Subseasonal Forecast Skill of Evaporative Demand, Soil Moisture, and Flash Drought Onset from Two Dynamic Models over the Contiguous United States

KYLE LESINGER,[a] DI TIAN,[a] AND HAILAN WANG[b]

[a] *Department of Crop, Soil, and Environmental Sciences, Auburn University, Auburn, Alabama*
[b] *Climate Prediction Center, NOAA/NWS/NCEP, College Park, Maryland*

(Manuscript received 28 July 2023, in final form 19 March 2024, accepted 27 March 2024)

ABSTRACT: Flash droughts are rapidly developing subseasonal climate extreme events that are manifested as suddenly decreased soil moisture, driven by increased evaporative demand and/or sustained precipitation deficits. Over each climate region in the contiguous United States (CONUS), we evaluated the forecast skill of weekly root-zone soil moisture (RZSM), evaporative demand (ET_o), and relevant flash drought (FD) indices derived from two dynamic models [Goddard Earth Observing System model V2p1 (GEOS-V2p1) and Global Ensemble Forecast System version 12 (GEFSv12)] in the Subseasonal Experiment (SubX) project between years 2000 and 2019 against three reference datasets: Modern-Era Retrospective Analysis for Research and Applications version 2 (MERRA-2), North American Land Data Assimilation System, phase 2 (NLDAS-2), and GEFSv12 reanalysis. The ET_o and its forcing variables at lead week 1 have moderate-to-high anomaly correlation coefficient (ACC) skill (~0.70–0.95) except downwelling shortwave radiation, and by weeks 3–4, predictability was low for all forcing variables (ACC < 0.5). RZSM (0–100 cm) for model GEFSv12 showed high skill at lead week 1 (~0.7–0.85 ACC) in the High Plains, West, Midwest, and South CONUS regions when evaluated against GEFSv12 reanalysis but lower skill against MERRA-2 and NLDAS-2 and ACC skill are still close to 0.5 for lead weeks 3–4, better than ET_o forecasts. GEFSv12 analysis has not been evaluated against in situ observations and has substantial RZSM anomaly differences when compared to NLDAS-2, and our analysis identified GEFSv12 reforecast prediction limit, which can maximally achieve ACC ~0.6 for RZSM forecasts between lead weeks 3 and 4. Analysis of major FD events reveals that GEFSv12 reforecast inconsistently captured the correct location of atmospheric and RZSM anomalies contributing to FD onset, suggesting the needs for improving the dynamic models' assimilation and initialization procedures to improve subseasonal FD predictability.

SIGNIFICANCE STATEMENT: Flash droughts are rapidly developing climate extremes which reduce soil moisture through enhanced evaporative demand and precipitation deficits, and these events can have large impacts on the ecosystem and crop health. We evaluated the subseasonal forecast skill of soil moisture and evaporative demand against three reanalysis datasets and found that evaporative demand skill was similar between forecasts and reanalyses while soil moisture skill is dependent on the reference dataset. Skill of evaporative demand decreases rapidly after week 1, while soil moisture skill declines more slowly after week 1. Case studies for the 2012, 2017, and 2019 United States flash droughts identified that forecasts could capture rapid decreases in soil moisture in some regions but not consistently, implying that long-lead forecasts still need improvements before being used in early warning systems. Improvements in flash drought predictability at longer lead times will require less biased initial conditions, better model parameterizations, and improved representations of large-scale teleconnections.

KEYWORDS: Drought; Evapotranspiration; Soil moisture; Forecast verification/skill; Model comparison; Model evaluation/performance

1. Introduction

Flash droughts (FDs) are rapidly developing subseasonal climate phenomena that reduce water availability and can pose long-lasting impacts on agriculture, ecosystems, and other natural and human systems globally. Widespread FD can cost billions of dollars in economic or ecological losses through crop and livestock losses, wildfire propagation, depletion of reservoirs, reduced streamflow, or increases in heat-related morbidity; therefore, it is important to accurately predict their occurrence to mitigate these harmful effects (NCEI 2020; McEvoy et al. 2020; Jin et al. 2019; Christian et al. 2020; Pendergrass et al. 2020; Banerjee et al. 2013; Gerken et al. 2018; Abadi et al. 2022). FD can be driven, from local or remote sources, by conditions which create or sustain high temperatures and high vapor pressure deficit and/or low initial soil moisture, which, in conjunction with below-normal precipitation, can promote land–atmosphere feedbacks causing rapid changes in evaporative demand (ET_o) and root-zone soil moisture (RZSM) leading to evaporative and water stress, respectively (Gerken et al. 2018; Schubert et al. 2021; Deangelis et al. 2020; Pendergrass et al. 2020; Otkin et al. 2022). Once stress is induced, plant growth and productivity are inhibited or damaged, which undermines ecosystems and agriculture, increases wildfire risks, and decreases air quality with substantial economic

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JHM-D-23-0124.s1.

Corresponding author: Di Tian, tiandi@auburn.edu

DOI: 10.1175/JHM-D-23-0124.1

losses incurred (Banerjee et al. 2013; NCEI 2020; Kunkel et al. 2018; McEvoy et al. 2020).

Traditionally, droughts are considered long-lasting and slowly developing events creating water deficits over land (Mishra and Singh 2010; Van Loon et al. 2016; Hao et al. 2018; Wilhite et al. 1985), whereas FDs are a newly recognized drought subclass, which intensify more rapidly compared to traditional droughts. FD development occurs at the Subseasonal to Seasonal Prediction project (S2S) time scale (~2–8 weeks), particularly during the warm season (March–November) in which crop growth is prevalent and temperatures are higher (Pendergrass et al. 2020). Numerous indices have been utilized to classify FD onset, and no single index generalizes all events across the United States (Osman et al. 2021, 2022). Commonly used FD indices include the soil moisture percentile drop (SMPD) in which FD onset is defined when a rapid decrease in RZSM occurs within 3 weeks or less (Osman et al. 2021; Ford and Labosier 2017) and heatwave/precipitation deficit FD classification by examining RZSM with concurrent changes in temperature, evapotranspiration (ET), and precipitation (Mo and Lettenmaier 2015, 2016), and other researchers have identified rapid changes with respect to standardized ratios between ET and ET_o with the evaporative stress index (ESI) or standardized evaporative stress ratio (SESR) (Christian et al. 2019; Otkin et al. 2014; Edris et al. 2023; Otkin et al. 2018). The evaporative demand drought index (EDDI), an early warning FD tool, details anomalous ET_o over a user-defined time window (Hobbins et al. 2016; McEvoy et al. 2016), and finally, satellites have revealed the total water content in vegetation or changes in 5-cm soil moisture, which can provide a near-real-time FD monitoring (Christian et al. 2022; Sehgal et al. 2021). While uncertainty exists regarding an optimal index that can be used for all regions and seasons, these indices exhibit strong signals for identifying high-impacting contiguous United States (CONUS) FDs as each index captures rapid changes occurring over land, which impact water availability and ecosystem stress (Osman et al. 2022).

FD onset can be induced by persistent atmospheric patterns associated with quasi-stationary Rossby waves producing alternating high and low pressure systems or by more isolated high pressure systems which reduce cloud cover and increase radiative heating of the surface; increase descending dry air, which compresses and warms air temperatures; and increase vapor pressure deficits, and these combined factors can increase ET_o and decrease RZSM through land–atmosphere coupling (Ford and Labosier 2017; Schubert et al. 2021; Deangelis et al. 2020; PaiMazumder and Done 2016; Hoerling et al. 2014; Jong et al. 2022). ET_o, also known as potential evapotranspiration, describes the atmospheric capacity to evaporate water from Earth's surface, and under conditions of high ET_o, RZSM can be rapidly decreased through enhanced ET, which increases water stress. Hobbins (2016) identified that high ET_o can be used as a proxy for low ET because of their complementary relationship under water-limited conditions, and J. Zhang et al. (2021) identified that high ET_o can induce plant stress despite concurrent high RZSM due to the coregulation of stomata by both RZSM and ET_o. The central United States has been identified as an area where there is a strong coupling between RZSM and precipitation, and ET_o can mediate these feedbacks by promoting aridity when RZSM becomes low (Koster et al. 2004a; Seneviratne et al. 2010). The relationship between RZSM, ET, ET_o, and precipitation feedbacks is complex and varies by topography, season, anomaly strength, and initial conditions, as well as being influenced by large-scale circulation patterns (Seneviratne et al. 2010; Guo and Dirmeyer 2013). For example, local negative precipitation anomalies and higher ET_o can occur if either water evaporated from the land surface is advected out of the system and enhanced by large-scale circulations which sustain the vapor pressure deficit or if RZSM matric potential becomes too negative and plants cannot exert a strong enough force to capture soil water which inhibits ET and increases the vapor pressure deficit (Wang et al. 2007; Brubaker et al. 1993; Gavande and Taylor 1967; Seneviratne et al. 2010). Under energy-limited conditions, an increase in ET_o causes ET to increase and RZSM to decrease, and once water becomes unavailable, the shift to water-limited conditions causes ET to decrease and ET_o to increase through enhanced surface heat fluxes. Under dry RZSM conditions, radiative energy is largely partitioned into sensible heat which heats air parcels and increases atmospheric water holding capacity, thereby increasing the vapor pressure deficit and ET_o (Zhou et al. 2019; Berg et al. 2016), but under wet RZSM conditions, evaporated water can decrease ET_o due to atmospheric moisture convergence, which forms clouds and increases precipitation likelihood (Findell et al. 2011). Studies have identified that antecedent RZSM conditions are lower and ET_o conditions are higher prior to FD onset; therefore, ET_o and RZSM are suitable indicators for FD prediction due to land–atmosphere coupling feedbacks, which promote rapid drying of soils (Christian et al. 2019; Osman et al. 2022).

Subseasonal flash drought prediction is challenging due to the weather–climate prediction gap (Mariotti et al. 2018). Weather forecasting is generally driven by initial atmospheric conditions (Vitart et al. 2017), whereas climate (seasonal) forecasting is strongly influenced by boundary conditions, such as sea surface temperatures or soil moisture (Koster et al. 2004b). At the subseasonal scale, the effects of forecast initial conditions are diminished and not long enough to be strongly influenced by oceanic boundary conditions. It has been identified that subseasonal forecast skill can be increased when models are initialized with dry or wet soil moisture anomalies and research agrees on the importance of land surface initialization and its effects on energy and water fluxes (Koster et al. 2004b; Deangelis et al. 2020; Koster et al. 2010; Dirmeyer et al. 2018; Liang and Yuan 2021; Koster et al. 2020). Additional challenges impacting subseasonal prediction skills include the biases in initial conditions combined with unresolved parameterizations of lower resolution models (Mariotti et al. 2018). To address these challenges, two multimodel subseasonal forecast projects have been initiated, including the S2S project and the Subseasonal Experiment (SubX) project, which generated retrospective forecasts (reforecasts) and real-time forecasts encapsulating ensemble forecasts from different coupled numerical models (Pegion et al. 2019; Vitart et al. 2017). These reforecasts have been used to better understand S2S predictability related to the Madden–Julian oscillation (MJO)

(Kim et al. 2019; Du et al. 2023), atmospheric rivers (Cao et al. 2021), precipitation (Li et al. 2021; Park and Kam 2023), temperature (Vitart and Robertson 2018), and cyclone activity (Zheng et al. 2021) with a general emphasis on understanding model representation of complex teleconnections, which drive physical processes and extreme phenomena. SubX evaluation of the 2012 central U.S. FD identified that most models at weeks 3–4 had low skill in predicting temperature and precipitation anomalies due to 1) bias in initialized soil moisture anomaly and 2) poor representation of the quasi-stationary cross-Pacific Rossby wave train that lingered over central United States (Deangelis et al. 2020). Lorenz et al. (2021) evaluated week 3–4 changes in both RZSM and the ESI using the European Centre for Medium-Range Weather Forecasts (ECMWF) model and identified low skill due in part to autocorrelation. Autocorrelation typically increases predictability because of dependence from previous days, but for rapid changes in RZSM or ESI, which are FD indicators, this autocorrelation diminishes prediction of extremes due to the models' relaxation toward climatology at longer leads. More specifically, the relaxation toward climatology implies that the model will more often predict the climatological mean state rather than abrupt deviations associated with FD. Lorenz et al. (2021) utilized ECMWF atmospheric conditions to forecast RZSM changes from soilMERGE reanalysis and identified moderate-to-high correlation (0.47–0.80) during June and August months between 2010 and 2018, but they did not examine soil moisture forecast skill directly derived from coupled subseasonal forecast models, which can help us improve understanding of model deficiencies.

When compared to observations, both reanalysis and forecasts can be biased due to model-specific parameterizations or initialization errors. For example, seasonal ET forecasts can be biased due to precipitation and net radiation biases within atmospheric general circulation models (AGCMs), which impact soil moisture memory and land–atmosphere feedback representations within a single land surface model (LSM) (Mahanama and Koster 2005), and research comparing several LSMs against in situ observations identified RZSM biases that vary by spatial and temporal time scales (Dirmeyer et al. 2016). When multiple LSMs were driven with the same meteorological forcings, large variability was observed between outputs indicating uncertainty of land–atmosphere interactions between models (Dirmeyer et al. 2006; Ashfaqur Rahman et al. 2018). LSM epistemic uncertainty has been noted to be related to model-specific parameterizations for factors such as dynamic demographic vegetation, land-use/land-cover change, urbanization, irrigation, and surface energy fluxes whose processes occur at subgrid scale and impact hydrological processes (Fisher and Koven 2020). Uncertainty has also been identified between reanalysis products, which use different AGCMs, specifically for precipitation and temperature extremes (Angélil et al. 2016). Therefore, an objective evaluation of model forecasts should include a comparison between multiple validation datasets to understand uncertainties associated with imperfect reanalyses.

Although research has explored model skill in predicting week 3–4 temperature, precipitation, evaporative stress, and RZSM anomalies associated with FD (Deangelis et al. 2020; Lorenz et al. 2021; Abatzoglou et al. 2023), the forecast skill of RZSM, ET_o, and FD onset within GMAO-V2p1 and Global Ensemble Forecast System version 12 (GEFSv12) subseasonal coupled dynamic models and across CONUS regions has not been comprehensively evaluated over a long period. DeAngelis et al. (2020) identified the importance of accurate land initialization when predicting the weeks 3–4 temperature, precipitation, and RZSM anomalies for 2012 drought but did not assess model performance in capturing rapid decreases in RZSM which indicate FD onset and their analysis was only for the 2012 central U.S. FD. Accurate prediction of rapid decreases in RZSM over a long time period can aid operational agencies in providing regional early warning alerts and mitigate potential impacts associated with FD. Due to the relationship between ET_o and RZSM during FD intensification (Hobbins et al. 2016), an investigation into the model representation of anomalous ET_o and its primary drivers can also provide additional early warning of FD onset and identify limitations of current forecasts. To improve our understanding of flash drought predictability from state-of-the-art numerical models at the subseasonal time scale, the objectives of this study are to 1) evaluate the forecast skill of RZSM and ET_o between weeks 1 and 4, 2) assess the utility of an RZSM-based index in predicting high-impacting FD events, and 3) identify differences in skill between verification datasets.

2. Data and methods

a. *CONUS regions*

To explore FD predictability under different regions in CONUS, we utilized the National Climate Assessment–Land Data Assimilation System (NCA–LDAS) USDM regions mask (Fig. 1). CONUS is separated into six regions, which include the West, Midwest, High Plains, South, Southeast, and Northeast regions with Wyoming and Colorado belonging to both the west and High Plains regions, respectively. The NCA–LDAS USDM mask was regridded from 0.125° to 1° resolution to match the SubX model resolution, and our analysis uses the regridded mask to average results within each region.

b. *Observational reference data*

For the observational reference, we used three different datasets, including National Aeronautics and Space Administration Modern-Era Retrospective Analysis for Research and Applications, version 2 (NASA MERRA-2; Gelaro et al. 2017), National Oceanic and Atmospheric Administration GEFSv12 reanalysis (NOAA GEFSv12; Hamill et al. 2022), hereafter referred to as GEFSv12-obs, and North American Land Data Assimilation System, phase 2 (NLDAS-2; Xia et al. 2012).

MERRA-2 is a global atmospheric reanalysis that assimilates satellite and conventional observations to represent the current state of the atmosphere and land and covers the period from 1980 to the present with a latency of ~3 weeks after the end of each month at an approximate resolution of 0.5° × 0.625° for 72 hybrid-eta model levels (or 42 pressure levels) from the surface to 0.01 hPa. MERRA-2 is produced with the Goddard Earth

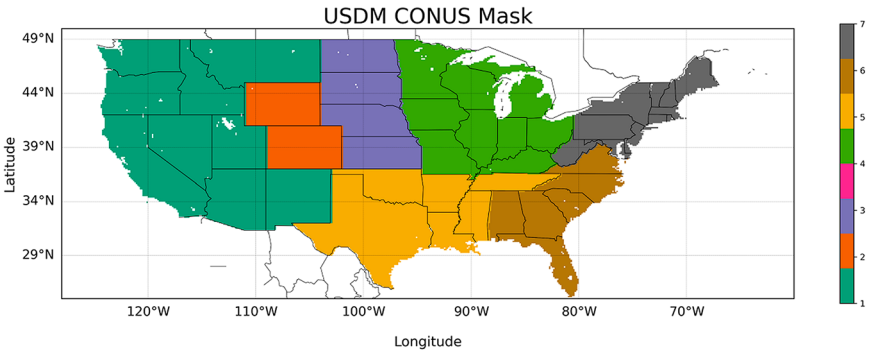

FIG. 1. Six USDM regions on NCA–LDAS grid. The West region (region 1) contains both Colorado and Wyoming (orange colored); the High Plains region (region 3) also contains Colorado and Wyoming; Midwest (region 4); South (region 5); Southeast (region 6); Northeast (region 7).

Observing System version 5.12.4 (GEOSv5.12.4) atmospheric data assimilation system, and the key components of the system are the GEOS AGCM and the Gridpoint Statistical Interpolation analysis scheme (Gelaro et al. 2017). Land surface variables are generated by forcing the Catchment LSM with bias-corrected precipitation, surface downward shortwave radiation, surface downward longwave radiation, humidity, temperature, wind, and surface pressure at the lowest model level taken from the GEOS AGCM (Reichle et al. 2017). We obtained daily mean fields for 2-m mean, maximum, and minimum temperature (T_{mean}, T_{max}, and T_{min}), 2-m dewpoint temperature, 10-m zonal wind speed, 10-m meridional wind speed, surface downwelling shortwave radiation, geopotential height at 200 hPa (z200), bias-corrected total precipitation, and RZSM (0–100 cm). All variables were bilinearly interpolated to a 1° × 1° grid over CONUS between 1 January 2000 and 31 December 2019. For all variables and grid cells, a 7-day rolling mean was applied to reduce day-to-day fluctuations, and anomalies were constructed by removing the climatology from the 7-day average. The climatology was created for each grid cell and each forecast date by taking the average of all values within a ±42-day window of the forecast date for each grid cell. Therefore, the daily climatology is the average of 85 days centered on the forecast date.

GEFSv12-obs is an atmosphere-only reanalysis with 590 variables produced four times daily at 0000, 0600, 1200, and 1800 UTC at a resolution of 0.25° × 0.25° between the years 2000 and 2019. The primary purpose for the creation of GEFSv12-obs was for GEFSv12 retrospective forecast (reforecast) initialization so that its initial conditions will be consistent with the operational data assimilation system Global Forecast System, version 16 (GFSv16), which produces real-time forecasts. The data assimilation procedure used is a hybrid 4D ensemble variation and includes the GFSv15.1 AGCM for model physics (Kleist and Ide 2015). Land surface variables are generated by forcing the GFS Noah LSM with precipitation, radiation, and temperature. Note that unlike MERRA-2, GEFSv12-obs does not bias correct its precipitation using observations before using it to drive Noah LSM, which may adversely affect the quality of GEFSv12-obs estimate of land surface state (e.g., soil moisture). We obtained the same variables as MERRA-2, and data were bilinearly interpolated to a 1° × 1° grid and restricted the domain to CONUS. Because each variable is generated at 0000, 0600, 1200, and 1800 UTC, we added all four time steps and took the mean for each variable (with the exception of T_{max}, T_{min}, and precipitation). For T_{max} and T_{min}, we obtained the maximum and minimum values of the four time steps, respectively. For precipitation, we obtained the accumulated sum for the four time steps. We generated a weighted average of 0–100-cm RZSM and created anomalies for all variables with the same methodology as MERRA-2 between 2000 and 2019.

NLDAS-2 is an offline data assimilation system using uncoupled land surface models which are driven by observation-based atmospheric forcings (Xia et al. 2012). Four different LSMs (Noah, Mosaic, Variable Infiltration Capacity, and Sacramento Soil Moisture Accounting) are forced with assimilated observations and model data from the North American Regional Reanalysis with each LSM having different model parameterizations to simulate hydrologic fluxes (Xia et al. 2012; Wood et al. 1997; Pinker et al. 2003). For our analysis, we utilized Noah, Mosaic, and Variable Infiltration Capacity models, which produce hourly forecasts over a portion of the North American continent, including CONUS and northern Mexico, which are generated at 1/8° resolution between 1979 and the present (Mitchell et al. 2004). Using an ensemble approach has been shown to produce better approximations than any individual model (Xia et al. 2014). Each LSM is forced by assimilated soil moisture initial state, wind, temperature, humidity, pressure, bias-corrected shortwave radiation (Pinker et al. 2003), longwave radiation, and bias-corrected total precipitation (same as MERRA-2). Additional forcings include a fraction of total precipitation that is convective and convective available potential energy, which is retrieved from the North American Regional Reanalysis. We obtained the same variables as MERRA-2 from NLDAS-2 forcing file A with the exception of z200, which was obtained from North American Regional Reanalysis. For RZSM, we obtained the 0–100 cm from Noah, Mosaic, and Variable Infiltration Capacity LSMs and took the mean of the three models. We created

TABLE 1. List of SubX models assessed, information about hindcast ensemble size, and initialization frequency. Additional details about the SubX models and references for individual models are provided in Pegion et al. (2019). All models are analyzed at a spatial resolution of 1° × 1°. Model component A = atmosphere, O = ocean, I = sea ice, and L = land.

Model	Model components	Ensemble members	Forecast length	Initialization frequency
NOAA GEFSv12	A, L	11	35 days	Every Wednesday
NASA GEOS-V2p1	A, O, I, L	4	45 days	Every 5 days

anomalies for all variables with the same methodology as MERRA-2 between 2000 and 2019.

c. *SubX retrospective forecasts (reforecasts)*

The SubX project is a multimodel subseasonal prediction experiment designed around operational requirements, and its unique feature is the inclusion of both research and operational models, which facilitates feedbacks between agencies to promote model development (Pegion et al. 2019). Eight global numerical models have produced more than 23 years of reforecasts and real-time forecasts (1999–2023) and are archived at the the International Research Institute for Climate and Society (IRI) Data Library (IRIDL) (https://iridl.ldeo.columbia.edu/SOURCES/.Models/.SubX/), or for NOAA GEFSv12 reforecast (GEFSv12-fcst), data are split between IRIDL and Amazon Web Services (https://noaa-gefs-retrospective.s3.amazonaws.com/index.html) (Guan et al. 2022). For our analysis, we use SubX models GEFSv12-fcst and Global Modeling and Assimilation Office Goddard Earth Observing System model V2p1 (GEOS-V2p1) (Table 1) because they are the only two forecast models that provide data needed for this study. Data retrieved include T_{mean}, T_{max}, T_{min}, accumulated precipitation, downwelling shortwave radiation, zonal and meridional wind speed, z200, 2-m dewpoint or specific humidity, and RZSM. Dewpoint and specific humidity were converted to actual vapor pressure. For each variable, anomalies were calculated using the same method as MERRA-2 for each ensemble member and lead time from weeks 1 to 4 and the average of all days within weeks 3 and 4.

Soil moisture data are available within four SubX models, including GEOS-V2p1, ESRL-Finite-Volume Cubed-Sphere Model r1p1 (FIMr1p1), Rosenstiel School of Marine and Atmospheric Science (RSMAS)-CCSM4, and GEFSv12-fcst. Vertically integrated soil moisture is provided within GEOS-V2p1, ESRL-FIMr1p1, and RSMAS-CCSM4 with depths ranging between 1.5 and 3.0 m across all grid cells, whereas volumetric soil moisture content is provided within GEFSv12-fcst for individual layers 0–10, 10–40, 40–100, and 100–200 cm. Although no specific rooting zone depth has been deemed optimal for FD identification within soil moisture indices, the upper 95th percentile rooting depth for 11 temperate agricultural crops was observed to be ~1.02 m (Fan et al. 2016), indicating that for FD analysis, SubX models providing vertically integrated soil moisture depths between 1.5 and 3.0 m may be too deep and unable to capture rapid changes in the layers in which crop roots predominately exist; therefore, only GEFSv12-fcst is considered for the soil-moisture-based FD analysis. Although FD is not specific to agriculture, during the growing season, this sector is highly impacted by moisture stress, and knowledge of rooting depths throughout the growing season can better inform the selection of RZSM datasets (Ford and Labosier 2017). In our analysis, we utilized GEFSv12-fcst volumetric soil moisture content and calculated a weighted average of the top three layers to obtain 0–100-cm RZSM.

GEOS-V2p1 was analyzed in its native spatial resolution of 1° × 1° across CONUS. GEFSv12-fcst produced data at 0.25° resolution (first 10 days) and 0.5° (11–35 days) and was bilinearly interpolated to 1° × 1°. Anomalies were calculated using the same method as MERRA-2 for each ensemble member and lead time from weeks 1 to 4 and the average of all days within weeks 3 and 4. While GEOS-V2p1 reforecasts are producing data up to the present day, GEFSv12-obs only spans 2000–19; therefore, all analyses were performed between 2000 and 2019.

d. *Soil moisture FD index*

To identify FD onset, the SMPD index was utilized describing rapid changes in RZSM from "above the 40th percentile to below the 20th percentile in 3 weeks or less." We also implemented a variant of this definition when RZSM drops from "above the 30th percentile to below the 10th percentile in 3 weeks or less" to examine FD onset that leads to more intense land surface impacts (Ford and Labosier 2017; Osman et al. 2021). FD duration has been suggested to persist for a minimum duration of 3 weeks in which drought impacts are more pronounced (Otkin et al. 2022), but this suggestion cannot be directly implemented into GEFSv12-fcst using the SMPD index. GEFSv12-fcst has a 5-week forecast, and classification of FD onset using the SMPD within 3 weeks only allows for maximum drought persistence of 2 weeks if FD intensification occurred over a 3-week time interval. It is possible for FD onset to occur within 2 weeks and the duration to last 3 weeks, but for our analysis, we are only interested in ensuring that the FD persists for at least 2 weeks. When longer lead forecasts become available, the FD duration methodology can be altered to the suggested minimum duration of 3 weeks.

For observational data, RZSM anomalies were converted into percentiles, and the three-dimensional (time, latitude, and longitude) time series was reformatted to a five-dimensional (initialized date, model, lead, latitude, and longitude) time series that matches the same lead time format as GEFSv12-fcst. The initialized date dimension for observations is merely a placeholder ensuring a common format because observations do not have an initialization date. For each reforecast initialized date and lead time, we selected the observation date which corresponds to the lead time difference from the initialized date. This produces a new dataset in which the observation dates are matched with the forecast lead dates at the

appropriate time step. Because we only have a single observation for each date and lead time, each model dimension has the same observational data. This preprocessing step is necessary because once FD onset occurs within the three-dimensional time series, the duration of the event persists until RZSM is above the lowest threshold. This duration continuation of FD leads to an inconsistent spatiotemporal comparison for identifying when RZSM depletes rapidly over the same time interval (results not shown). Our reformatting of observational data into the same lead time format as GEFSv12-fcst ensures that the classification of FD onset is equitably compared. FD was classified when the RZSM percentile decreased from above the highest threshold (40th or 30th percentile) to below the lowest threshold (20th or 10th percentile) within 3 weeks or less. For GEFSv12-fcst, we calculated the anomaly for each ensemble member, converted it to percentiles, and classified FD onset to identify probabilistic skill. We also computed the ensemble mean of anomaly which was converted into percentiles and FD classified. For both reforecast and reanalyses, once FD onset is classified, the drought persists into future days if RZSM is below the lowest threshold.

e. Evaporative demand (reference evapotranspiration)

To investigate atmospheric conditions that increase plant stress through excessive moisture loss or reduce plant efficiency through stomatal closures, we computed ET_o using the Food and Agriculture Organization (FAO) of the United States Irrigation and Drainage Paper 56 (FAO-56) Penman–Monteith equation (Allen et al. 1998). For our applications, the Penman–Monteith ET_o method determines ET_o from a short-grass, well-watered reference surface and provides a standard for comparison of ET_o in different periods of the year or in other regions that is independent of crop type, crop development, and management practices. The ET_o reference crop surface assumes a crop height of 0.12 m, the fixed surface resistance of 70 s m^{-1}, and an albedo coefficient of 0.23. The FAO-56 Penman–Monteith ET_o equation is

$$ET_o = \frac{0.408\Delta(R_n - G) + \gamma\dfrac{900}{T + 273}u_2(e_s - e_a)}{\Delta + \gamma(1 + 0.34u_2)},$$

where ET_o is the reference evapotranspiration (mm day^{-1}), R_n is the net radiation, G is the soil heat flux, e_s is the saturated vapor pressure, e_a is the actual vapor pressure, $(e_s - e_a)$ is the vapor pressure deficit of the air, u_2 is the 2-m mean wind speed, T is the 2-m mean temperature, Δ is the slope of the saturation vapor pressure curve, and γ is the psychrometric constant. The 10-m uas and vas are combined to produce wind speed and interpolated to 2 m using the logarithmic conversion factor (Allen et al. 1998). Due to the diurnal cycle of heat transfer, G was assumed to be zero. For detailed descriptions of each input and parameter estimation, the readers refer to Allen et al. (1998).

f. Mean absolute error

To compare differences between reanalyses, we compute the mean absolute error (MAE) between each set of reanalyses for each grid cell and observation date between 2000 and 2019.

$$\mathrm{MAE}_i = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|,$$

where i is the grid cell, j is the observation date, y is the first reanalysis, and $\hat{y}$ is the comparison reanalysis.

g. Anomaly correlation coefficient

One metric to evaluate deterministic forecast is the anomaly correlation coefficient (ACC) which describes the correlation between a forecast anomaly relative to forecast climatology and a verifying observation anomaly relative to observation climatology. ACC indicates the strength of association for both direction and magnitude of forecasts relative to observations with values falling between +1 and −1.

$$\mathrm{ACC}_i = \frac{\sum_{m=1}^{M}(f_m - \overline{f})(o_m - \overline{o})}{\left[\sum_{m=1}^{M}(f_m - \overline{f})^2\sum_{m=1}^{M}(o_m - \overline{o})^2\right]^{1/2}},$$

where i is the grid cell, M is the total number of time samples, f_m is the forecast, $\overline{f}$ is the forecast climatology, o_m is the observation, and $\overline{o}$ is the observation climatology. According to the ECMWF, ACC values approaching +1 indicate strong positive agreement where forecasts have value; ACC values $> +0.6$ indicate synoptic prediction skill for large weather patterns; ACC values approaching +0.5 indicate the forecast errors are similar to errors from a forecast of the climatological average; and ACC values around 0 indicate the forecast has no value (Santoalla 2018). For each grid cell, we calculated the ACC using the mean of all ensemble members for the average of individual weeks 1–4 as well as the average of all days between weeks 3 and 4.

h. True positive, false positive, and critical success index

FD onset is a binary categorical event, where 1 indicates FD is present and 0 indicates FD is absent. Forecasts are evaluated using a contingency table in which each forecast lead time is verified against observations from the same day where the random variable true positive (TP) indicates the forecast and observation both classified FD, false positive (FP) when the forecast predicted FD but the observation did not, false negative (FN) when the forecast forecasted no FD when observations classified FD, and true negative (TN) when both forecast and observations had no FD. Using these random variables from forecasts allows for assessment of overall skill in correctly or incorrectly predicting FD and for identification of areas where forecast performance needs to be improved. The true positive rate (TPR), also known as sensitivity, is a performance measure commonly used to assess the percentage of correct forecast predictions of a categorical time series compared to all predicted events, including events that did not occur in observations [World Weather Research Programme (WWRP) 2017].

$$\mathrm{TPR}_i(\%) = \frac{\sum \mathrm{TP}}{\sum \mathrm{TP} + \sum \mathrm{FN}} \times 100,$$

where i is the grid cell and TPR values range from 0% to 100%, where 100% indicates that all flash droughts were correctly identified with no FN forecasts. We calculated TPR for each grid cell and averaged results by region for each SMPD variant [(i) decrease in RZSM from above the 40th to below the 20th percentile and (ii) decrease in RZSM from above the 30th to below the 10th percentile]. The false positive rate (FPR) assesses the percentage of predictions that were incorrectly classified as positive and is the ratio of incorrectly forecasted events when FD occurred to the total number of actual negative events (WWRP 2017). FPR measures how frequently a forecast wrongly predicts FD when it did not occur:

$$\mathrm{FPR}_i(\%) = \frac{\sum \mathrm{FP}}{\sum \mathrm{FP} + \sum \mathrm{TN}} \times 100,$$

where i is the grid cell and FPR results were averaged by region for each SMPD variant. The critical success index (CSI) is used to measure the accuracy of correctly predicted forecasts when TN has been removed (Schaefer 1990):

$$\mathrm{CSI}_i(\%) = \frac{\sum \mathrm{TP}}{\sum \mathrm{TP} + \sum \mathrm{FP} + \sum \mathrm{FN}} \times 100,$$

where i is the grid cell and results are averaged by region for each SMPD variant. CSI is dependent on climatological frequency with lower CSI percentages obtained for rarer events with a score of 100% indicating a perfect forecast (WWRP 2017). The CSI has been applied for various meteorological events such as tornadoes, drought, and winter thunderstorms (Gerapetritis and Pelissier 2004; Schaefer 1990; Hobeichi et al. 2022), but currently, there is no CSI threshold which informs users of forecast skill across all meteorological events, and expert judgment must be employed. For FD, there is currently no threshold for TPR, FPR, or CSI, which indicates that forecasts are highly skillful, but rather higher percentages implicitly refer to forecasts, which more accurately forecast FD.

3. Results

a. Forecast skill of ET_o and its forcing variables

For each SubX model, CONUS region, season, and variable (including ET_o, z200, T_{mean}, T_{max}, T_{min}, wind speed, downwelling shortwave radiation, accumulated precipitation, and actual vapor pressure), we examined anomaly skill of the ensemble mean for each lead time from the weekly average of individual weeks 1–4 as well as the average of days between weeks 3 and 4. Week 1 skill was high (~0.60–0.95 ACC) for ET_o, z200, T_{mean}, T_{max}, T_{min}, wind speed, and actual vapor pressure (avp) for both SubX models when referenced against GEFSv12-obs, MERRA-2, and NLDAS-2 (Fig. 2; Figs. S1–S6 in the online supplemental material). GEFSv12-fcst is more skillful than GEOS-V2p1 for ET_o, z200, shortwave radiation, and T_{mean} (spring and fall seasons). GEOS-V2p1 is more skillful for T_{mean} during summer, T_{min} for the Southeast and Northeast regions, and precipitation for all regions during the spring. Radiation was less skillful (~0.50–0.68 ACC) with the highest skill agreement between GEFSv12-fcst and GEFSv12-obs for most regions and seasons (Fig. S7). Precipitation skill was variable between SubX models and reanalyses (~0.25–0.65 ACC), and during spring, GEOS-V2p1 was more skillful for all regions, but GEFSv12-fcst has higher skill for summer and fall (Fig. S8). The fall season was most skillful for precipitation at lead week 1.

Between weeks 3 and 4, summer has higher skill for z200, T_{mean}, T_{max}, T_{min}, and avp (Figs. S1–S4 and S6). ET_o and shortwave radiation skill is highest in the fall season (Fig. 2; Fig. S7), and wind speed has the lowest skill in the fall (Fig. S5). During summer, high variability exists between SubX models and reanalyses for T_{max} and ET_o with GEFSv12-fcst having the highest skill. Shortwave radiation also has high variability between SubX models and observations during fall. The highest skill is observed during summer for T_{mean}, T_{max}, T_{min}, and z200 (~0.5 ACC), which is the threshold for determining when forecast errors equal errors when using the climatological average. ACC for ET_o ranges from 0.00 to 0.55; shortwave radiation ranges from 0.0 to 0.5; wind speed ranges from 0.0 to 0.2; and precipitation ranges from 0.0 to 0.15. For all variables, no region is observably more skillful than other regions, but Southeast ET_o is least skillful during summer and the Northeast is least skillful during spring. Averaging anomalies between weeks 3 and 4 has induced slightly higher skill than individual weeks 3 or 4, while the averaging only increases skill beyond the 0.5 threshold for ET_o during fall, z200 during summer, and T_{mean}, T_{max}, and T_{min} during summer. We further examined the ET_o skill of individual grid cells because averaging by region masks individual grid cells and may not be suitable for understanding spatial forecast variability (Fig. S9). The ET_o skill is >0.5 during individual weeks 3–4 for the Northwest during summer and fall, the Northeast coastal regions during fall, and portions of the Southwest during fall.

b. Comparison of RZSM between reanalyses

We first investigate the representation of seasonal RZSM anomalies between MERRA-2, NLDAS-2, and GEFSv12-obs to assess similarities and differences over spatial and temporal scales (Fig. 3). The mean absolute error between MERRA-2 and NLDAS-2 is the smallest in all comparisons, and results indicate that spring season differences are higher in the Southeast and West regions, while the summer season has higher differences in the central United States. The largest difference exists when comparing MERRA-2 and GEFSv12-obs with the highest difference occurring during spring in the eastern and western regions, while summer has the highest difference in the central United States and a smaller difference during fall in the central United States. The mean absolute error between GEFSv12-obs and NLDAS-2 shows a high difference in the western United States during spring, central United States during summer, and central/eastern United States during fall. When compared to NLDAS-2, GEFSv12-obs performs better than MERRA-2 in the spring in the eastern United States, and GEFSv12-obs performs similarly to MERRA-2 in summer and fall in the West and Northeast.

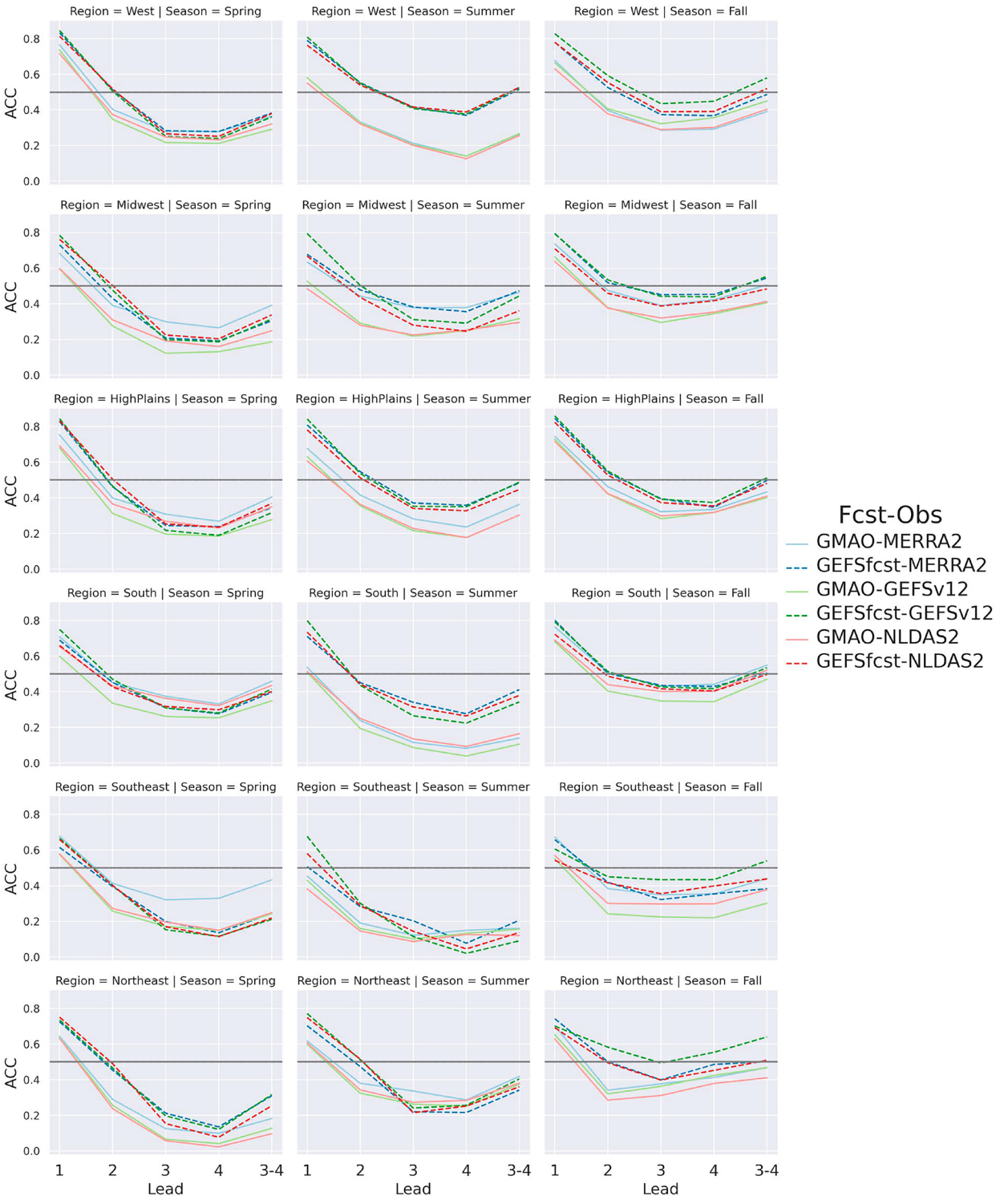

FIG. 2. FAO-56 Penman–Monteith ET_o ACC skill for GMAO-V2p1 and GEFSv12-fcst when referenced against GEFSv12-obs, NLDAS-2, and MERRA-2 reanalyses. Data are separated by season, region, and weekly leads 1–4 as well as the average of weekly leads 3–4 for years 2000–19. The spring season contains months of March, April, and May; the summer season contains months of June, July, and August; and the fall season contains months of September, October, and November.

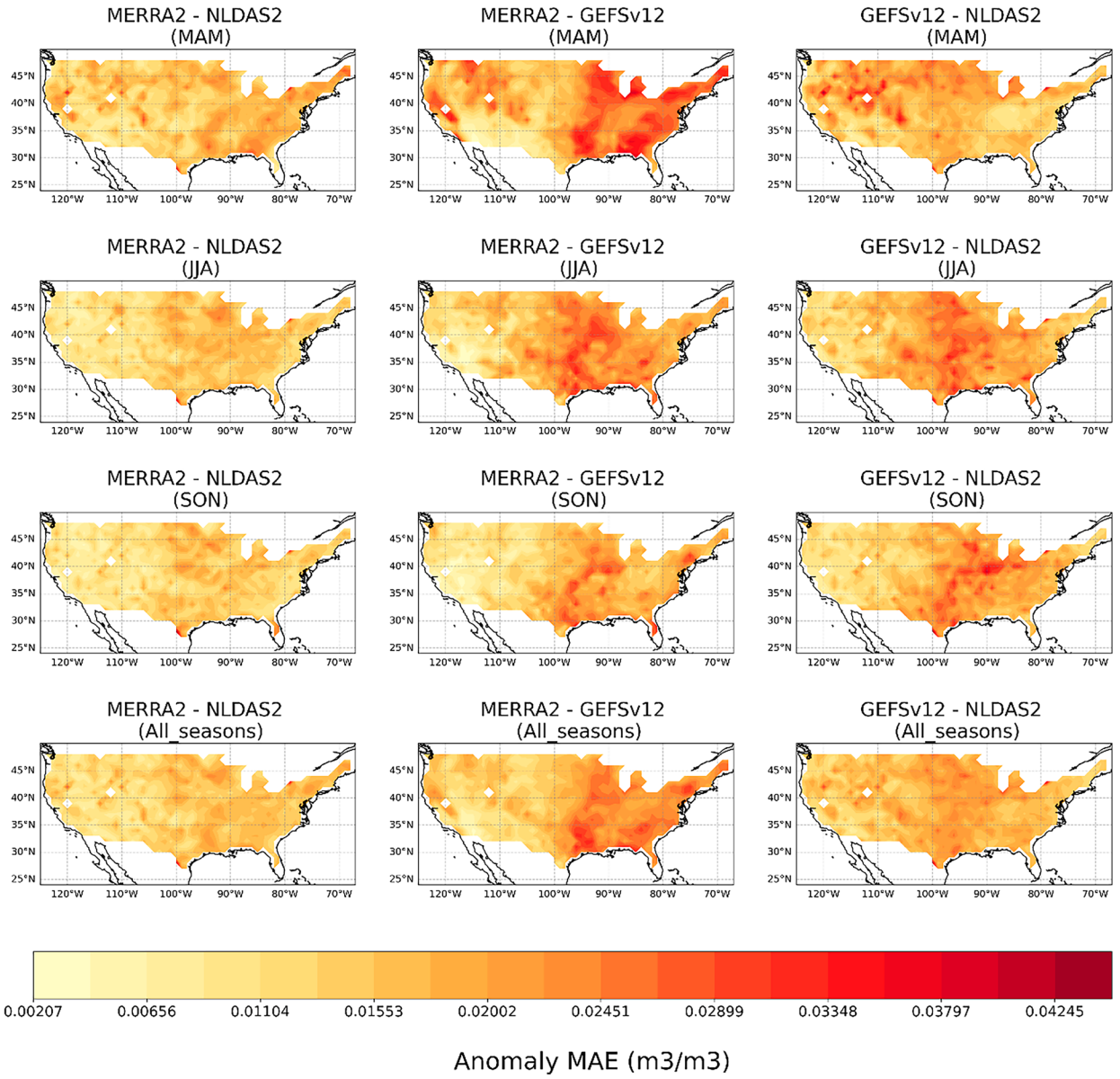

FIG. 3. RZSM anomaly MAE between MERRA-2, NLDAS-2, and GEFSv12-obs. MAE is computed between each reanalysis and comparisons separated by season with MAM (spring) being March, April, and May; JJA (summer) being June, July, and August; SON (fall) being September, October, and November; and All_seasons being all months between March and November.

c. *RZSM forecast skill*

GEFSv12-fcst was more correlated with GEFSv12-obs than with MERRA-2 or NLDAS-2 with week 1 differences in ACC between reanalyses ranging from 0.15 to 0.50 with the highest differences observed in the Midwest, Northeast, and Southeast regions (Fig. 4). Weeks 3–4 ACC differences against three reanalyses are minimized and ranged from 0.00 to 0.25. GEFSv12-fcst against GEFSv12-obs has the highest skill across all leads. For week 1, ACC skill ranged between 0.75 and 0.85 across regions with minimal seasonal differences. Between weeks 3 and 4, the High Plains and South regions retained the highest skill of all regions (~0.6–0.7 ACC in summer and fall) followed by the Midwest, West, and Southeast regions (~0.45–0.60 ACC), and the Northeast was the least skillful (~0.29–0.37 ACC). The skill of anomalies for the average of weeks 3–4 was slightly higher than the skill of individual week 3 or 4 for all regions/seasons.

We further examined the RZSM skill of individual grid cells because averaging by region masks individual grid cells and may not be suitable for understanding spatial forecast variability, and we only examined results between GEFSv12-fcst when referenced against GEFSv12-obs (Fig. 5). For week 1, the central and western United States has the highest skill during spring and summer (0.85–0.95 ACC), and during fall,

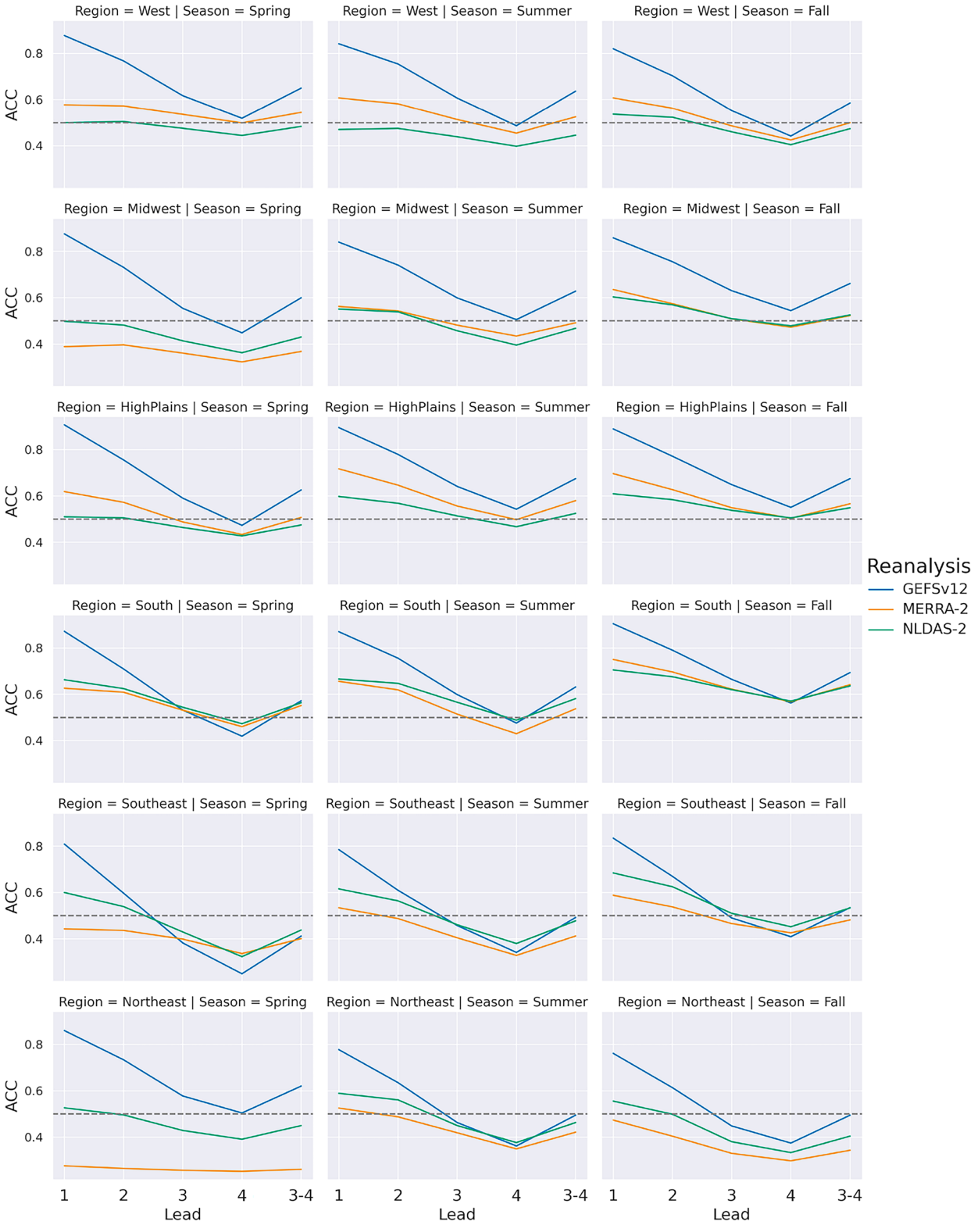

FIG. 4. RZSM ACC skill from GEFSv12-fcst when referenced against MERRA-2, GEFSv12-obs, and NLDAS-2 reanalysis. Data are separated by region, season, and weekly leads 1–4 as well as the average of leads 3–4. The spring season contains months of MAM, the summer season contains months of JJA, and the fall season contains months of SON.

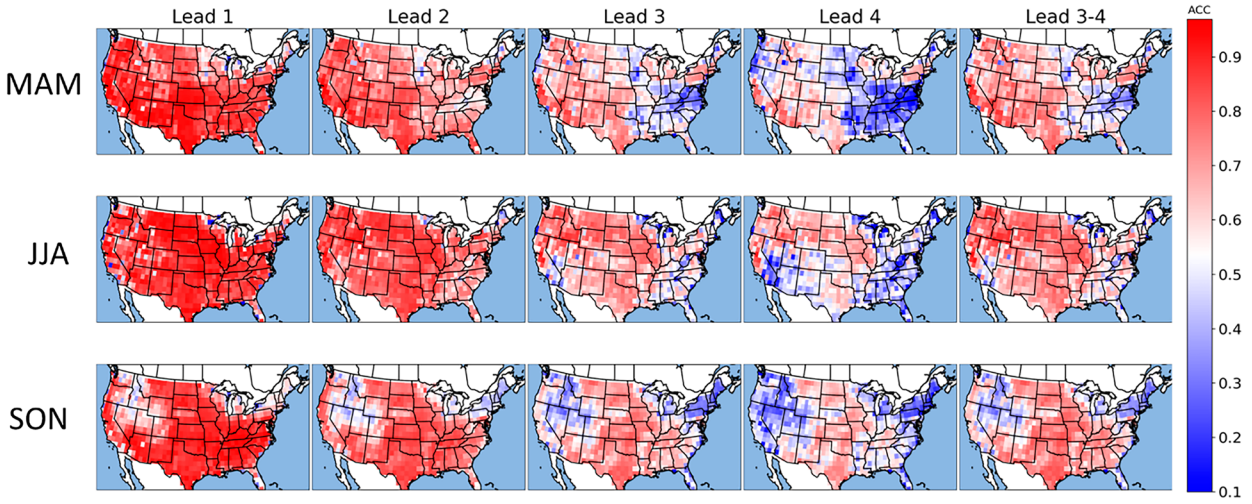

FIG. 5. GEFSv12-fcst RZSM ACC when referenced against GEFSv12-obs. Data are separated by season and lead times 1–4 and the average of leads 3–4.

the Southeast has similar skills. During weeks 3–4, the highest skill is observed in the western United States and Northeast during spring, the central United States and Northwest during summer, and the central United States and Southeast during fall. The state of Florida and the southeast Atlantic coast have low skill. The average of weeks 3–4 has higher skill than individual week 3 or 4 with spatial skill patterns similar to week 3. For all regions during summer, reforecast week 1 anomalies are generally wetter than the observation anomalies, which is indicated by the median value being below zero, and week 4 anomalies during fall tend to be drier than observations (Fig. 6). Density plot comparisons between GEFSv12-fcst, MERRA-2, and NLDAS-2 indicate that MERRA-2 and NLDAS-2 have larger anomaly differences for weekly leads 1 and 4 when referenced against GEFSv12-obs with the High Plains and west regions having the smallest bias among all three reforecasts (Figs. S10 and S11). For GEFSv12-fcst dry anomalies, the skill was assessed using soil moisture percentile thresholds for U.S. Drought Monitor drought categories (Table S1) with results indicating for weeks 3 and 4, very dry anomalies between the 0th and 10th percentile are more predictable (~0.50–0.85 ACC) than anomalies between the 11th and 30th percentile (~0.20–0.55 ACC) (Fig. 7). Across all leads, skill was lowest when anomalies were between the 21st and 30th percentiles.

d. *Soil moisture flash drought index*

We first assessed TPR, FPR, and CSI rates between GEFSv12-obs and GEFSv12-fcst for weeks 3 and 4 for the SMPD index using the ensemble mean when RZSM drops from above the 40th to below the 20th percentile in 3 weeks or less (Fig. 8). TPR percentages range from 4% to 33% across regions and seasons with the Northeast having the highest TPR during spring (28%–33%). The West has similar TPR across all seasons but slightly higher during summer (15%–18%), the Midwest has the highest TPR in spring (13%–15%), the High Plains has the highest TPR during summer (15%–18%), the South has the highest TPR during spring (16%–19%), and the Southeast has the highest TPR during fall (17%). FPR was always less than TPR for each region with the fall season generally having higher FPR across all regions with percentages ranging from 1% to 5%. CSI percentages range from 1% to 15% across regions and

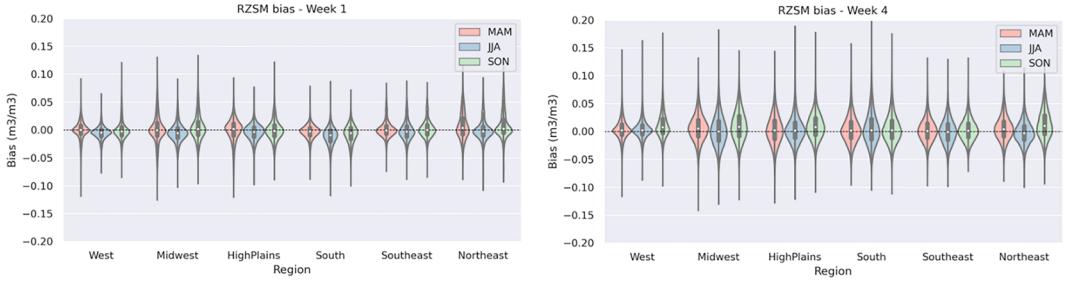

FIG. 6. GEFSv12-fcst RZSM anomaly biases for weeks 1 and 4 when referenced against GEFSv12-obs. Data are separated by season and region. Bias is calculated by subtracting reforecast anomaly from reanalysis anomaly. Negative values indicate reforecast conditions were wetter than reanalysis, and positive values indicate reforecast conditions were drier than reanalysis. The white dots for each region and season are the median.

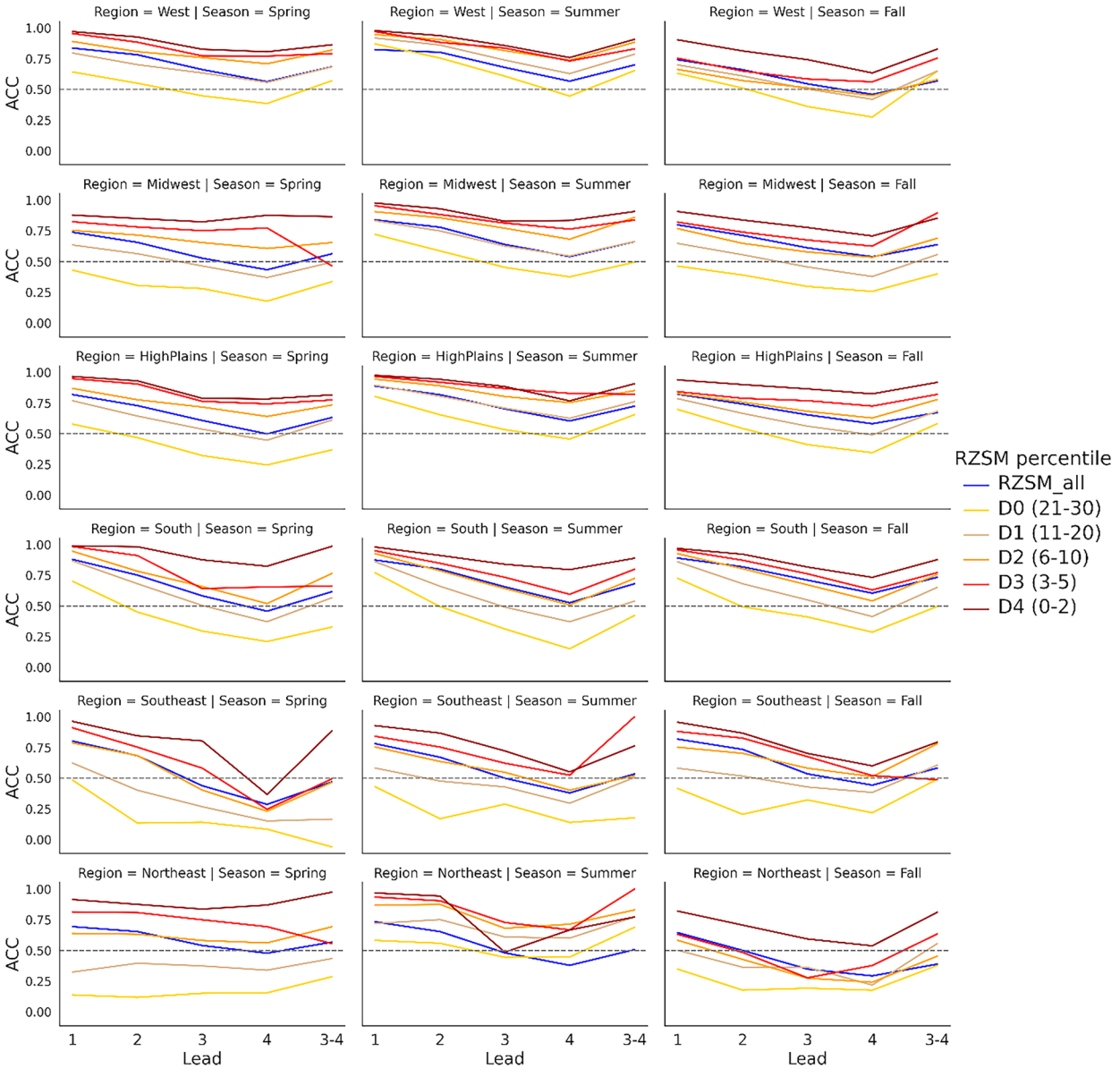

FIG. 7. GEFSv12-fcst ACC skill between 2000 and 2019 separated by season, region, and weekly lead time when referenced against GEFSv12-obs. RZSM_all displays the skill for all anomalies, D0 is the skill when anomaly is between the 21st and 30th percentile, D1 is the skill when anomaly is between the 11th and 20th percentile, D2 is the skill when anomaly is between the 6th and 10th percentile, D3 is the skill when anomaly is between the third and fifth percentile, and D4 is the skill when anomaly is between the zeroth and second percentile.

seasons with the Northeast having the highest CSI during spring. TPR, FPR, and CSI for SMPD from above 30th to below 10th percentile were similarly distributed, but rates were always lower for all three performance measures (Fig. S12). Probabilistically, we observe that taking the ensemble mean (Fig. 8) for the "above 40th to below 20th," SMPD criteria are more skillful for almost all regions, seasons, and leads when compared to using individual models and classifying FD when at least 30%, 50%, or 70% of the ensemble members agree (Figs. S13–S15). When evaluating results between GEFSv12-fcst, MERRA-2, and NLDAS-2 (Figs. S16 and S17), TPR was similar to GEFSv12-obs (Fig. 8), but all regional and seasonal CSI rates were less than GEFSv12-obs indicating poorer agreement.

e. *Flash drought onset case studies*

To first establish the agreement between MERRA-2, NLDAS-2, and GEFSv12-obs, we evaluate the progression of RZSM anomalies during May–July 2012, and we discuss findings only within the central United States (Figs. S18–S23).

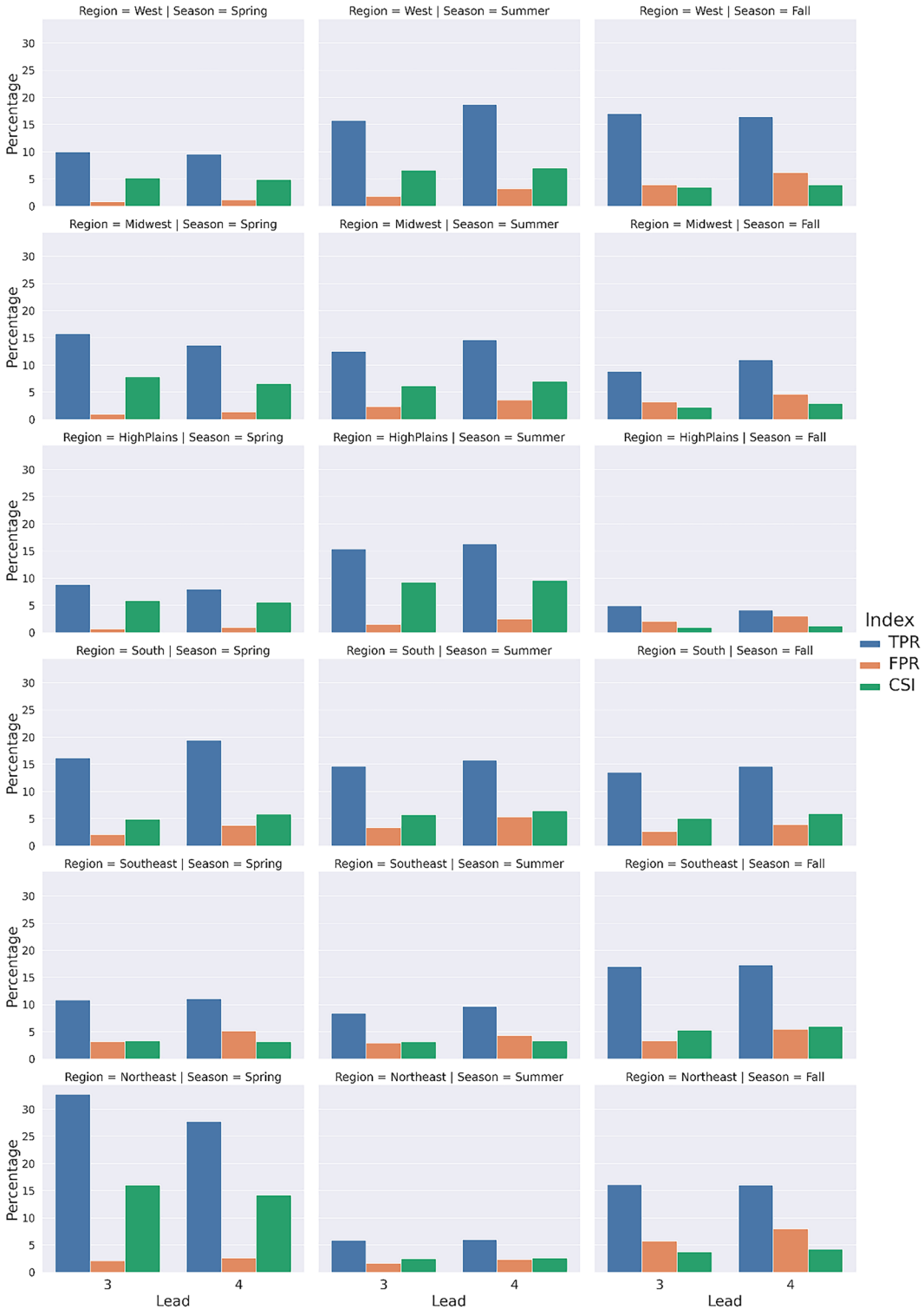

FIG. 8. TPR, FPR, and CSI for the SMPD index when RZSM dropped from above the 40th percentile to below the 20th percentile in 3 weeks or less. Data are separated by region, season, and weekly leads 3 and 4.

The 2012 central U.S. FD began in May and continued progression through July across various parts of central CONUS with the lowest RZSM occurring around mid-July (Lesinger and Tian 2022; Jin et al. 2019; Deangelis et al. 2020; Basara et al. 2019; Hoerling et al. 2014). May 2012 anomalies varied across reanalyses (Fig. S18), but in general, we observe that during the first week of May, all three reanalyses have similar RZSM spatial anomalies. Using NLDAS-2 as the ground truth, the difference in May anomalies is more pronounced between GEFSv12-obs than with MERRA-2, indicating that GEFSv12-obs can be highly divergent and is not strongly representative of conditions in which other reanalysis generally agree (Fig. S19). During week 2, MERRA-2 and NLDAS-2 are in agreement, while GEFSv12-obs had more wet anomalies in the east-central United States; during week 3, all reanalyses are similar with the exception of a strong wet tongue in GEFSv12-obs in the south-central United States; week 4 had closer agreement between MERRA-2 and NLDAS-2 in the east-central United States, while NLDAS-2 and GEFSv12-obs were more similar in the west-central United States; and week 5 had closer agreement between GEFSv12-obs and NLDAS-2. During weeks 3–5 west of the 96th meridian, MERRA-2 underestimated RZSM anomalies, whereas GEFSv12-obs overestimated anomalies.

During June 2012 weeks 1–4 (Fig. S20), spatial RZSM anomalies were similar between all three reanalyses with the exception of wet tongues in GEFSv12-obs that were present in northern and southern portions of the central United States. Assuming NLDAS-2 to be the ground truth, MERRA-2 had higher similarities in anomaly strength and direction than GEFSv12-obs (Fig. S21). During July 2012, wet tongues were not present in GEFSv12-obs, and we observed that all three reanalyses had strong agreement in the location of the highest RZSM dry anomalies in the central United States (Fig. S22). In the central United States, GEFSv12-obs overestimated RZSM dry anomalies, and MERRA-2 underestimated anomalies (Fig. S21). Overall, we observe that the progression of spatial RZSM anomalies, as well as intensities, was varied between reanalyses and that GEFSv12-obs had performance issues regarding wet tongues in June, but GEFSv12-obs May and July anomalies were well captured in the central United States. Across CONUS, GEFSv12-obs had several instances in which regional anomalies were of the opposite sign when compared to MERRA-2 or NLDAS-2 such as in the lower High Plains region (Fig. S20). MERRA-2 captured strong negative RZSM anomalies in the east-central United States and generally underestimated west-central U.S. negative anomalies.

Our evaluation of GEFSv12-obs and GEFSv12-fcst is beneficial for understanding the maximum achievable skill of GEFSv12-fcst under perfect forecast conditions, and we exclusively analyzed FD onset for major FD events in 2012, 2017, and 2019 between GEFSv12-obs and GEFSv12-fcst because of higher long-lead skill (Fig. 4), adequate performance in representing central U.S. RZSM anomalies (Figs. S18–S23), and higher flash drought metric scores (Fig. 8; Figs. S16 and S17). It should be noted that our methodology only captures the rapid decrease in RZSM which indicates FD and does not provide information on preexisting areas in FD. We chose to investigate three different initialized dates separated by 3 weeks using the SMPD variant "above the 40th to below the 20th percentile in 3 weeks or less" (Fig. 9). Because of land surface heterogeneity across CONUS, choosing initialization dates separated by 3 weeks allows for better visualization of FD onset, which affected different regions at different times.

For initialization on 9 May 2012 and forecast lead week 3 (Fig. 9a), the SMPD index was able to capture FD across Louisiana, Arkansas, and Iowa, which was likely induced by low precipitation and high ET_o associated with a high pressure ridge, but the upper Midwest was not predicted correctly due to the slight misplacement and intensity of the ridge when compared to observations. Week 4 predictions show a continuation of week 3 FD in most areas, but a high number of FP classifications occurred due to the misplacement of the high and low pressure z200 anomalies, which decreased ET_o and provided a small amount of precipitation in certain regions. The large number of FP classifications that are grouped around TP classifications does provide some reassurance that the model can identify areas where FD may be induced, but the forecast currently lacks grid cell specificity. For initialization on 30 May 2012 and forecast lead 3 (Fig. 9b), FD was correctly classified for a few grid cells in the Midwest but missed other surrounding areas due to the misalignment of z200 anomalies as well as ET_o intensity and precipitation improperly forecasted. Week 4 predictions accurately forecast some FD in the eastern Midwest, but we observe that these areas were already classified as being in FD during week 3 indicating that the forecast was 1 week late in correctly predicting FD. For initialization on 20 June 2012 and forecast lead 3 (Fig. 9c), portions of the central United States are correctly predicted, but FD development in the southern United States was incorrectly predicted due to no precipitation being forecasted, which was present in observations. The misrepresented precipitation in week 3 continued to impact FD forecasts during week 4 in the South with a large area falsely identified as being in FD, and in the Northwest, FD was falsely identified because week 4 precipitation was incorrectly forecasted.

The 2017 northern Great Plains FD is noted to have a decrease in RZSM from above the 80th to below the 17th percentile between 18 May 2017 and 8 June 2017 with low precipitation being the primary driver, and regions most affected include eastern Montana and western North Dakota (Hoell et al. 2019). Initialization on 19 April 2017 showed almost no SMPD signal except for lead 4 in northern Minnesota with large differences between reforecast and reanalysis for ET_o, z200, and precipitation anomalies (Fig. 10a). Reforecast initialization on 17 May 2017 also failed to predict FD in the northern High Plains with the z200 ridge misaligned and precipitation incorrectly forecasted (Fig. 10b). We observe that GEFSv12-fcst predicted a large area of FD in the Southeast for forecast lead 4, which did not occur due to heavy precipitation within observations during forecast lead 3, which increased RZSM. No other initializations captured signals for the northern High Plains for dates between 18 May and

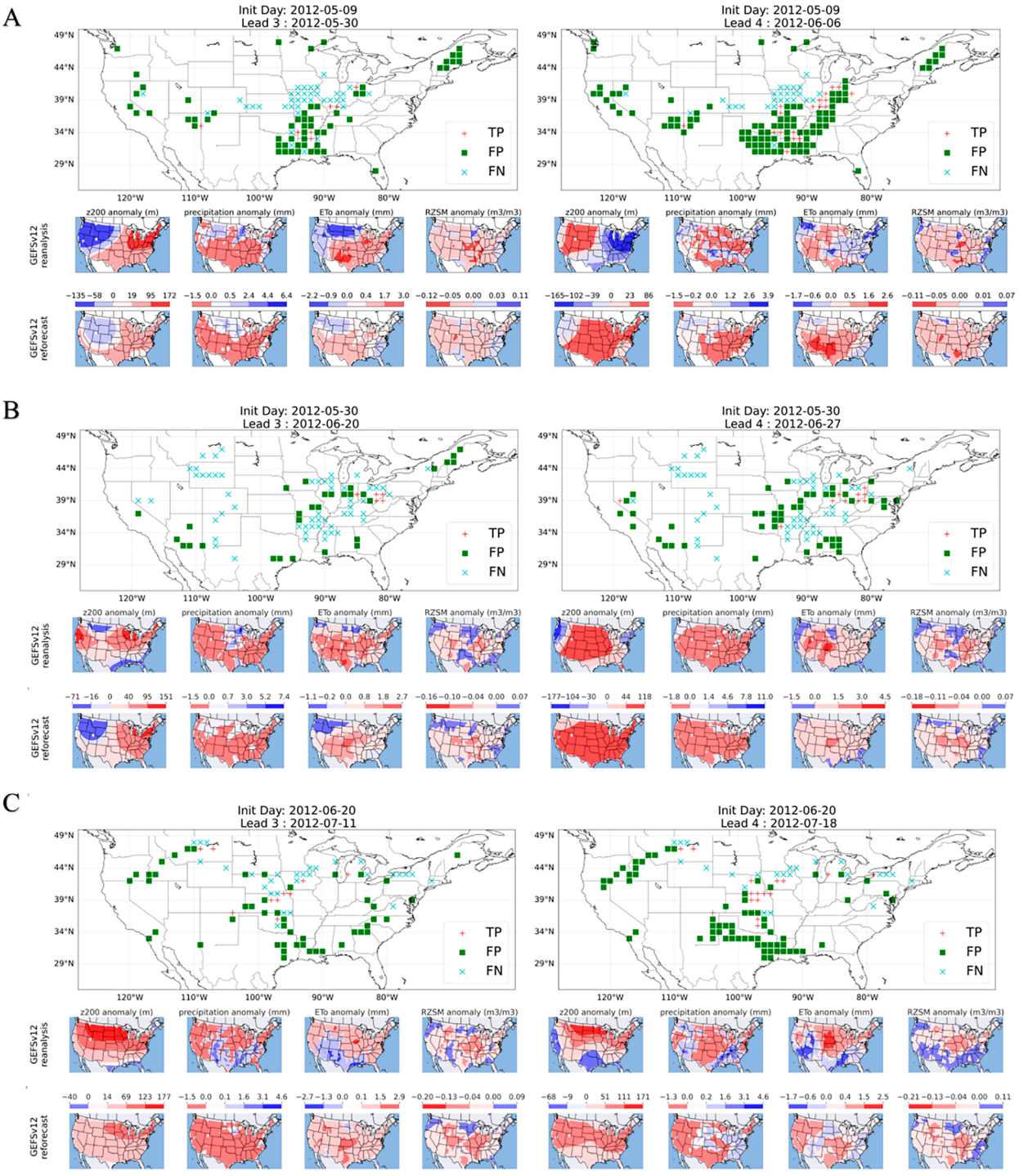

FIG. 9. FD onset classification for lead weeks 3 and 4 separated by 3-week initialization dates: (a) 9 May 2012, (b) 30 May 2012, and (c) 20 Jun 2012. Average weekly anomalies of z200, precipitation, ET_o, and RZSM are displayed beneath each TP, FP, and FN plot.

8 June 2017. Drought conditions worsened across the Northwest during July 2017, and we observed a moderate agreement for initialization on 21 June 2017 for lead 3 for z200, precipitation, ET_o, and RZSM and a strong positive agreement for FD classification at forecast lead 4 (Fig. 10c).

The 2019 Southeast FD began intensification in mid-September (Schubert et al. 2021), and we investigated weekly initializations between 21 August and 4 September 2019. Initialized forecasts on 21 August 2019 identify that GEFSv12-fcst missed the early stages of FD at weekly lead 3 due to the misplacement of the

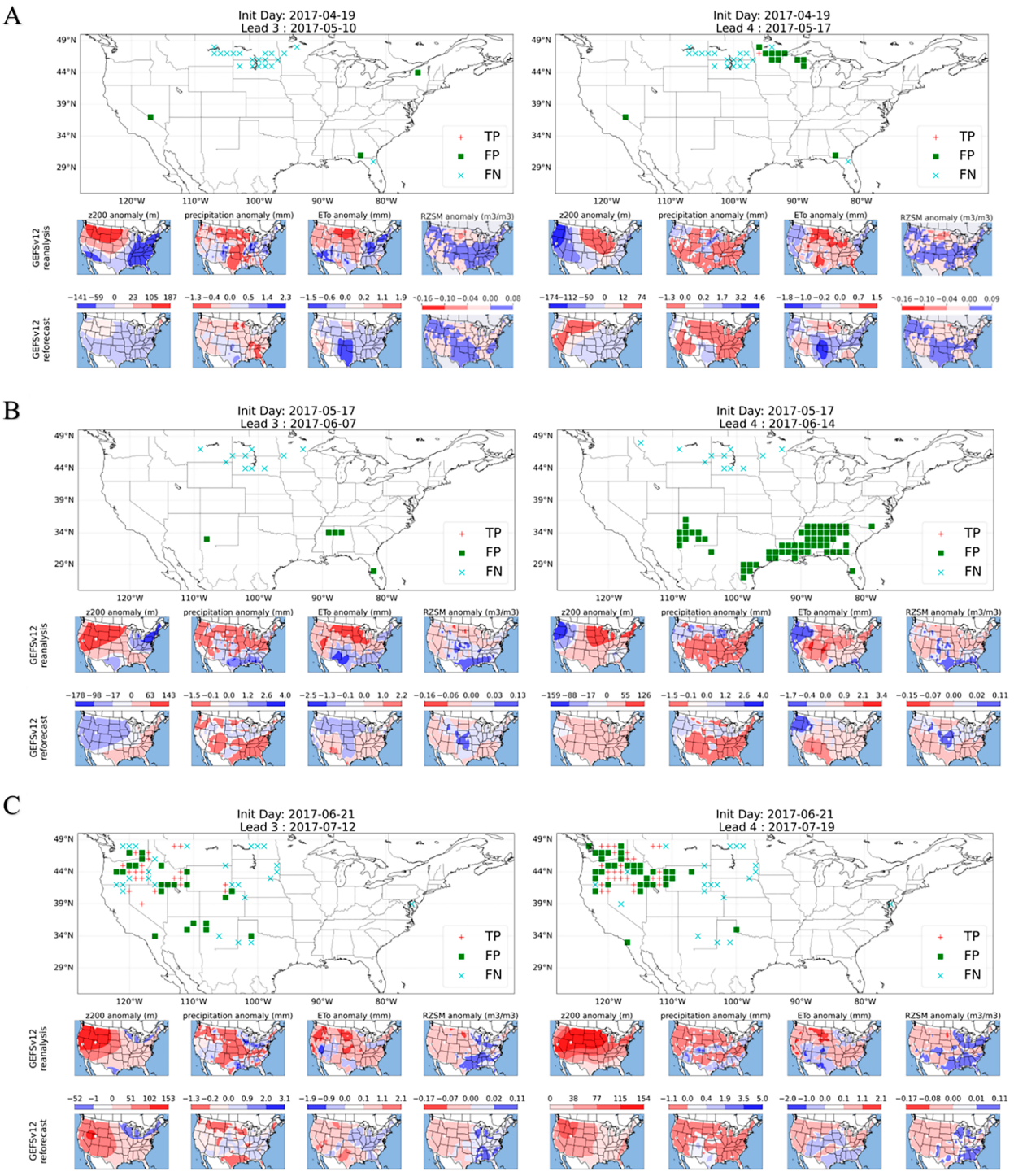

FIG. 10. As in Fig. 9, but for initialization dates (a) 19 Apr 2017, (b) 17 May 2017, and (c) 21 Jun 2017.

high pressure ridge further north than observed, which directly impacted ET_o forecasts (Fig. 11a). Week 4 forecasts also had the z200 ridge intensity misrepresented, and ET_o was incorrectly forecasted. For initialized forecasts on 28 August and 4 September 2019, we observed no FD signal with all anomalies (except precipitation) not well represented in the Southeast (Figs. 11b,c).

4. Discussion

a. *Factors influencing RZSM forecast skill*

We identified a large variation in RZSM skill when evaluating GEFSv12-fcst against MERRA-2, GEFSv12-obs, and NLDAS-2, with the highest skill obtained when GEFSv12-fcst was referenced

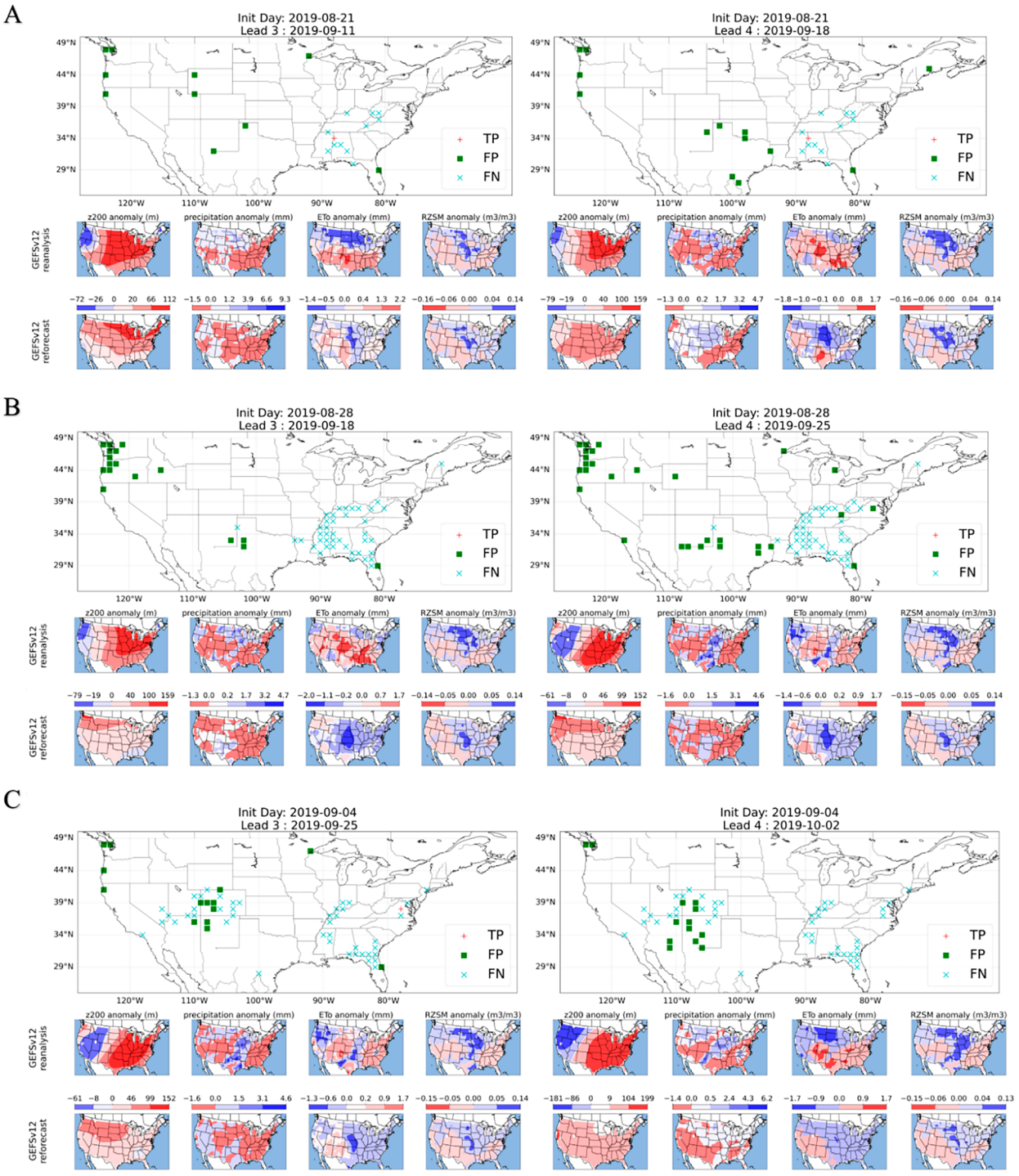

FIG. 11. As in Fig. 9, but for initialization dates (a) 21 Aug 2019, (b) 28 Aug 2019, and (c) 4 Sep 2019.

against GEFSv12-obs (Fig. 4). The large RZSM skill variation between reanalyses and forecasts results from differences between 1) AGCMs, 2) atmospheric and land state initializations, 3) LSMs, and 4) model spatial resolution. We observe that different reforecast–reanalysis outputs from different AGCMs created minor skill differences for z200, T_{mean}, T_{min}, actual vapor pressure, precipitation, and wind speed with larger skill differences in shortwave radiation, ET_o, and T_{max}. This implies that many variables are similarly predicted across reforecasts, but because some of these deterministic outputs are used as background forcings within LSMs, atmospheric deviations can inhibit RZSM forecast fidelity through mechanisms

such as misidentifying convective triggers associated with precipitation (Suhas and Zhang 2014) or improper representation of cloud cover and associated radiative feedbacks (Arakawa 2004; Park et al. 2006). Therefore, long-term statistics of atmospheric components may provide some relevant information concerning the degree to which RZSM forecasts are skillful, but because flux biases can negatively impact forecasts, more research is needed to verify the sources of highly biased predictions for early lead subseasonal forecasts, particularly for precipitation and radiation, which are primary components of the global water and energy cycle influencing drought. Research has identified spatiotemporal variations between boundary and initial conditions, which increase monthly precipitation skill over North America and Northern Hemisphere between weeks 1 and 6 (Sun et al. 2022; Dutra et al. 2021), but an in-depth analysis into sources of precipitation or radiation biases within subdaily subseasonal forecasts (e.g., GEFSv12-fcst) has not been quantified and could be explored in future research to improve long-lead predictions.

While estimates of RZSM between reanalyses are subject to the performance of their individual AGCM and LSM in simulating land surface processes, the accuracy of atmospheric and land surface forcing initial conditions also plays a role (Materia et al. 2014; Koster et al. 2010). MERRA-2 and NLDAS-2 use CPC bias-corrected precipitation, whereas GEFSv12-obs uses CPC nonbias-corrected Merged Analysis of Precipitation, which may adversely impact GEFv12-obs estimate of RZSM and potentially lead to less accurate RZSM forecasts. GEFSv12-fcst is more correlated with GEFSv12-obs, and the high skill of GEFSv12-obs against GEFSv12-fcst is expected due to the common AGCM and LSM used in these two products, whereas MERRA-2 and NLDAS-2 have different AGCMs and LSMs. One limitation of GEFSv12-fcst is that GEFSv12-obs initial land state forcing includes neither direct data assimilation of RZSM nor insertion of RZSM initial state from GLDAS reanalysis, meaning that RZSM is simulated throughout the entire reanalysis (Hamill et al. 2022). Because RZSM is simulated, this implies that the GEFSv12-fcst theoretical maximum attainable skill for week lead 4 is ~0.6 ACC and that future analysis can verify the actual skill based on ground-based observations (Fig. 4). Simulated RZSM could imply that GEFSv12-obs land surface initial conditions are of poorer quality, but we have identified that GEFSv12-obs RZSM initial state has low differences for many regions when compared to NLDAS-2 (Figs. S18–S23). NLDAS-2 is assumed to be the most accurate reanalysis over CONUS due to high spatial and temporal resolution, focused regional calibration, and bias-corrected precipitation forcings (Xia et al. 2012, 2014). Assuming NLDAS-2 to be the least biased reanalysis, GEFSv12-obs has less bias in the eastern United States during spring season when compared to MERRA-2, and when averaged over all warm seasons, biases between NLDAS-2/GEFSv12-obs and NLDAS-2/MERRA-2 are similar in the eastern United States (Fig. 3). But high biases exist between GEFSv12-obs and NLDAS-2 in the central United States, indicating that GEFSv12-obs and GEFSv12-fcst may be less credible in these regions. We have shown that GEFSv12-obs can potentially provide low-bias RZSM initial forcings into GEFSv12-fcst LSM, but the effect from forcing with nonbias-corrected precipitation and nonassimilated RZSM may impact long-lead forecast quality and interpretations of forecast integrity.

The GEFSv12-fcst Noah LSM has several limiting factors, including the inability to account for subgrid-scale heterogeneity, which is necessary to account for processes such as runoff, infiltration, and interception (Torres-Rojas et al. 2022; Gehne et al. 2019). Additionally, parameterizations within the LSM such as soil hydraulic conductivity, stomatal resistance, or vegetation fraction are fixed and not calibrated for individual regions which leads to the growth of biases and has impacts on energy and mass exchanges between land and atmosphere (Gehne et al. 2019). Improvements in parameterizations should increase predictability, but previous research which downscaled and bias-corrected SubX models and forced a Noah-multiparameterization model showed that drought onset along the U.S. West Coast for lead weeks 1–2 had usable drought onset skill, while lead weeks 3–4 had little to no skill regardless of drought severity (Su et al. 2023). They observed that precipitation skill for all models after downscaling was <0.3 ACC, which negatively impacted drought forecasts, and this implies that FD forecasts would also be hindered by inaccurate precipitation forecasts and that improving the LSM must be jointly improved with the AGCM.

Spatial resolution can impact the fidelity of reanalyses and subsequent forecasts by better resolving energy, mass, momentum, and carbon fluxes. GEFSv12-fcst is currently the highest-resolution SubX model for days 0–10, and the skill performance is noticeably higher than GMAO-V2p1 (which produces a 0.5° × 0.5° forecast for all lead times) for atmospheric variables, including z200, T_{mean}, T_{max}, wind speed, actual vapor pressure, shortwave radiation, and precipitation (Figs. S2, S3, and S5–S8). GEFSv12-fcst has also been identified to have higher probabilistic skill when compared with other SubX models for both precipitation and temperature across different global regions, indicating AGCM parameterizations are likely better resolved and allow for higher predictability (Robertson et al. 2023). It has been shown that models with 50-km horizontal spatial resolution produce more accurate precipitation forecasts and more realistic weeks 3–4 predictions of teleconnections when compared with 104-km horizontal resolution models (de Andrade et al. 2019), and additional researchers identified improvements from higher-resolution models in representing stratosphere and ocean processes, which are important for LSM forecasts (Seo et al. 2014; Roff et al. 2011; Roberts et al. 2022; Richter et al. 2020). Higher-resolution precipitation estimates increase LSM accuracy through more realistic land surface storage fluxes, and stratospheric circulation anomalies can influence tropospheric precipitation variability across CONUS through chemical–radiative–dynamical feedbacks. Stratosphere–atmosphere coupling in GEFSv12-fcst has been identified to agree well with the ECMWF ERA5 reanalysis with enhanced MJO and North Atlantic Oscillation predictive skill, which impact CONUS precipitation variability (Lawrence et al. 2023). The importance of stratosphere–troposphere feedbacks has been identified in the midlatitudes with spring stratospheric ozone

anomalies disrupting the atmospheric radiation balance and cause poleward shifts of the midlatitude jet (Ma et al. 2019; Smith and Polvani 2014). Therefore, a better resolved stratosphere can enhance precipitation forecasts and potentially improve predictions of FD onset related to precipitation deficits.

Atmospheric initial conditions are most important for short-term subseasonal forecasts, but air–ocean thermodynamic feedbacks are important for extended multiweek forecasts (Li et al. 2001; Ling et al. 2015). GEFSv12-fcst only has a coupled atmosphere and land model with the ocean boundary conditions derived from a two-tiered sea surface temperature and near sea surface temperature approach (Zhu et al. 2018, 2017; Guan et al. 2022). This limits predictability because of the impact ocean anomalies have on precipitation predictability primarily during El Niño–Southern Oscillation (ENSO) and MJO phases (Sun et al. 2022; Tian et al. 2017; Li et al. 2021; Vigaud et al. 2017). Ocean models with increasing resolutions are better able to simulate subsurface heat fluxes and have increased precipitation predictability along with improved MJO simulations stressing the importance of air–ocean feedbacks (Seo et al. 2014). The next GEFSv12-fcst update is scheduled to include a fully coupled atmosphere–ocean–wave–ice model, and improvements in RZSM and FD forecast skill can be assessed when this becomes available.

b. *Errors influencing RZSM forecast skill*

For predicting RZSM within GEFSv12-fcst Noah LSM, the three primary forcings include temperature, radiation, and precipitation, and their accumulated errors jointly affect RZSM forecasts. Except for radiation in the fall season and temperature during summer (Figs. S2 and S7), no primary forcings show substantial predictability at longer leads which suggests weeks 3–4 RZSM errors are driven by inadequate representations of all primary forcings with precipitation introducing the largest errors (Fig. S8). When GEFSv12-fcst is referenced against all reanalyses, it has higher precipitation skill than GMAO-V2p1 for summer and fall seasons for most regions, indicating that short-lead LSM predictions in these regions could also be more skillful and GEFSv12-fcst could be a resource for water resource managers. However, we do observe that RZSM skill between GEFSv12-obs and GEFSv12-fcst at week 1 has the same skill during spring and fall seasons, yet precipitation skill has large differences (Fig. 4; Fig. S8). In spring, precipitation ACC is ~0.3–0.4, while RZSM ACC is ~0.8–0.9, but in fall, precipitation ACC is ~0.50–0.65, while RZSM ACC is still ~0.8–0.9. This may indicate the relative importance of initialized RZSM state and persistence of RZSM anomalies, and the effect of low-quality precipitation forecasts does not begin to degrade RZSM forecasts until after week 1. Although RZSM skill plateaus at ~0.5 ACC for lead weeks 3–4, this skill is significantly better than precipitation (<0.3 ACC), indicating the added value of using RZSM predictions for FD forecasts rather than using precipitation. Solar radiation provides energy inputs for land, atmosphere, and ocean processes (Kalogirou 2004), while cloud cover and feedback processes which directly impact surface radiation are difficult to estimate in models, which impacts RZSM and precipitation predictability (Klein and Hall 2015). Temperature also suffers from a lack of predictability in the subseasonal time scale with research improving short-lead predictions over the climatological baseline, but long-lead deterministic forecasts are still highly unpredictable (He et al. 2021). Therefore, improvements to LSMs must identify ways to reduce errors from all three sources because of their impact on long-term predictability.

Another error-inducing mechanism is RZSM initialization error, which has been shown to impact predictions of seasonal soil moisture and streamflow (Koster et al. 2014) and subseasonal temperature and precipitation anomalies associated with the 2012 Central U.S. FD (DeAngelis et al. 2020; Liang and Yuan 2021). Initialization biases can stem from interpolation issues or observation error, and if initialization errors are large enough and combined with errors from primary drivers, skillful long-lead forecasts become difficult to achieve (Ni-Meister et al. 2005). It has been identified that under near-perfect initial conditions using high-resolution models, midlatitude weather has intrinsic predictability limits between 10 and 15 days for synoptic precipitation and temperature systems due to the chaotic nature of the atmosphere, which cannot be improved upon, which may limit skill SubX RZSM forecasts regardless of initial condition (Zhang et al. 2019).

c. *Flash drought predictability*

RZSM was used as our primary FD indicator because of its success in classifying major events using the SMPD framework (Lesinger and Tian 2022; Osman et al. 2021). The SMPD index has low-to-medium skill in predicting rapid changes in RZSM for the 2012 central United States and portions of the 2017 northern High Plains FD but no skill for 2019 Southeast FD (Figs. 9–11). Larger ensemble size models have been shown to improve probabilistic prediction (Tippett and Barnston 2008; Richardson 2001; Smith et al. 2015), which should be considered for future research on probabilistic FD forecasts. Although GEFSv12-fcst successfully classified FD in some CONUS areas, the TPR, FPR, and CSI rates do not support confidence in relying explicitly on GEFSv12-fcst for issuing FD early warning alerts. While the CSI can be a useful evaluation criterion for early warning, there is currently no threshold which indicates the trustworthiness of forecasts, and it has been noted that for winter storm warnings, the FPR has the greatest impact on CSI skill and that optimization of the CSI does not necessarily lead to more effective warnings (Gerapetritis and Pelissier 2004). A comparison study of metrics may be needed to evaluate FD prediction skill as it pertains to FD intensity and severity. Accurate FD onset predictions can allow regions ample mitigation time which can reduce cost-related impacts (Shreve and Kelman 2014), but conversely, false positive forecasts can incur unneeded costs and reduce user trust on future forecasts (Burgeno and Joslyn 2020; de Elía 2022).

RZSM initial conditions were shown to be important for predicting weeks 3–4 temperature and precipitation anomalies, and it has been identified that less-biased RZSM forcings reduce temperature and precipitation biases (DeAngelis et al. 2020). DeAngelis et al. (2020) identified the increase in skill was largely dependent on initial dry soil conditions, which amplified

land–atmosphere feedbacks and were better captured in the model. From a bottom-up perspective, improved local-scale modeling has immediate consequences on reducing mesoscale and synoptic errors, and the reduction of errors can improve model representation of global teleconnections such as Rossby wave trains. Therefore, it can be inferred that RZSM and associated land–atmosphere feedbacks influence teleconnection skill and FD predictability. Although GEFSv12-fcst dry anomalies were most predictable (Fig. 7), it is difficult to estimate the impact of initialized RZSM on forecast quality due to GEFSv12-obs forcing having considerable anomaly deviations when compared to NLDAS-2 (Fig. 3). RZSM is known to have high autocorrelation (persistence), which can impart predictability from initial conditions, but this persistence varies by region and season (Jacobs et al. 2020; Kumar et al. 2019; Koster and Suarez 2001). We identified that ACC skill of deeper soil layers is higher than shallow layers (Fig. S24), and this is related to longer soil moisture memory as identified by Kumar et al. (2019) who also found that deeper layer soil moisture anomalies can propagate into shallower layers and can influence drought on interannual time scales. Research has not fully explored CONUS FD onset and its relation to lagged anomalies in individual soil layers, and this may provide an additional source of FD predictability.

It is important to note that the 2012 and 2019 CONUS FDs had preexisting dry conditions present before FD onset, but FD intensification was amplified by propagating Rossby waves, but it is not known if both events had similar source regions for wave initiation. The 2019 Southeast FD had wave initiations from anomalous SSTs in the Indian Ocean, which increased deep convection and forced anticyclonic flow over the Southeast (Schubert et al. 2021), while the 2012 central U.S. FD wave origins have not yet been identified. We observed that z200 anomalies were not well forecasted for the 2019 Southeast drought (Fig. 11), while the 2017 northern High Plains z200 anomalies were better predicted (Fig. 10). Future research can identify the origination source of Rossby waves during other notable FD events, wave propagating features, and characteristics of the forecast models, which enable accurate predictions (e.g., North Pacific jet stream representation). It should be noted that the USDM did classify the Northwest as being abnormally dry during July 2017 (Fig. S25), which is in agreement with GEFSv12-fcst predictions of decreased RZSM (Fig. 11), but moderate drought was never officially classified. Summer in the Northwest is climatologically the driest when compared to all seasons, and dry conditions during this time period may be normal and have lesser environmental impacts (Guirguis and Avissar 2008). However, under future warming scenarios, rapid drying associated with FD is projected to become more frequent which may place these regions at higher risk (Christian et al. 2023). These findings indicate that the current SMPD framework may not be universally suitable during all seasons because of regional climatologies and the projected impacts, suggesting a potential need for localized FD classifications.

d. *Skill of ET_o and comparison with RZSM*

We identified that atmospheric variable skill was moderately similar between SubX models against the three reanalyses with summertime z200 and temperature having the highest ACC values at leads 3–4 (~0.25–0.5) (Figs. S1 and S2). These values lie below the ECMWF recommended 0.5 ACC threshold in which prediction errors are empirically similar to forecast errors based on the climatological average; thus, predictability is still minimal. CONUS is differentially impacted by ET_o drivers with research identifying regional drivers of ET_o being temperature in the eastern United States and along the West Coast, humidity in the western and Southeast United States, radiation in the Southeast, and wind speed in the Southwest (Hobbins 2016). Although FD is not explicitly represented by ET_o, poor representation at leads 3–4 may hinder the fidelity of early warning forecasts, which could identify regions susceptible to FD. We show that GEFSv12-fcst generally has the highest model skill during summer and fall for all regions and GMAO-V2p1 has higher model skill during spring. Additional increases in skill are obtained by taking the average of weeks 3–4 anomalies, and these forecasts can be useful for agriculture or water resource managers whom are reliant on biweekly accurate predictions (Fig. 2). Abatzoglou et al. (2023) identified that subseasonal ET_o forecast skill varied between April and October, and we find strong agreement with our results for week 3 regarding higher skill in the South, central, and western United States (Fig. S9). Seasonal ET_o forecast skill is known to be increased when forecasts are initialized during moderate and strong ENSO events, indicating some portion of predictability stems from initial state of tropical Pacific SSTs (McEvoy et al. 2016). Future research can explore the increase in model skill during initialized ENSO phases as well as other modes of variability, including the MJO, North Atlantic Oscillation, or Pacific–North American pattern, which can influence CONUS weather patterns through atmospheric blocking (Lupo 2021).

RZSM generally shows higher skill than ET_o. It is worth noting that RZSM influences energy and mass exchanges between the land and atmosphere, which may influence ET_o, while a positive relationship between RZSM and ET_o skill does not appear consistently across all regions. For weeks 3–4 during fall, higher RZSM skill is accompanied by higher ET_o skill for Midwest, South, and Southeast regions, and this same pattern emerges during summer for the Midwest and High Plains (Figs. 2 and 3). Regional variations are likely the result of climatological relationships between RZSM and ET. Koster (2004a) identified that stronger coupling between RZSM and precipitation occurs in summer in the central United States. The central United States is in a transitional climate regime between semiarid and humid conditions, and seasonal variations based on RZSM mean state can influence ET and ET_o. More humid regions like the Southeast have energy-limited regimes most seasons, and ET_o and RZSM coupling is typically of a lower magnitude, and it has been previously identified that fall season FDs are more common in the Southeast during fall and associated with La Niña (Brown and Degaetano 2013; Lesinger and Tian 2022). The

stronger skill association between ET_o and RZSM during fall in the Southeast and summer in both the Midwest and High Plains is jointly accompanied with increased FD predictability, which may indicate the source of these models' better predictions (Fig. 8). The desiccation of soils during summer and land–atmosphere feedbacks may serve to increase ET_o in the Southeast in the fall season, and this is better captured by models when compared to other seasons. Given the joint effects of RZSM and ET_o on crop yields (Rigden et al. 2020), future research could develop a joint index combining RZSM and ET_o for FD monitoring and prediction.

e. *Predictability by seasons as influenced by teleconnections*

For accurate hydroclimate predictions at the subseasonal scale, the representation of global teleconnections is necessary because predictability imparted from initial conditions diminishes at longer leads. Specific regions which generate teleconnections influencing CONUS climate include the Indian, Pacific, and Atlantic Oceans as well as the Arctic stratospheric polar vortex. In general, winter surface and circulation anomalies have higher predictability imparted by sudden stratospheric warmings (SSWs) causing deviations in normal stratospheric polar vortex (SPV) state impacting jet stream motion. SSWs can be promoted equally by El Niño and La Niña phases which introduce Rossby waves into the troposphere, breaking down the SPV (Butler and Polvani 2011). Additional winter skill within the troposphere is obtained during MJO and ENSO events, which strongly modulate CONUS weather (Tripathi et al. 2015; Wang et al. 2020; Toniazzo and Scaife 2006). Summer predictability is inhibited by lower amplitude MJO and ENSO events and a weakened SPV, which is less influenced by the troposphere. We observe that for z200 and temperature, summer is slightly more predictable than spring or fall for weeks 3–4 forecasts, but SubX model variability is higher. The higher predictability in FD onset in summer and fall in the western United States may be due to this region being climatologically dry with low soil moisture variability with ENSO impacting FD severity during fall (Guirguis and Avissar 2008; Lesinger and Tian 2022), and for the Southeast, the fall highest predictability may be associated with ENSO, Pacific decadal oscillation, Western Hemisphere warm pool, or the Southern Oscillation index (Sehgal and Sridhar 2018). Seasonal spring forecasts are noted to be inhibited by a "spring predictability barrier" in which forecasts initialized during spring produce less skillful results for future seasons—particularly for predicting ENSO, which is a leading mode of interannual variability across the globe (Jin et al. 2022; Duan and Wei 2013; Bushuk et al. 2022). The jet stream transitions rapidly in between winter and spring; this process may not be well captured within models, but under conditions where ENSO state and North Pacific jet structure is adequately resolved in models, then springtime mass transport processes are improved (Breeden et al. 2021). The spring predictability barrier has been noted to be more prevalent during the growth phases of El Niño (Duan and Wei 2013), and our observations across different variables and seasons may imply that the spring predictability barrier could be a relevant issue within subseasonal forecasts and can be caused by jet stream shifts that may be influenced by ENSO. Better representation of ocean–troposphere–stratosphere interactions will be needed to further enhance FD predictability. It has been identified that during summer, when El Niño is transitioning to La Niña, a warming signal occurs over the central United States and is related to suppressed convection in the Pacific Ocean and development of propagating Rossby waves (Anderson et al. 2017; Jong et al. 2020). These signals may be useful indicators of future FD likelihood and can be explored in future work as researchers continue to locate windows of opportunity for enhanced predictive skill.

Previous research has identified that for most CONUS regions, FD occurs more often during Atlantic multidecadal oscillation positive phase and during La Niña due to increased Atlantic convection, which provides drier air to the Pacific Ocean through background easterly winds and increasing the Pacific Walker circulation (Lesinger and Tian 2022). Other oscillations which slightly increase FD frequency include the North Atlantic Oscillation, Arctic Oscillation, and Pacific decadal oscillation. One teleconnection not explored for FD predictability is the East Asian monsoon which is known to cause alterations in the North Pacific jet stream and increase CONUS rainfall variability by promoting anticyclonic blocking over regions such as the U.S. Great Plains (Lopez et al. 2019; Zhu and Li 2018). These modes of variability are not comprehensive, and additional circulations could exist across the globe and should be further explored to improve FD predictability. Our assessment only considers two SubX models (GMAO-V2p1 and GEFSv12-fcst), mainly for deterministic forecast skill. There are a total of eight SubX models available that generate large ensemble members. This information could be explored in the future for probabilistic FD forecasts.

5. Conclusions

In this study, we comprehensively analyzed the subseasonal forecast skill of RZSM, ET_o, and flash drought onset, as well as the primary forcing variables in two SubX models. Results indicate that drier RZSM anomalies have the highest skill, while the predictability can be attributed to persistence from initial conditions, which does not always lead to accurate prediction of rapid decreases in RZSM. RZSM accuracy is inhibited by poor predictions of temperature, radiation, and precipitation which drive land surface model fluxes with temperature being the most skillful variable followed by radiation and precipitation. The large difference between RZSM skill against three reanalyses were more pronounced than atmospheric forcing variables, highlighting the impact of verification dataset on perceived RZSM skill. GEFSv12 reanalysis shows high biases in the central United States when compared to MERRA-2 and NLDAS-2, but biases are smaller in the eastern United States. The evaluation of GEFSv12 reforecast against the GEFSv12 reanalyses revealed its prediction limit, which can achieve ACC of RZSM forecasts as high as 0.6 for lead weeks 3–4. ET_o is another indicator of flash drought with skill generally lower compared to RZSM forecasts due to

unskillful forecasts of radiation, temperature, wind speed, and vapor pressure. RZSM forecast skill often remains higher than ET_o through week 3 or 4, given the longer memory of soil moisture, suggesting that it could be a better proxy used to predict flash droughts, rather than ET_o. In the western United States, during summer and spring, ET_o has higher skill than RZSM and can be useful as a regional flash drought early warning indicator. Case studies of major United States FDs indicate their onset is rarely predicted at weeks 3–4, which is mainly due to limited skill of precipitation, 200-hPa geopotential height, and ET_o at longer leads. To further improve flash drought forecasts, future research must jointly improve skill of temperature, precipitation, and radiation to provide better land surface model forcings. Accurate flash drought forecasts allow for early warning alerts from operational agencies to effectively mitigate drought impacts. While subseasonal forecasts have potential to inform users of impending extreme climate events, more research is needed to improve dynamical forecast systems by reducing model biases and improving forecast initializations, which currently impede accurate flash drought forecasts, and better understand predictability limit of flash drought events.

Acknowledgments. Research support is provided by the NSF Research Traineeship Program (DGE-1922687), the NSF CAREER Award (EAR-2144293), and the Hatch program of the USDA National Institute of Food and Agriculture (NIFA) (Accession 1012578). We would like to acknowledge high-performance computing support from Cheyenne (https://doi.org/10.5065/D6RX99HX) provided by the NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. Additionally, this work was completed in part with resources provided by the Auburn University Easley Cluster.

Data availability statement. All the data used in this study are freely available from the referenced sources. The Python scripts that were used for the analyses are available from the authors upon request.

REFERENCES

Abadi, A. M., Y. Gwon, M. O. Gribble, J. D. Berman, R. Bilotta, M. Hobbins, and J. E. Bell, 2022: Drought and all-cause mortality in Nebraska from 1980 to 2014: Time-series analyses by age, sex, race, urbanicity and drought severity. *Sci. Total Environ.*, **840**, 156660, https://doi.org/10.1016/j.scitotenv.2022.156660.

Abatzoglou, J. T., D. J. McEvoy, N. J. Nauslar, K. C. Hegewisch, and J. L. Huntington, 2023: Downscaled subseasonal fire danger forecast skill across the contiguous United States. *Atmos. Sci. Lett.*, **24**, e1165, https://doi.org/10.1002/asl.1165.

Allen, R. G., L. S. Pereira, D. Raes, and M. Smith, 1998: Crop evapotranspiration: Guidelines for computing crop water requirements. FAO Irrigation and Drainage Paper 56, 300 pp., www.fao.org/docrep/X0490E/X0490E00.htm.

Anderson, W., R. Seager, W. Baethgen, and M. Cane, 2017: Life cycles of agriculturally relevant ENSO teleconnections in North and South America. *Int. J. Climatol.*, **37**, 3297–3318, https://doi.org/10.1002/joc.4916.

Angélil, O., and Coauthors, 2016: Comparing regional precipitation and temperature extremes in climate model and reanalysis products. *Wea. Climate Extremes*, **13**, 35–43, https://doi.org/10.1016/j.wace.2016.07.001.

Arakawa, A., 2004: The cumulus parameterization problem: Past, present, and future. *J. Climate*, **17**, 2493–2525, https://doi.org/10.1175/1520-0442(2004)017<2493:RATCPP>2.0.CO;2.

Ashfaqur Rahman, M., M. Almazroui, M. Nazrul Islam, E. O'Brien, and A. E. Yousef, 2018: The role of land surface fluxes in Saudi-KAU AGCM: Temperature climatology over the Arabian Peninsula for the period 1981–2010. *Atmos. Res.*, **200**, 139–152, https://doi.org/10.1016/j.atmosres.2017.10.011.

Banerjee, O., R. Bark, J. Connor, and N. D. Crossman, 2013: An ecosystem services approach to estimating economic losses associated with drought. *Ecol. Econ.*, **91**, 19–27, https://doi.org/10.1016/j.ecolecon.2013.03.022.

Basara, J. B., J. I. Christian, R. A. Wakefield, J. A. Otkin, E. H. Hunt, and D. P. Brown, 2019: The evolution, propagation, and spread of flash drought in the central United States during 2012. *Environ. Res. Lett.*, **14**, 084025, https://doi.org/10.1088/1748-9326/ab2cc0.

Berg, A., and Coauthors, 2016: Land–atmosphere feedbacks amplify aridity increase over land under global warming. *Nat. Climate Change*, **6**, 869–874, https://doi.org/10.1038/nclimate3029.

Breeden, M. L., A. H. Butler, J. R. Albers, M. Sprenger, and A. O'Neil Langford, 2021: The spring transition of the North Pacific jet and its relation to deep stratosphere-to-troposphere mass transport over western North America. *Atmos. Chem. Phys.*, **21**, 2781–2794, https://doi.org/10.5194/acp-21-2781-2021.

Brown, P. J., and A. T. Degaetano, 2013: Trends in U.S. surface humidity, 1930–2010. *J. Appl. Meteor. Climatol.*, **52**, 147–163, https://doi.org/10.1175/JAMC-D-12-035.1.

Brubaker, K. L., D. Entekhabi, and P. S. Eagleson, 1993: Estimation of continental precipitation recycling. *J. Climate*, **6**, 1077–1089, https://doi.org/10.1175/1520-0442(1993)006<1077:EOCPR>2.0.CO;2.

Burgeno, J. N., and S. L. Joslyn, 2020: The impact of weather forecast inconsistency on user trust. *Wea. Climate Soc.*, **12**, 679–694, https://doi.org/10.1175/WCAS-D-19-0074.1.

Bushuk, M., and Coauthors, 2022: Mechanisms of regional Arctic sea ice predictability in two dynamical seasonal forecast systems. *J. Climate*, **35**, 4207–4231, https://doi.org/10.1175/JCLI-D-21-0544.1.

Butler, A. H., and L. M. Polvani, 2011: El Niño, La Niña, and stratospheric sudden warmings: A reevaluation in light of the observational record. *Geophys. Res. Lett.*, **38**, L13807, https://doi.org/10.1029/2011GL048084.

Cao, Q., S. Shukla, M. J. Deflorio, F. M. Ralph, and D. P. Lettenmaier, 2021: Evaluation of the subseasonal forecast skill of floods associated with atmospheric rivers in coastal western U.S. watersheds. *J. Hydrometeor.*, **22**, 1535–1552, https://doi.org/10.1175/JHM-D-20-0219.1.

Christian, J. I., J. B. Basara, J. A. Otkin, and E. D. Hunt, 2019: Regional characteristics of flash droughts across the United States. *Environ. Res. Commun.*, **1**, 125004, https://doi.org/10.1088/2515-7620/ab50ca.

——, ——, E. D. Hunt, J. A. Otkin, and X. Xiao, 2020: Flash drought development and cascading impacts associated with the 2010 Russian heatwave. *Environ. Res. Lett.*, **15**, 094078, https://doi.org/10.1088/1748-9326/ab9faf.

——, ——, L. E. L. Lowman, X. Xiao, D. Mesheske, and Y. Zhou, 2022: Flash drought identification from satellite-based land surface water index. *Remote Sens. Appl.*, **26**, 100770, https://doi.org/10.1016/j.rsase.2022.100770.

——, and Coauthors, 2023: Global projections of flash drought show increased risk in a warming climate. *Commun. Earth Environ.*, **4**, 165, https://doi.org/10.1038/s43247-023-00826-1.

de Andrade, F. M., C. A. S. Coelho, and I. F. A. Cavalcanti, 2019: Global precipitation hindcast quality assessment of the Subseasonal to Seasonal (S2S) prediction project models. *Climate Dyn.*, **52**, 5451–5475, https://doi.org/10.1007/s00382-018-4457-z.

DeAngelis, A. M., H. Wang, R. D. Koster, S. D. Schubert, Y. Chang, and J. Marshak, 2020: Prediction skill of the 2012 U.S. Great Plains flash drought in Subseasonal Experiment (SubX) models. *J. Climate*, **33**, 6229–6253, https://doi.org/10.1175/JCLI-D-19-0863.1.

de Elía, R., 2022: The false alarm/surprise trade-off in weather warnings systems: An expected utility theory perspective. *Environ. Syst. Decis.*, **42**, 450–461, https://doi.org/10.1007/s10669-022-09863-1.

Dirmeyer, P. A., X. Gao, M. Zhao, Z. Guo, T. Oki, and N. Hanasaki, 2006: GSWP-2: Multimodel analysis and implications for our perception of the land surface. *Bull. Amer. Meteor. Soc.*, **87**, 1381–1398, https://doi.org/10.1175/BAMS-87-10-1381.

——, and Coauthors, 2016: Confronting weather and climate models with observational data from soil moisture networks over the United States. *J. Hydrometeor.*, **17**, 1049–1067, https://doi.org/10.1175/JHM-D-15-0196.1.

——, 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.

Du, D., A. C. Subramanian, W. Han, H.-H. Wei, B. B. Sarojini, M. Balmaseda, and F. Vitart, 2023: Assessing the impact of ocean in situ observations on MJO propagation across the Maritime Continent in ECMWF subseasonal forecasts. *J. Adv. Model. Earth Syst.*, **15**, e2022MS003044, https://doi.org/10.1029/2022MS003044.

Duan, W., and C. Wei, 2013: The 'spring predictability barrier' for ENSO predictions and its possible mechanism: Results from a fully coupled model. *Int. J. Climatol.*, **33**, 1280–1292, https://doi.org/10.1002/joc.3513.

Dutra, E., F. Johannsen, and L. Magnusson, 2021: Late spring and summer subseasonal forecasts in the Northern Hemisphere midlatitudes: Biases and skill in the ECMWF model. *Mon. Wea. Rev.*, **149**, 2659–2671, https://doi.org/10.1175/MWR-D-20-0342.1.

Edris, S. G., J. B. Basara, J. I. Christian, E. D. Hunt, J. A. Otkin, S. T. Salesky, and B. G. Illston, 2023: Analysis of the critical components of flash drought using the standardized evaporative stress ratio. *Agric. For. Meteor.*, **330**, 109288, https://doi.org/10.1016/j.agrformet.2022.109288.

Fan, J., B. McConkey, H. Wang, and H. Janzen, 2016: Root distribution by depth for temperate agricultural crops. *Field Crops Res.*, **189**, 68–74, https://doi.org/10.1016/j.fcr.2016.02.013.

Findell, K. L., P. Gentine, B. R. Lintner, and C. Kerr, 2011: Probability of afternoon precipitation in eastern United States and Mexico enhanced by high evaporation. *Nat. Geosci.*, **4**, 434–439, https://doi.org/10.1038/ngeo1174.

Fisher, R. A., and C. D. Koven, 2020: Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. *J. Adv. Model. Earth Syst.*, **12**, e2018MS001453, https://doi.org/10.1029/2018MS001453.

Ford, T. W., and C. F. Labosier, 2017: Meteorological conditions associated with the onset of flash drought in the eastern United States. *Agric. For. Meteor.*, **247**, 414–423, https://doi.org/10.1016/j.agrformet.2017.08.031.

Gavande, S. A., and S. A. Taylor, 1967: Influence of soil water potential and atmospheric evaporative demand on transpiration and the energy status of water in plants. *Agron. J.*, **59**, 4–7, https://doi.org/10.2134/agronj1967.00021962005900010002x.

Gehne, M., T. M. Hamill, G. T. Bates, P. Pegion, and W. Kolczynski, 2019: Land surface parameter and state perturbations in the global ensemble forecast system. *Mon. Wea. Rev.*, **147**, 1319–1340, https://doi.org/10.1175/MWR-D-18-0057.1.

Gelaro, R., and Coauthors, 2017: The Modern-Era Retrospective Analysis for Research and Applications, version 2 (MERRA-2). *J. Climate*, **30**, 5419–5454, https://doi.org/10.1175/JCLI-D-16-0758.1.

Gerapetritis, H., and J. M. Pelissier, 2004: On the behavior of the critical success index. Eastern Region Tech. Attachment 2004-03, 6 pp., https://www.weather.gov/media/erh/ta2004-03.pdf.

Gerken, T., G. T. Bromley, B. L. Ruddell, S. Williams, and P. C. Stoy, 2018: Convective suppression before and during the United States Northern Great Plains flash drought of 2017. *Hydrol. Earth Syst. Sci.*, **22**, 4155–4163, https://doi.org/10.5194/hess-22-4155-2018.

Guan, H., and Coauthors, 2022: GEFSv12 reforecast dataset for supporting subseasonal and hydrometeorological applications. *Mon. Wea. Rev.*, **150**, 647–665, https://doi.org/10.1175/MWR-D-21-0245.1.

Guirguis, K. J., and R. Avissar, 2008: A precipitation climatology and dataset intercomparison for the western United States. *J. Hydrometeor.*, **9**, 825–841, https://doi.org/10.1175/2008JHM832.1.

Guo, Z., and P. A. Dirmeyer, 2013: Interannual variability of land–atmosphere coupling strength. *J. Hydrometeor.*, **14**, 1636–1646, https://doi.org/10.1175/JHM-D-12-0171.1.

Hamill, T. M., and Coauthors, 2022: The reanalysis for the Global Ensemble Forecast System, version 12. *Mon. Wea. Rev.*, **150**, 59–79, https://doi.org/10.1175/MWR-D-21-0023.1.

Hao, Z., V. P. Singh, and Y. Xia, 2018: Seasonal drought prediction: Advances, challenges, and future prospects. *Rev. Geophys.*, **56**, 108–141, https://doi.org/10.1002/2016RG000549.

He, S., X. Li, T. DelSole, P. Ravikumar, and A. Banerjee, 2021: Sub-seasonal climate forecasting via machine learning: Challenges, analysis, and advances. *Proc. 35th AAAI Conf. on Artificial Intelligence*, Online, AAAI Press, 169–177, https://doi.org/10.1609/aaai.v35i1.16090.

Hobbins, M. T., 2016: The variability of ASCE standardized reference evapotranspiration: A rigorous, CONUS-wide decomposition and attribution. *Trans. ASABE*, **59**, 561–576, https://doi.org/10.13031/trans.59.10975.

——, A. Wood, D. J. McEvoy, J. L. Huntington, C. Morton, M. Anderson, and C. Hain, 2016: The evaporative Demand Drought Index. Part I: Linking drought evolution to variations in evaporative demand. *J. Hydrometeor.*, **17**, 1745–1761, https://doi.org/10.1175/JHM-D-15-0121.1.

Hobeichi, S., G. Abramowitz, J. P. Evans, and A. Ukkola, 2022: Toward a robust, impact-based, predictive drought metric. *Water Resour. Res.*, **58**, e2021WR031829, https://doi.org/10.1029/2021WR031829.

Hoell, A., J. Perlwitz, and J. K. Eischeid, 2019: The causes, predictability, and historical context of the 2017 U.S. Northern Great Plains Drought. Drought Assessment Tech. Rep., 27 pp., https://repository.library.noaa.gov/view/noaa/23003.

Hoerling, M., J. Eischeid, A. Kumar, R. Leung, A. Mariotti, K. Mo, S. Schubert, and R. Seager, 2014: Causes and predictability of the 2012 Great Plains drought. *Bull. Amer. Meteor. Soc.*, **95**, 269–282, https://doi.org/10.1175/BAMS-D-13-00055.1.

Jacobs, E. M., L. E. Bertassello, and P. S. C. Rao, 2020: Drivers of regional soil water storage memory and persistence. *Vadose Zone J.*, **19**, e20050, https://doi.org/10.1002/vzj2.20050.

Jin, C., X. Luo, X. Xiao, J. Dong, X. Li, J. Yang, and D. Zhao, 2019: The 2012 flash drought threatened US Midwest agroecosystems. *Chin. Geogr. Sci.*, **29**, 768–783, https://doi.org/10.1007/s11769-019-1066-7.

Jin, Y., Z. Liu, and W. Duan, 2022: The different relationships between the ENSO spring persistence barrier and predictability barrier. *J. Climate*, **35**, 6207–6218, https://doi.org/10.1175/JCLI-D-22-0013.1.

Jong, B.-T., M. Ting, R. Seager, and W. B. Anderson, 2020: ENSO teleconnections and impacts on U.S. summertime temperature during a multiyear La Niña life cycle. *J. Climate*, **33**, 6009–6024, https://doi.org/10.1175/JCLI-D-19-0701.1.

——, M. Newman, and A. Hoell, 2022: Subseasonal meteorological drought development over the central United States during spring. *J. Climate*, **35**, 2525–2547, https://doi.org/10.1175/JCLI-D-21-0435.1.

Kalogirou, S. A., 2004: Solar thermal collectors and applications. *Prog. Energy Combust. Sci.*, **30**, 231–295, https://doi.org/10.1016/j.pecs.2004.02.001.

Kim, H., M. A. Janiga, and K. Pegion, 2019: MJO propagation processes and mean biases in the SubX and S2S reforecasts. *J. Geophys. Res. Atmos.*, **124**, 9314–9331, https://doi.org/10.1029/2019JD031139.

Klein, S. A., and A. Hall, 2015: Emergent constraints for cloud feedbacks. *Curr. Climate Change Rep.*, **1**, 276–287, https://doi.org/10.1007/s40641-015-0027-1.

Kleist, D. T., and K. Ide, 2015: An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part II: 4DEnVar and hybrid variants. *Mon. Wea. Rev.*, **143**, 452–470, https://doi.org/10.1175/MWR-D-13-00350.1.

Koster, R. D., and M. J. Suarez, 2001: Soil moisture memory in climate models. *J. Hydrometeor.*, **2**, 558–570, https://doi.org/10.1175/1525-7541(2001)002<0558:SMMICM>2.0.CO;2.

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

——, and Coauthors, 2004b: Realistic initialization of land surface states: Impacts on subseasonal forecast skill. *J. Hydrometeor.*, **5**, 1049–1063, https://doi.org/10.1175/JHM-387.1.

——, and Coauthors, 2010: Contribution of land surface initialization to subseasonal forecast skill: First results from a multimodel experiment. *Geophys. Res. Lett.*, **37**, L02402, https://doi.org/10.1029/2009GL041677.

——, G. K. Walker, S. P. P. Mahanama, and R. H. Reichle, 2014: Soil moisture initialization error and subgrid variability of precipitation in seasonal streamflow forecasting. *J. Hydrometeor.*, **15**, 69–88, https://doi.org/10.1175/JHM-D-13-050.1.

——, S. D. Schubert, A. M. Deangelis, A. M. Molod, and S. P. Mahanama, 2020: Using a simple water balance framework to quantify the impact of soil moisture initialization on subseasonal evapotranspiration and air temperature forecasts. *J. Hydrometeor.*, **21**, 1705–1722, https://doi.org/10.1175/JHM-D-20-0007.1.

Kumar, S., M. Newman, Y. Wang, and B. Livneh, 2019: Potential reemergence of seasonal soil moisture anomalies in North America. *J. Climate*, **32**, 2707–2734, https://doi.org/10.1175/JCLI-D-18-0540.1.

Kunkel, K. E., D. R. Easterling, B. C. Stewart, T. K. Maycock, C. W. Avery, K. L. M. Lewis, and D. R. Reidmiller, 2018: *Impacts, Risks, and Adaptation in the United States: The Fourth National Climate Assessment*. Vol. 2. U.S. Global Change Research Program, 69–112 pp.

Lawrence, Z. D., D. Elsbury, A. H. Butler, J. Perlwitz, J. R. Albers, L. M. Ciasto, and E. Ray, 2023: Evaluation of processes related to stratosphere–troposphere coupling in GEFSv12 subseasonal hindcasts. *Mon. Wea. Rev.*, **151**, 1735–1755, https://doi.org/10.1175/MWR-D-22-0283.1.

Lesinger, K., and D. Tian, 2022: Trends, variability, and drivers of flash droughts in the contiguous United States. *Water Resour. Res.*, **58**, e2022WR032186, https://doi.org/10.1029/2022WR032186.

Li, W., R. Yu, H. Liu, and Y. Yu, 2001: Impacts of diurnal cycle of SST on the intraseasonal variation of surface heat flux over the western Pacific warm pool. *Adv. Atmos. Sci.*, **18**, 793–806, https://doi.org/10.1007/BF03403503.

Li, Y., D. I. Tian, and H. Medina, 2021: Multimodel subseasonal precipitation forecasts over the contiguous United States: Skill assessment and statistical postprocessing. *J. Hydrometeor.*, **22**, 2581–2600, https://doi.org/10.1175/JHM-D-21-0029.1.

Liang, M., and X. Yuan, 2021: Critical role of soil moisture memory in predicting the 2012 central United States flash drought. *Front. Earth Sci.*, **9**, 615969, https://doi.org/10.3389/feart.2021.615969.

Ling, T., M. Xu, X.-Z. Liang, J. X. L. Wang, and Y. Noh, 2015: A multilevel ocean mixed layer model resolving the diurnal cycle: Development and validation. *J. Adv. Model. Earth Syst.*, **7**, 1680–1692, https://doi.org/10.1002/2015MS000476.

Lopez, H., S. K. Lee, S. Dong, G. Goni, B. Kirtman, R. Atlas, and A. Kumar, 2019: East Asian monsoon as a modulator of U.S. Great Plains heat waves. *J. Geophys. Res. Atmos.*, **124**, 6342–6358, https://doi.org/10.1029/2018JD030151.

Lorenz, D. J., J. A. Otkin, B. Zaitchik, C. Hain, and M. C. Anderson, 2021: Predicting rapid changes in Evaporative Stress Index (ESI) and soil moisture anomalies over the continental United States. *J. Hydrometeor*, **22**, 3017–3036, https://doi.org/10.1175/JHM-D-20-0289.1.

Lupo, A. R., 2021: Atmospheric blocking events: A review. *Ann. N. Y. Acad. Sci.*, **1504**, 5–24, https://doi.org/10.1111/NYAS.14557.

Ma, X., F. Xie, J. Li, X. Zheng, W. Tian, R. Ding, C. Sun, and J. Zhang, 2019: Effects of Arctic stratospheric ozone changes on spring precipitation in the northwestern United States. *Atmos. Chem. Phys.*, **19**, 861–875, https://doi.org/10.5194/acp-19-861-2019.

Mahanama, S. P. P., and R. D. Koster, 2005: AGCM biases in evaporation regime: Impacts on soil moisture memory and land–atmosphere feedback. *J. Hydrometeor.*, **6**, 656–669, https://doi.org/10.1175/JHM446.1.

Mariotti, A., P. M. Ruti, and M. Rixen, 2018: Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. *npj Climate Atmos. Sci.*, **1**, 4, https://doi.org/10.1038/s41612-018-0014-z.

Materia, S., A. Borrelli, A. Bellucci, A. Alessandri, P. di Pietro, P. Athanasiadis, A. Navarra, and S. Gualdi, 2014: Impact of atmosphere and land surface initial conditions on seasonal forecasts of global surface temperature. *J Climate*, **27**, 9253–9271, https://doi.org/10.1175/JCLI-D-14-00163.1.

McEvoy, D. J., J. L. Huntington, M. T. Hobbins, A. Wood, C. Morton, M. Anderson, and C. Hain, 2016: The evaporative

demand drought index. Part II: CONUS-wide assessment against common drought indicators. *J. Hydrometeor.*, **17**, 1763–1779, https://doi.org/10.1175/JHM-D-15-0122.1.

——, D. W. Pierce, J. F. Kalansky, D. R. Cayan, and J. T. Abatzoglou, 2020: Projected changes in reference evapotranspiration in California and Nevada: Implications for drought and wildland fire danger. *Earth's Future*, **8**, e2020EF001736, https://doi.org/10.1029/2020EF001736.

Mishra, A. K., and V. P. Singh, 2010: A review of drought concepts. *J. Hydrol.*, **391**, 202–216, https://doi.org/10.1016/j.jhydrol.2010.07.012.

Mitchell, K. E., and Coauthors, 2004: The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *J. Geophys. Res.*, **109**, D07S90, https://doi.org/10.1029/2003JD003823.

Mo, K. C., and D. P. Lettenmaier, 2015: Heat wave flash droughts in decline. *Geophys. Res. Lett.*, **42**, 2823–2829, https://doi.org/10.1002/2015GL064018.

——, and ——, 2016: Precipitation deficit flash droughts over the United States. *J. Hydrometeor.*, **17**, 1169–1184, https://doi.org/10.1175/JHM-D-15-0158.1.

NCEI, 2020: U.S. billion-dollar weather and climate disasters. NCEI, accessed 4 May 2021, https://www.ncdc.noaa.gov/billions/.

Ni-Meister, W., J. P. Walker, and P. R. Houser, 2005: Soil moisture initialization for climate prediction: Characterization of model and observation errors. *J. Geophys. Res.*, **110**, D13111, https://doi.org/10.1029/2004JD005745.

Osman, M., B. F. Zaitchik, H. S. Badr, J. I. Christian, T. Tadesse, J. A. Otkin, and M. C. Anderson, 2021: Flash drought onset over the contiguous United States: Sensitivity of inventories and trends to quantitative definitions. *Hydrol. Earth Syst. Sci.*, **25**, 565–581, https://doi.org/10.5194/hess-25-565-2021.

——, and Coauthors, 2022: Diagnostic classification of flash drought events reveals distinct classes of forcings and impacts. *J. Hydrometeor.*, **23**, 275–289, https://doi.org/10.1175/JHM-D-21-0134.1.

Otkin, J. A., M. C. Anderson, C. Hain, and M. Svoboda, 2014: Examining the relationship between drought development and rapid changes in the evaporative stress index. *J. Hydrometeor.*, **15**, 938–956, https://doi.org/10.1175/JHM-D-13-0110.1.

——, Y. Zhong, D. Lorenz, M. C. Anderson, and C. Hain, 2018: Exploring seasonal and regional relationships between the Evaporative Stress Index and surface weather and soil moisture anomalies across the United States. *Hydrol. Earth Syst. Sci.*, **22**, 5373–5386, https://doi.org/10.5194/hess-22-5373-2018.

——, and Coauthors, 2022: Getting ahead of flash drought: From early warning to early action. *Bull. Amer. Meteor. Soc.*, **103**, E2188–E2202, https://doi.org/10.1175/BAMS-D-21-0288.1.

PaiMazumder, D., and J. M. Done, 2016: Potential predictability sources of the 2012 U.S. drought in observations and a regional model ensemble. *J. Geophys. Res. Atmos.*, **121**, 12 581–12 592, https://doi.org/10.1002/2016JD025322.

Park, C.-K., and J. Kam, 2023: Sub-Seasonal Experiment (SubX) Model-based assessment of the prediction skill of recent multi-year South Korea droughts. *Asia-Pac. J. Atmos. Sci.*, **59**, 69–82, https://doi.org/10.1007/s13143-022-00307-z.

Park, S., M. A. Alexander, and C. Deser, 2006: The impact of cloud radiative feedback, remote ENSO forcing, and entrainment on the persistence of North Pacific Sea surface temperature anomalies. *J. Climate*, **19**, 6243–6261, https://doi.org/10.1175/JCLI3957.1.

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.

Pendergrass, A. G., and Coauthors, 2020: Flash droughts present a new challenge for subseasonal-to-seasonal prediction. *Nat. Climate Change*, **10**, 191–199, https://doi.org/10.1038/s41558-020-0709-0.

Pinker, R. T., and Coauthors, 2003: Surface radiation budgets in support of the GEWEX Continental-Scale International Project (GCIP) and the GEWEX Americas Prediction Project (GAPP), including the North American Land Data Assimilation System (NLDAS) project. *J. Geophys. Res.*, **108**, 8844, https://doi.org/10.1029/2002JD003301.

Reichle, R. H., C. S. Draper, Q. Liu, M. Girotto, S. P. P. Mahanama, R. D. Koster, and G. J. M. De Lannoy, 2017: Assessment of MERRA-2 land surface hydrology estimates. *J. Climate*, **30**, 2937–2960, https://doi.org/10.1175/JCLI-D-16-0720.1.

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.*, **127**, 2473–2489, https://doi.org/10.1002/qj.49712757715.

Richter, J. H., and Coauthors, 2020: Subseasonal prediction with and without a well-represented stratosphere in CESM1. *Wea. Forecasting*, **35**, 2589–2602, https://doi.org/10.1175/WAF-D-20-0029.1.

Rigden, A. J., N. D. Mueller, N. M. Holbrook, N. Pillai, and P. Huybers, 2020: Combined influence of soil moisture and atmospheric evaporative demand is important for accurately predicting US maize yields. *Nat. Food*, **1**, 127–133, https://doi.org/10.1038/s43016-020-0028-7.

Roberts, C. D., M. A. Balmaseda, S. Tietsche, and F. Vitart, 2022: Sensitivity of ECMWF coupled forecasts to improved initialization of the ocean mesoscale. *Quart. J. Roy. Meteor. Soc.*, **148**, 3694–3714, https://doi.org/10.1002/qj.4383.

Robertson, A. W., and Coauthors, 2023: A multimodel real-time system for global probabilistic subseasonal forecasts of precipitation and temperature. *Wea. Forecasting*, **38**, 921–935, https://doi.org/10.1175/WAF-D-22-0160.1.

Roff, G., D. W. J. Thompson, and H. Hendon, 2011: Does increasing model stratospheric resolution improve extended-range forecast skill? *Geophys. Res. Lett.*, **38**, L05809, https://doi.org/10.1029/2010GL046515.

Santoalla, D., 2018: Anomaly correlation coefficient: Relative skill of IFS models. Accessed 29 December 2022, https://confluence.ecmwf.int/display/FUG/Anomaly+Correlation+Coefficient.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575, https://doi.org/10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.

Schubert, S. D., Y. Chang, A. M. Deangelis, H. Wang, and R. D. Koster, 2021: On the development and demise of the fall 2019 southeast U.S. flash drought: Links to an extreme positive IOD. *J. Climate*, **34**, 1701–1723, https://doi.org/10.1175/JCLI-D-20-0428.1.

Sehgal, V., and V. Sridhar, 2018: Effect of hydroclimatological teleconnections on the watershed-scale drought predictability in the southeastern United States. *Int. J. Climatol.*, **38**, e1139–e1157, https://doi.org/10.1002/joc.5439.

——, N. Gaur, and B. P. Mohanty, 2021: Global flash drought monitoring using surface soil moisture. *Water Resour. Res.*, **57**, e2021WR029901, https://doi.org/10.1029/2021WR029901.

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.

Seo, H., A. C. Subramanian, A. J. Miller, and N. R. Cavanaugh, 2014: Coupled impacts of the diurnal cycle of sea surface temperature on the Madden–Julian oscillation. *J. Climate*, **27**, 8422–8443, https://doi.org/10.1175/JCLI-D-14-00141.1.

Shreve, C. M., and I. Kelman, 2014: Does mitigation save? Reviewing cost-benefit analyses of disaster risk reduction. *Int. J. Disaster Risk Reduct.*, **10**, 213–235, https://doi.org/10.1016/j.ijdrr.2014.08.004.

Smith, K. L., and L. M. Polvani, 2014: The surface impacts of Arctic stratospheric ozone anomalies. *Environ. Res. Lett.*, **9**, 074015, https://doi.org/10.1088/1748-9326/9/7/074015.

Smith, L. A., H. Du, E. B. Suckling, and F. Niehörster, 2015: Probabilistic skill in ensemble seasonal forecasts. *Quart. J. Roy. Meteor. Soc.*, **141**, 1085–1100, https://doi.org/10.1002/qj.2403.

Su, L., Q. Cao, S. Shukla, M. Pan, and D. P. Lettenmaier, 2023: Evaluation of subseasonal drought forecast skill over the coastal western United States. *J. Hydrometeor.*, **24**, 709–726, https://doi.org/10.1175/JHM-D-22-0103.1.

Suhas, E., and G. J. Zhang, 2014: Evaluation of trigger functions for convective parameterization schemes using observations. *J. Climate*, **27**, 7647–7666, https://doi.org/10.1175/JCLI-D-13-00718.1.

Sun, L., M. P. Hoerling, J. H. Richter, A. Hoell, A. Kumar, and J. W. Hurrell, 2022: Attribution of North American subseasonal precipitation prediction skill. *Wea. Forecasting*, **37**, 2069–2085, https://doi.org/10.1175/WAF-D-22-0076.1.

Tian, D., E. F. Wood, and X. Yuan, 2017: CFSv2-based subseasonal precipitation and temperature forecast skill over the contiguous United States. *Hydrol. Earth Syst. Sci.*, **21**, 1477–1490, https://doi.org/10.5194/hess-21-1477-2017.

Tippett, M. K., and A. G. Barnston, 2008: Skill of multimodel ENSO probability forecasts. *Mon. Wea. Rev.*, **136**, 3933–3946, https://doi.org/10.1175/2008MWR2431.1.

Toniazzo, T., and A. A. Scaife, 2006: The influence of ENSO on winter North Atlantic climate. *Geophys. Res. Lett.*, **33**, L24704, https://doi.org/10.1029/2006GL027881.

Torres-Rojas, L., N. Vergopolan, J. D. Herman, and N. W. Chaney, 2022: Towards an optimal representation of sub-grid heterogeneity in land surface models. *Water Resour. Res.*, **58**, e2022WR032233, https://doi.org/10.1029/2022WR032233.

Tripathi, O. P., A. Charlton-Perez, M. Sigmond, and F. Vitart, 2015: Enhanced long-range forecast skill in boreal winter following stratospheric strong vortex conditions. *Environ. Res. Lett.*, **10**, 104007, https://doi.org/10.1088/1748-9326/10/10/104007.

Van Loon, A. F., and Coauthors, 2016: Drought in the anthropocene. *Nat. Geosci.*, **9**, 89–91, https://doi.org/10.1038/ngeo2646.

Vigaud, N., A. W. Robertson, and M. K. Tippett, 2017: Multimodel ensembling of subseasonal precipitation forecasts over North America. *Mon. Wea. Rev.*, **145**, 3913–3928, https://doi.org/10.1175/MWR-D-17-0092.1.

Vitart, F., and A. W. Robertson, 2018: The sub-seasonal to Seasonal Prediction Project (S2S) and the prediction of extreme events. *npj Climate Atmos. Sci.*, **1**, 3, https://doi.org/10.1038/s41612-018-0013-0.

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

Wang, D., Y. Kang, and S. Wan, 2007: Effect of soil matric potential on tomato yield and water use under drip irrigation condition. *Agric. Water Manage.*, **87**, 180–186, https://doi.org/10.1016/j.agwat.2006.06.021.

Wang, Z., T. Li, J. Gao, and M. Peng, 2020: Enhanced winter and summer trend difference of Madden–Julian oscillation intensity since 1871. *Int. J. Climatol.*, **40**, 6369–6381, https://doi.org/10.1002/joc.6586.

Wilhite, D. A., M. H. Glantz, and M. H. And Glantz, 1985: Understanding the drought phenomenon: The role of definitions. Drought Mitigation Center Faculty Publications, 17 pp., https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1019&context=droughtfacpub.

Wood, E. F., D. Lettenmaier, X. Liang, B. Nijssen, and S. W. Wetzel, 1997: Hydrological modeling of continental-scale basins. *Annu. Rev. Earth Planet. Sci.*, **25**, 279–300, https://doi.org/10.1146/annurev.earth.25.1.279.

WWRP, 2017: Forecast verification methods across time and space scales. *Seventh Int. Verification Methods Workshop*, Berlin, Germany, WWRP/WCRP, https://cawcr.gov.au/projects/verification/.

Xia, Y., and Coauthors, 2012: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *J. Geophys. Res.*, **117**, D03109, https://doi.org/10.1029/2011JD016048.

——, J. Sheffield, M. B. Ek, J. Dong, N. Chaney, H. Wei, J. Meng, and E. F. Wood, 2014: Evaluation of multi-model simulated soil moisture in NLDAS-2. *J. Hydrol.*, **512**, 107–125, https://doi.org/10.1016/j.jhydrol.2014.02.027.

Zhang, F., Y. Qiang Sun, L. Magnusson, R. Buizza, S.-J. Lin, J.-H. Chen, and K. Emanuel, 2019: What is the predictability limit of midlatitude weather? *J. Atmos. Sci.*, **76**, 1077–1091, https://doi.org/10.1175/JAS-D-18-0269.1.

Zhang, J., and Coauthors, 2021: Sustainable irrigation based on co-regulation of soil water supply and atmospheric evaporative demand. *Nat. Commun.*, **12**, 5549, https://doi.org/10.1038/s41467-021-25254-7.

Zheng, C., E. K.-M. Chang, H. Kim, M. Zhang, and W. Wang, 2021: Subseasonal prediction of wintertime Northern Hemisphere extratropical cyclone activity by SubX and S2S models. *Wea. Forecasting*, **36**, 75–89, https://doi.org/10.1175/WAF-D-20-0157.1.

Zhou, S., and Coauthors, 2019: Land–atmosphere feedbacks exacerbate concurrent soil drought and atmospheric aridity. *Proc. Natl. Acad. Sci. USA*, **116**, 18 848–18 853, https://doi.org/10.1073/pnas.1904955116.

Zhu, Y., X. Zhou, M. Peña, W. Li, C. Melhauser, and D. Hou, 2017: Impact of sea surface temperature forcing on weeks 3 and 4 forecast skill in the NCEP global ensemble forecasting system. *Wea. Forecasting*, **32**, 2159–2174, https://doi.org/10.1175/WAF-D-17-0093.1.

——, and Coauthors, 2018: Toward the improvement of subseasonal prediction in the national centers for environmental prediction global ensemble forecast system. *J. Geophys. Res. Atmos.*, **123**, 6732–6745, https://doi.org/10.1029/2018JD028506.

Zhu, Z., and T. Li, 2018: Amplified contiguous United States summer rainfall variability induced by East Asian monsoon interdecadal change. *Climate Dyn.*, **50**, 3523–3536, https://doi.org/10.1007/s00382-017-3821-8.