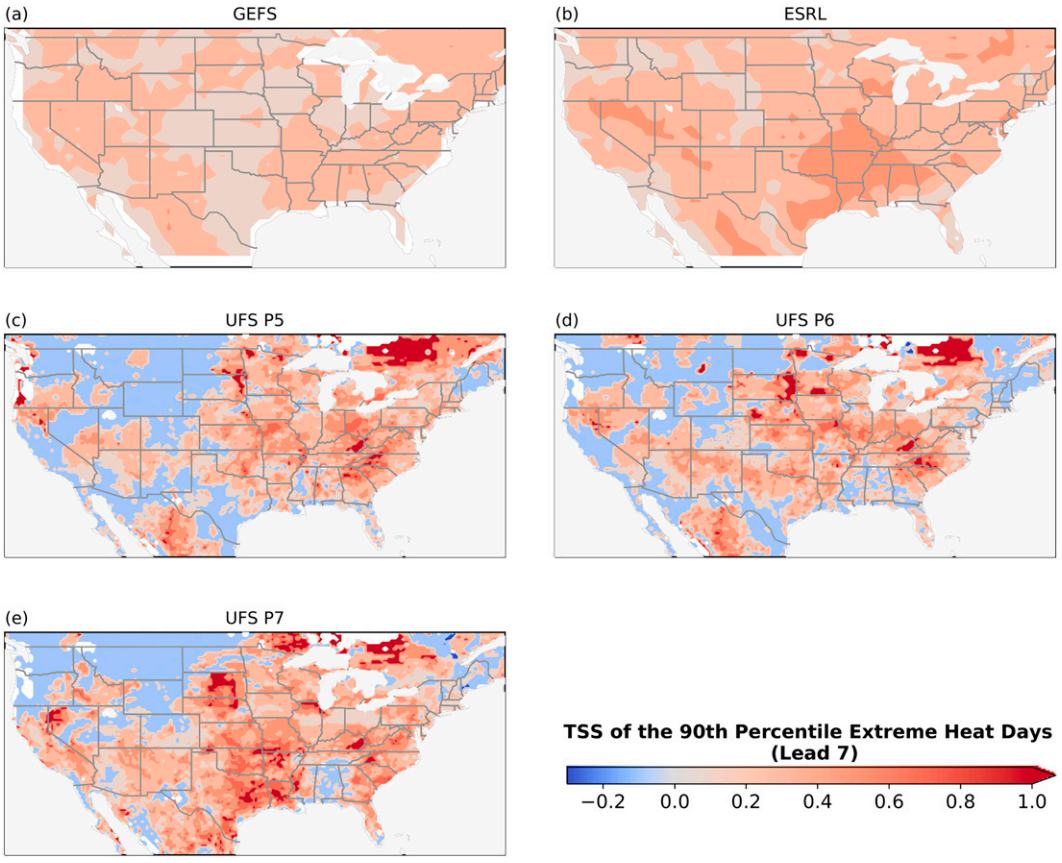

FIG. 10. Maps of the TSS of extreme heat days over CONUS for lead day 7. Blue color indicates areas with skill worse than random guess.

either hits or false positives outperforms the forecasts when there are misses or correct negative skill for soil moisture, for up to a week of forecast in GEFS and ESRL and up to 2 weeks for the UFS prototypes. In other words, the models are better at predicting extreme heat when they are in the hypersensitive regime, regardless of whether that state concurs with the reanalysis. The uncertainties indicate that the separation is usually significant.

The GEFS model (Fig. 12a) performs the worst when it misses the dry soil moisture state, but all other models oddly perform worst for the correct negatives. It is possible that in these models, although the soil moisture is not dry enough to exceed the breakpoint, they miss some other factor which influences the skill of EHD prediction. The high performance of the "hit" composite is heartening, in that it suggests an improvement in model soil moisture initialization to better capture genuinely dry days could further improve EHD skill. However, the elevated skill for false positives suggests there is more to the problem than land surface initialization.

The comparable ETS scores (Fig. S8) behave similarly to TSS scores. An exception is the ESRL model's ETS scores (Fig. S8b), which are more skillful both when it correctly represents dryness relative to breakpoints and when it does not (i.e., when ERA5 indicates a hypersensitive regime). Otherwise, all models show more lead times with the significantly highest skill for soil moisture hits than for any other category.

4. Discussion and conclusions

In this study, the skill of several subseasonal forecast models has been assessed to better understand how their representation of land–atmosphere interactions and soil moisture initialization influence their extreme heat prediction capabilities over the conterminous United States. The models have been compared to ERA5, which is used as validation. They are assessed on their representation of thresholds of soil moisture below which the near-surface atmospheric temperature becomes more sensitive to drying, which we call breakpoints. The models' ability to predict the 90th percentile of maximum temperature defined herein as extreme heat days (EHD) is also analyzed. Composite studies have been employed to parse possible connections between accurate characterization of breakpoints and EHD predictive skill.

Breakpoints occur when drying soils lead to the shutting down of evaporation and its associated cooling effect, and the transfer of net radiation from land to atmosphere occurs mainly through sensible heat flux (Benson and Dirmeyer 2021). The decline of soil moisture across the breakpoint indicates a shift in the sensitivity of land–atmosphere coupling in which the atmosphere moves from a moderately sensitive regime, where temperature gradually increases as soil dries, to a hypersensitive regime where this response intensifies (the slope of piecewise linear regression increases significantly). Most of the models

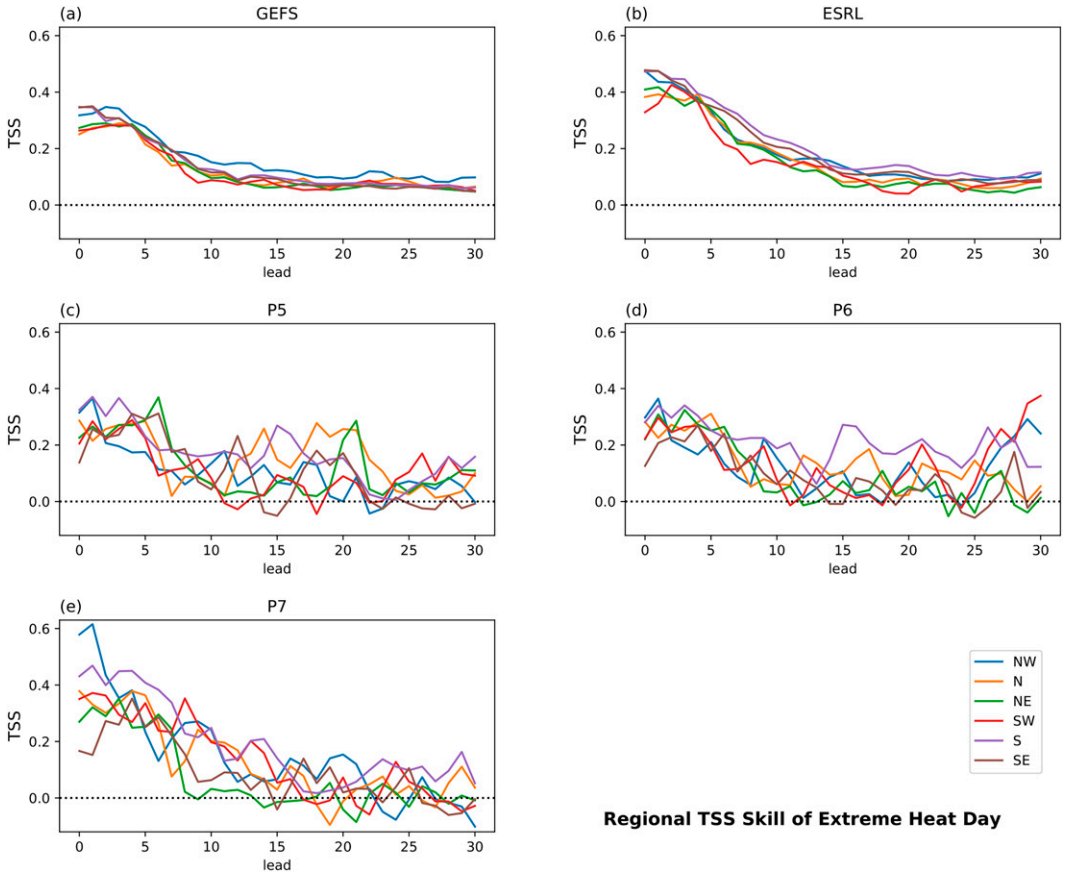

FIG. 11. Regional dependency of the TSS for extreme heat days across forecast lead days.

involved in this study appear to replicate the continental-scale pattern of breakpoint distribution across the United States: low soil moisture values at the breakpoint in the Southwest and higher values in the Northeast. The breakpoint maps mirror the maps of the summer soil moisture climatology (Fig. 2) and although this result is seemingly obvious, it emphasizes the importance of the role of proper land surface representativeness in forecast models and their potential to improve or hinder the predictability of extreme heat as well as associated soil moisture driven drought. Locations pass in and out of the hypersensitive regime depending on the moisture content in the soil, but generally the forecast models effectively capture the tendency of western regions to spend more days of the JJA summer in the hypersensitive regime than regions in the east (Fig. 3), as seen in the reanalysis. A major driver of the model performance appears to be soil moisture initialization.

The exact values of volumetric soil moisture for the breakpoints of the models were not compared. Instead, the models were assessed to determine whether their soil moisture content was below the breakpoint on the same dates as ERA5 at each location. This approach was taken because the relative position of soil moisture to the breakpoint is relevant for land–atmosphere interactions. Ultimately, LSM soil moisture is a derived parameter that describes the response of the model to forcings and parameters, with fluxes being more important quantities (Koster and Milly 1997). Hence, directly comparing the breakpoint estimates of the models to the reanalysis could be misleading. Therefore, the soil moisture content of each model is validated relative to its own breakpoints in order to quantify the relative responsiveness of each model to the variations of its soil moisture.

The number of days either side of the breakpoint can vary from location to location and model to model, and rarely matches the climatological proportions from reanalysis. Such biases affect skill metrics. The best attainable skill scores when the number of days on either side of the breakpoint does not match the ERA5 validation data have been determined (Fig. 4), and they strongly affect interpretation of the results. TSS (Fig. 5) and ETS (Fig. S4) scores are only fair at the time of forecast initialization, and they decline as the forecast lead time increases. The models' skill is much higher when normalized by their best attainable skill, therefore accounting for the models' differences in representation of soil moisture climatology relative to the breakpoint. Regionally, the models generally performed best in the U.S. Northwest and worst in the Southwest (Fig. 7).

Turning to forecasts of extreme heat, models mostly represent the mean state of maximum temperature properly at initialization when compared to ERA5 (Fig. 8). While there are systemic temperature biases in the models (Fig. S5), they are

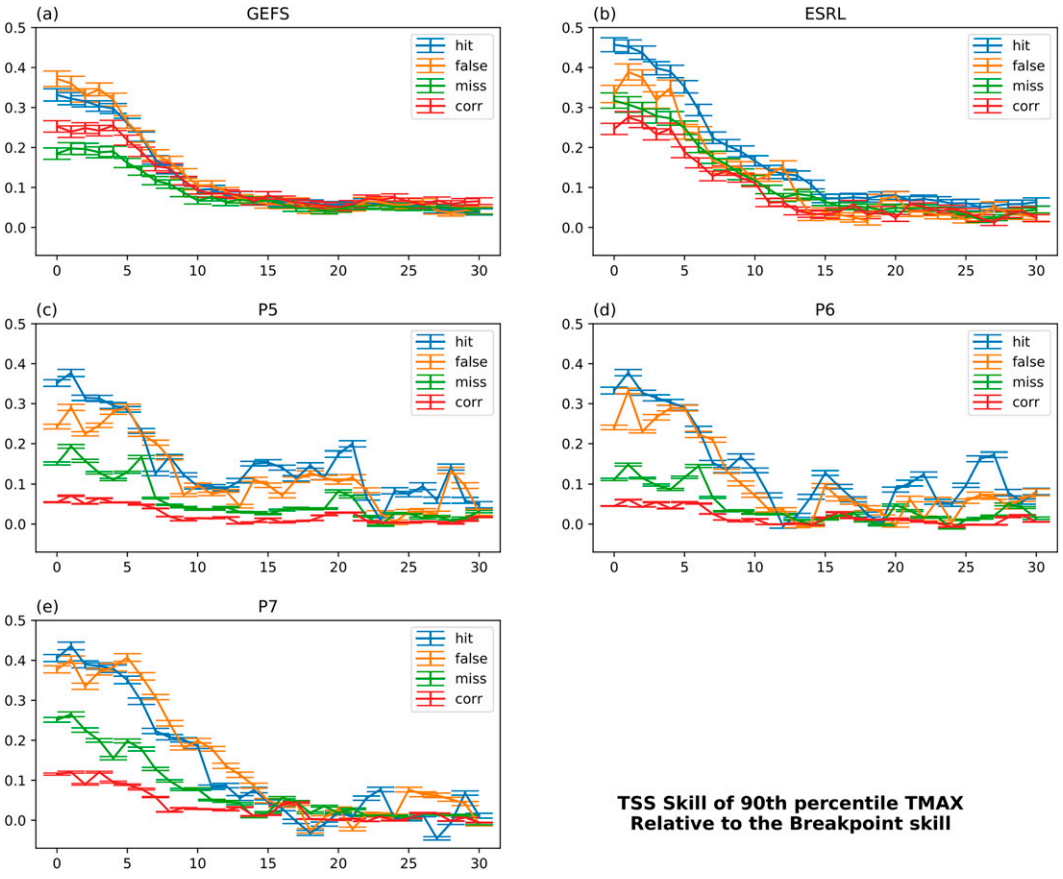

FIG. 12. TSS for extreme heat day skill spatially averaged over the study region at each lead day for (a)–(c) SubX models and (d)–(f) UFS prototypes with forecasts grouped into composites based on whether the soil moisture on that day, relative to its breakpoint, scored a hit (blue), a false positive (orange), a miss (green), or a correct negative (red).

not the sole cause of the lack of predictive skill. The models' performance in EHD forecasts starts out skillful but decays more rapidly than their soil moisture breakpoint skill. The UFS prototypes tend toward good skill in the eastern part of the United States and poorer skill in the northern Rockies and Northwest (Fig. 10). The map of extreme heat days reveals that the prototypes fail to capture extreme heat in the northern Great Plains but are skillful over the Mississippi basin. They struggle to replicate the magnitude of extreme heat days in regions that have complex topography like the Rocky Mountains.

There is noticeable improvement in skill across the evolution of the prototypes from UFS P5 to P7. However, inland areas in the west of the study region that perform poorly in the representation of dry extremes are also poor for extreme heat forecasts. A more accurate representation of these regime shifts could lead to higher accuracy in extreme heat prediction. It is plausible that the problems with soil moisture initialization also play a role in the rapid decay of EHD skill in the models. Another potential reason could be misplaced breakpoint values in forecast models due to problems with model physics. Such problems could affect not only initialization on the correct side of the breakpoint but also models' ability to effectively represent the difference in sensitivity between the sensitive to the hypersensitive regimes.

The relationship between soil dryness relative to breakpoints and EHD forecast skill was quantified using a composite analysis. The EHD forecast skill was categorized into composite fields for when the skill of soil dryness relative to the breakpoint was a hit, miss, false alarm, and correct negative, and calculated for each lead day. All models performed better at predicting extreme heat in the first five to eight days when they properly initialized soil moisture relative to its breakpoint, and most models displayed significantly better performance when they were drier, regardless of the observed state (Fig. 12). Correctly predicting soil moisture on the dry side of the breakpoint plays a significant role in improving EHD forecast skill, yet simply having a dry model soil is seen to improve EHD forecast skill even up to the second week (Fig. 12 and Fig. S8). The evident connection between dry soil and EHD skill is encouraging as it illustrates the potential for better predictability of extreme heat amid numerous efforts aimed at the improvement of soil moisture observations and accurate data collection. More analysis or potentially model

sensitivity studies would be necessary to disentangle the contributing factors, whether improvement of soil moisture initialization relative to the breakpoint, model breakpoint representation, or land–atmosphere coupling would contribute most to advancement in EHD predictability.

Overall, the GEFS and ESRL models seem to do relatively better overall at extreme heat prediction than estimating soil moisture dryness relative to the breakpoint. The UFS P7 on the other hand, seems to be an improvement over earlier prototypes particularly in the depiction of extreme heat. This advancement could be ascribed to the changes in the land surface model and boundary layer parameterizations in the most recent of the prototypes used for this analysis.

To summarize, the evaluation of the S2S models and UFS prototypes reiterates the need for the focus on soil moisture initialization and boundary layer physics in regard to the land surface. Through breakpoint analysis, the results in this study provide another methodology in which land surface hydrology and land–atmosphere interactions can be examined in forecast models using statistical methods to elucidate the soil moisture–temperature between the land and the atmosphere in these models. From the standpoint of model validation and development, the soil moisture breakpoint presents several challenges. The breakpoint itself is largely determined by soil and vegetation properties that control wilting point, within the domain of LSMs. But the position of any given day above or below the breakpoint is determined by both land and atmosphere, and the ability of their corresponding models to reach the proper balance in the water cycle. The impact of breakpoint transitions on extreme heat sensitivity brings in the energy cycle of both land and atmosphere. In a changing climate, regions are more likely to shift into the hypersensitive regime in the warmer months, leading to the possibility of more severe heatwaves (Seneviratne et al. 2013; Berg and Sheffield 2018; Schoof et al. 2019; Hirsch et al. 2021).

Beyond improved model physics, realistic initialization of soil moisture has been shown to improve atmospheric predictability over the boreal summer (Guo et al. 2011; Koster et al. 2011; Dirmeyer et al. 2018). Koster et al. (2009a) reasoned against the direct transfer of soil moisture values between models because it leads to model-inconsistent initializations and consequently loss of skill, particularly in seasonal and longer time scale forecasts of meteorological quantities. That study argued for the initialization of soil moisture using adjusted model statistics, based on anomalies scaled by model means and standard deviations. Results from this study provide justification for the next step: consideration of the physics of the critical points of soil moisture that demark different coupling regimes between land and atmosphere (Hsu and Dirmeyer 2021). Ensuring that soil moisture is initialized correctly relative to the breakpoint separating the sensitive from the hypersensitive regimes, as well as the higher critical soil moisture value that separates water-limited from energy-limited regimes (not examined in this study) can help improve prediction skill in the S2S models.

One caveat regarding the methodology used in this study is that piecewise regression could be convolving other connections in the soil moisture–temperature relationship not central to this analysis, or that are unexplained by the presumed physical processes linking soil dryness to heat. One example of this is over the Northeast U.S. region where estimated breakpoint values are more likely indicators of changes in evapotranspiration and the predisposition of that region to precipitation changes (Koster et al. 2004). If soil moisture in a humid location never drops below the local breakpoint value, it will not be present in the data and the piecewise regression will seize on some other artifact in the data. It is also important to remember that these models were validated against the ERA5 and, while reliable, it is not an infallible representation of Earth system processes, having some level of model dependence of its own.

It is evident through this study and others that drier soil conditions contribute to extreme heat through the changes in the partitioning of surface fluxes (Miralles et al. 2012; Wulff and Domeisen 2019). This connection between soil moisture and temperature extremes provides a potential source of improved predictability of hot extremes especially at subseasonal to seasonal (S2S) time scales where informed decisions for policy makers are pertinent. S2S forecast models appear to have room for improvement as this relationship between soil moisture and temperature extremes is currently not well reproduced. Finally, while this study was carried out over CONUS, the methodology can be applied elsewhere characterizing the nature of land–atmosphere interactions in various regional climates and diagnosing model strengths and limitations across those regions.

Acknowledgments. This research was supported by National Oceanographic and Atmospheric Administration Grant NA16OAR4310095 and National Aeronautics and Space Administration Grant 80NSSC21K1801. We acknowledge the agencies that support the SubX system, and we thank the climate modeling groups (Environment and Climate Change Canada, NASA, NOAA/NCEP, NRL, and the University of Miami) for producing and making available their model output. NOAA/MAPP, ONR, NASA, NOAA/NWS jointly provided coordinating support and led development of the SubX system.

Data availability statement. The ERA5 data are openly available at ECMWF. Further documentation of the data can be found here: https://confluence.ecmwf.int/display/CKB/ERA5. All SubX data used in this research are made available through the International Research Institute for Climate and Society (IRI) Data Library at Columbia University http://iridl.ldeo.columbia.edu/SOURCES/.Models/.SubX/. The NOAA Unified Forecast System Subseasonal to Seasonal Prototypes was accessed from https://registry.opendata.aws/noaa-ufs-s2s.

REFERENCES

Barlage, M., M. Tewari, F. Chen, G. Miguez-Macho, Z. L. Yang, and G. Y. Niu, 2015: The effect of groundwater interaction in North American regional climate simulations with WRF/Noah-MP. *Climatic Change*, **129**, 485–498, https://doi.org/10.1007/s10584-014-1308-8.

——, F. Chen, R. Rasmussen, Z. Zhang, and G. Miguez-Macho, 2021: The importance of scale-dependent groundwater

processes in land–atmosphere interactions over the central United States. *Geophys. Res. Lett.*, **48**, e2020GL092171, https://doi.org/10.1029/2020GL092171.

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

Berg, A., and J. Sheffield, 2018: Soil moisture–evapotranspiration coupling in CMIP5 models: Relationship with simulated climate and projections. *J. Climate*, **31**, 4865–4878, https://doi.org/10.1175/JCLI-D-17-0757.1.

Denissen, J. M. C., A. J. Teuling, M. Reichstein, and R. Orth, 2020: Critical soil moisture derived from satellite observations over Europe. *J. Geophys. Res. Atmos.*, **125**, e2019JD031672, https://doi.org/10.1029/2019JD031672.

Dirmeyer, P. A., 2011: The terrestrial segment of soil moisture–climate coupling. *Geophys. Res. Lett.*, **38**, L16702, https://doi.org/10.1029/2011GL048268.

——, C. A. Schlosser, and K. L. Brubaker, 2009: Precipitation, recycling and land memory: An integrated analysis. *J. Hydrometeor.*, **10**, 278–288, https://doi.org/10.1175/2008JHM1016.1.

——, S. Halder, and R. Bombardi, 2018: On the harvest of predictability from land states in a global forecast model. *J. Geophys. Res. Atmos.*, **123**, 13 111–13 127, https://doi.org/10.1029/2018JD029103.

——, G. Balsamo, E. M. Blyth, R. Morrison, and H. M. Cooper, 2021: Land–atmosphere interactions exacerbated the drought and heatwave over northern Europe during summer 2018. *AGU Adv.*, **2**, e2020AV000283, https://doi.org/10.1029/2020AV000283.

Ek, M. B., and Coauthors, 2003: Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J. Geophys. Res. Atmos.*, **108**, 8851, https://doi.org/10.1029/2002JD003296.

Ford, T. W., and S. M. Quiring, 2014: In situ soil moisture coupled with extreme temperatures: A study based on the Oklahoma Mesonet. *Geophys. Res. Lett.*, **41**, 4727–4734, https://doi.org/10.1002/2014GL060949.

Gevaert, A. I., D. G. Miralles, R. A. M. de Jeu, J. Schellekens, and A. J. Dolman, 2018: Soil moisture–temperature coupling in a set of land surface models. *J. Geophys. Res. Atmos.*, **123**, 1481–1498, https://doi.org/10.1002/2017JD027346.

Guo, Z., P. A. Dirmeyer, and T. DelSole, 2011: Land surface impacts on subseasonal and seasonal predictability. *Geophys. Res. Lett.*, **38**, L24812, https://doi.org/10.1029/2011GL049945.

Haghighi, E., D. J. Short Gianotti, R. Akbar, G. D. Salvucci, and D. Entekhabi, 2018: Soil and atmospheric controls on the land surface energy balance: A generalized framework for distinguishing moisture-limited and energy-limited evaporation regimes. *Water Resour. Res.*, **54**, 1831–1851, https://doi.org/10.1002/2017WR021729.

Herold, N., J. Kala, and L. V. Alexander, 2016: The influence of soil moisture deficits on Australian heatwaves. *Environ. Res. Lett.*, **11**, 064003, https://doi.org/10.1088/1748-9326/11/6/064003.

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

Hirsch, A. L., N. N. Ridder, S. E. Perkins-Kirkpatrick, and A. Ukkola, 2021: CMIP6 multimodel evaluation of present-day heatwave attributes. *Geophys. Res. Lett.*, **48**, e2021GL095161, https://doi.org/10.1029/2021GL095161.

Hsu, H., and P. A. Dirmeyer, 2021: Nonlinearity and multivariate dependencies in land–atmosphere coupling. *Water Resour. Res.*, **57**, e2020WR028179, https://doi.org/10.1029/2020WR028179.

Koster, R. D., and P. C. D. Milly, 1997: The interplay between transpiration and runoff formulations in land surface schemes used with atmospheric models. *J. Climate*, **10**, 1578–1591, https://doi.org/10.1175/1520-0442(1997)010<1578:TIBTAR>2.0.CO;2.

——, and Coauthors, 2004: Regions of strong coupling between soil moisture and precipitation. *Science*, **305**, 1138–1140, https://doi.org/10.1126/science.1100217.

——, Z. Guo, P. A. Dirmeyer, R. Yang, K. Mitchell, and M. J. Puma, 2009a: On the nature of soil moisture in land surface models. *J. Climate*, **22**, 4322–4335, https://doi.org/10.1175/2009JCLI2832.1.

——, S. D. Schubert, and M. J. Suarez, 2009b: Analyzing the concurrence of meteorological droughts and warm periods, with implications for the determination of evaporative regime. *J. Climate*, **22**, 3331–3341, https://doi.org/10.1175/2008JCLI2718.1.

——, and Coauthors, 2011: The second phase of the Global Land–Atmosphere Coupling Experiment: Soil moisture contributions to subseasonal forecast skill. *J. Hydrometeor.*, **12**, 805–822, https://doi.org/10.1175/2011JHM1365.1.

Lhotka, O., J. Kyselý, and A. Farda, 2018: Climate change scenarios of heat waves in Central Europe and their uncertainties. *Theor. Appl. Climatol.*, **131**, 1043–1054, https://doi.org/10.1007/s00704-016-2031-3.

Lorenz, R., E. B. Jaeger, and S. I. Seneviratne, 2010: Persistence of heat waves and its link to soil moisture memory. *Geophys. Res. Lett.*, **37**, L09703, https://doi.org/10.1029/2010GL042764.

Miralles, D. G., M. J. van den Berg, A. J. Teuling, and R. A. M. de Jeu, 2012: Soil moisture-temperature coupling: A multiscale observational analysis. *Geophys. Res. Lett.*, **39**, L21707, https://doi.org/10.1029/2012GL053703.

Muggeo, V. M., 2008: Segmented: An R package to fit regression models with broken-line relationships. *R News*, **8** (1), 20–25.

Muñoz-Sabater, J., H. Lawrence, C. Albergel, P. Rosnay, L. Isaksen, S. Mecklenburg, Y. Kerr, and M. Drusch, 2019: Assimilation of SMOS brightness temperatures in the ECMWF Integrated Forecasting System. *Quart. J. Roy. Meteor. Soc.*, **145**, 2524–2548, https://doi.org/10.1002/qj.3577.

Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. *Bull. Amer. Meteor. Soc.*, **100**, 2043–2060, https://doi.org/10.1175/BAMS-D-18-0270.1.

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

Peterson, T. C., and Coauthors, 2013: Monitoring and understanding changes in heat waves, cold waves, floods, and droughts in the United States: State of knowledge. *Bull. Amer. Meteor. Soc.*, **94**, 821–834, https://doi.org/10.1175/BAMS-D-12-00066.1.

Qiu, J., W. T. Crow, and G. S. Nearing, 2016: The impact of vertical measurement depth on the information content of soil moisture for latent heat flux estimation. *J. Hydrometeor.*, **17**, 2419–2430, https://doi.org/10.1175/JHM-D-16-0044.1.

Salamanca, F., Y. Zhang, M. Barlage, F. Chen, A. Mahalov, and S. Miao, 2018: Evaluation of the WRF-urban modeling system coupled to Noah and Noah-MP land surface models over a semiarid urban environment. *J. Geophys. Res. Atmos.*, **123**, 2387–2408, https://doi.org/10.1002/2018JD028377.

Schoof, J. T., S. C. Pryor, and T. W. Ford, 2019: Projected changes in United States regional extreme heat days derived from bivariate

quantile mapping of CMIP5 simulations. *J. Geophys. Res. Atmos.*, **124**, 5214–5232, https://doi.org/10.1029/2018JD029599.

Schwingshackl, C., M. Hirschi, and S. I. Seneviratne, 2017: Quantifying spatiotemporal variations of soil moisture control on surface energy balance and near-surface air temperature. *J. Climate*, **30**, 7105–7124, https://doi.org/10.1175/JCLI-D-16-0727.1.

Sehgal, V., N. Gaur, and B. P. Mohanty, 2021: Global surface soil moisture drydown patterns. *Water Resour. Res.*, **57**, e2020WR027588, https://doi.org/10.1029/2020WR027588.

Seneviratne, S. I., T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling, 2010: Investigating soil moisture–climate interactions in a changing climate: A review. *Earth-Sci. Rev.*, **99**, 125–161, https://doi.org/10.1016/j.earscirev.2010.02.004.

——, and Coauthors, 2013: Impact of soil moisture-climate feedbacks on CMIP5 projections: First results from the GLACE-CMIP5 experiment. *Geophys. Res. Lett.*, **40**, 5212–5217, https://doi.org/10.1002/grl.50956.

Sillmann, J., and Coauthors, 2017: Understanding, modeling and predicting weather and climate extremes: Challenges and opportunities. *Wea. Climate Extremes*, **18**, 65–74, https://doi.org/10.1016/j.wace.2017.10.003.

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: *Survey of Common Verification Methods in Meteorology*. WMO, 114 pp.

Stephenson, D. B., B. Casati, C. A. T. Ferro, and C. A. Wilson, 2008: The extreme dependency score: A non-vanishing measure for forecasts of rare events. *Meteor. Appl.*, **15**, 41–50, https://doi.org/10.1002/met.53.

Tartaglione, N., 2010: Relationship between precipitation forecast errors and skill scores of dichotomous forecasts. *Wea. Forecasting*, **25**, 355–365, https://doi.org/10.1175/2009WAF2222211.1.

Vautard, R., and Coauthors, 2013: The simulation of European heat waves from an ensemble of regional climate models within the EURO-CORDEX project. *Climate Dyn.*, **41**, 2555–2575, https://doi.org/10.1007/s00382-013-1714-z.

Woodcock, F., 1976: The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Wea. Rev.*, **104**, 1209–1214, https://doi.org/10.1175/1520-0493(1976)104<1209:TEOYFF>2.0.CO;2.

Wu, J., and P. A. Dirmeyer, 2020: Drought demise attribution over CONUS. *J. Geophys. Res. Atmos.*, **125**, e2019JD031255, https://doi.org/10.1029/2019JD031255.

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

Xue, Y., and Coauthors, 2021: Development of a coupled subseasonal-to-seasonal prediction model using community-based unified forecast system for NCEP operations. *2021 EGU General Assembly*, Online, European Geosciences Union, Abstract EGU21-5722, https://doi.org/10.5194/egusphere-egu21-5722.