Contribution of Meteorological Downscaling to Skill and Precision of Seasonal Drought Forecasts

RYAN A. ZAMORA, BENJAMIN F. ZAITCHIK

JOHNS HOPKINS UNIVERSITY, BALTIMORE, MARYLAND, USA

MATTHEW RODELL, AUGUSTO GETIRANA, SUJAY KUMAR

HYDROLOGICAL SCIENCES LABORATORY, NASA GODDARD SPACE FLIGHT CENTER, GREENBELT, MARYLAND, USA

KRISTI ARSENAULT

SCIENCE APPLICATIONS INTERNATIONAL CORPORATION, MCLEAN, VA, USA AND HYDROLOGICAL SCIENCES LABORATORY, NASA GODDARD SPACE FLIGHT CENTER, GREENBELT, MARYLAND, USA

ETHAN GUTMANN

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH, BOULDER, COLORADO, USA

Corresponding author: Ryan A. Zamora, ryan.zamora@jhu.edu

ABSTRACT

Research in meteorological prediction on sub-seasonal to seasonal (S2S) timescales has seen growth in recent years. Concurrent with this, demand for seasonal drought forecasting has risen. While there is obvious synergy between these fields, S2S meteorological forecasting has typically focused on low resolution global models, while the development of drought can be sensitive to the local expression of weather anomalies and their interaction with local surface properties and processes. This suggests that downscaling might play an important role in the application of meteorological S2S forecasts to skillful forecasting of drought. Here, we apply the Generalized Analog Regression Downscaling (GARD) algorithm to downscale meteorological hindcasts from the NASA Goddard Earth Observing System (GEOS) global S2S forecast system. Downscaled meteorological fields are then applied to drive offline simulations with the Catchment Land Surface Model (CLSM) to forecast United States Drought Monitor (USDM) style drought indicators derived from simulated surface hydrology variables. We compare the representation of drought in these downscaled hindcasts to hindcasts that are not downscaled, using the North American Land Data Assimilation System Phase 2 (NLDAS-2) dataset as an observational reference. We find that downscaling using GARD improves hindcasts of temperature and temperature anomalies, but the results for precipitation are mixed and generally small. Overall, GARD downscaling led to improved hindcast skill for total drought across the Contiguous United States (CONUS), and improvements were greatest for extreme (D3) and exceptional (D4) drought categories.

1. INTRODUCTION

Droughts are one of the most damaging climate phenomena, impacting agriculture, water resources, ecosystems, and human health at local to regional scale. Economic impacts are also significant. In the United States, the cost of drought to the economy since 1980 has been estimated to be at least $249 billion, accounting for costs from 26 unique events (**NCEI 2020**). Early warning of drought has the potential to reduce these costs, through improved preparedness and more rapid impact response (Wilhite and Svoboda, 2000), yet most seasonal drought prediction systems are currently working only in a research capacity, with few dynamically-based systems applied operationally to drought preparation and management.

Research and interest in sub-seasonal to seasonal (S2S) meteorological prediction has risen over the past two decades, as is evident in the appearance of multiple intermodel comparison projects and operational forecast ensembles. These include the North American Multimodel Ensemble (NMME; **Kirtman et al. 2014**), the S2S Prediction Project (**Vitart et al. 2017**), and the Subseasonal Experiment (SubX; **Pegion et al. 2019**). Objectives of these efforts include testing the predictability of various climate phenomena and application to regional climate outlooks (**Thober et al. 2015, Sossa et al. 2017, Vitart 2017, Kim et al. 2019a**). The bulk of these efforts have focused on global dynamically-based prediction systems, with local applications for weather hazards, agricultural impacts, and water resource management (among other sectors) generally mediated by local expert interpretation rather than by a formal customization or downscaling of the global forecast models.

Recent research on S2S time scales has spanned a number of different subjects, such as the Madden Julian Oscillation (MJO; **Jenney et al. 2019, Kim et al. 2019b, Wang et al. 2020**), aerosols (**Benedetti and Vitart 2018**), and tropical cyclones (**Robertson et al. 2020**). As a number

of key impacts of S2S climate variability involve or are caused by changes in hydrology, interest in hydrological S2S prediction has also increased (**Wood et al. 2002, Luo and Wood 2007, Shukla et al. 2014, Wanders and Wood 2016, Arsenault et al. 2020, DeAngelis et al. 2020, Pendergrass et al. 2020**). Hydrological S2S forecasts generally rely on atmospheric S2S forecasts to provide meteorological forcing data. As the simulated hydrological impacts of meteorological conditions can be highly sensitive to spatial and temporal resolution (**Berne et al. 2004**), this raises the question of whether it is important to downscale the output of global S2S forecast systems before applying them for hydrological prediction. The answer to this question depends on the characteristics of the modeling system, the climate and hydrology of the regions of interest, and on the applications context—forecasts focused on problems of hydropower reservoir management, for example, might have different needs from those focused on agricultural drought.

Since 2012, the NASA/GSFC Hydrological Sciences Laboratory has collaborated with the United States National Drought Mitigation Center (NDMC) to provide drought monitoring products informed by the Gravity Recovery and Climate Experiment (GRACE) and GRACE-Follow On (GRACE-FO) satellite missions. In this system, GRACE-derived estimates of terrestrial water storage anomalies are assimilated into the Catchment Land Surface Model (CLSM) to provide continuous monitoring of drought and wet anomalies in near-surface soil moisture, root zone soil moisture, and shallow groundwater (**Houborg et al. 2012, Li et al. 2019**). This information is disseminated publicly by NDMC and also provided directly to authors of the U.S. Drought Monitor as an input for their expert-informed drought maps. In monitoring mode, this system draws meteorology from the North American Land Data Assimilation System Phase 2 (NLDAS-2; **Xia et al. 2012**). Recently, the system has been enhanced to provide seasonal forecasts in addition to drought and wetness monitoring (**Getirana et al. 2020**). In forecast mode, the system

89 draws meteorology from the NASA Global Earth Observation System (GEOS) subseasonal to
90 seasonal forecast ensemble (**Molod et al. 2012**). The objective is to provide season-ahead forecasts
91 of the same drought/wetness indicators that are offered in monitoring mode.

92 Here, we assess the importance of downscaling GEOS forecast meteorological fields for
93 use in this system. Where **Getirana et al. (2020)** bias-corrected but did not downscale GEOS
94 forcing, we test an implementation of the forecast system that applies the Generalized Analog and
95 Regression Downscaling (GARD; **Gutmann et al. 2021**) tool to GEOS surface meteorological
96 fields. GARD implements a hybrid analog-regression approach to downscaling, in which multiple
97 input predictor variables can be used to estimate each variable at each grid cell. First, an analog
98 approach is used to select a group of analog days from the training period for each day to be
99 predicted. These analog days are then used to train a multi-variable regression, and this regression
100 is applied to predict a downscaled value on the day of interest. Compared to commonly used
101 statistical disaggregation techniques like Bias Correction and Spatial Disaggregation (BCSD;
102 **Wood et al. 2002**), GARD allows the downscaling process to be informed by multiple variables
103 at multiple scales, potentially taking advantage of skill in an atmospheric forecast even when the
104 model's prediction of a particular target variable (e.g., local precipitation) has large errors. In the
105 context of seasonal forecasts, the use of GARD is novel in that it offers the potential to improve a
106 dynamically-based seasonal forecast by correcting for systematic errors in the large-scale forecast
107 related to both the spatial placement and temporal variability of weather processes.

108 Given the large uncertainty in subseasonal-to-seasonal atmospheric forecasts, and the
109 potential for errors in meteorology to have nonlinear impacts on hydrological simulations, skillful
110 downscaling of a global forecast system could be quite important when forecasting drought

111 indices. This paper evaluates this sensitivity for the case of an operational drought and wetness
112 monitoring system for the United States.
113
114 **2. DATA**
115 GARD, like most statistical downscaling methods, requires a reliable high resolution
116 "observational" dataset with a sufficiently long record that includes the meteorological forcing
117 variables being downscaled, along with a training dataset that is representative of the spatially
118 coarse forecast data that need to be downscaled. In our case, we need to downscale the 9 surface
119 meteorological variables required by the CLSM: air temperature at 2-meters above the surface
120 (*T2M*), U and V wind components at 10 meters above the surface (*U10M* and *V10M*, respectively),
121 specific humidity at 2 meters above the surface (*Q2M*), surface pressure (*PS*), surface downward
122 longwave and shortwave radiation (*LWS* and *SLRSF*, respectively), total precipitation
123 (*PRECTOT*), and convective precipitation (*CNPRCP*).
124
125 *2.1. Observational Dataset*
126 The "observational" dataset we use in this study is NLDAS-2. The data has a fine spatial
127 grid (0.125°), frequent temporal resolution (hourly), and has data starting in 1979 with near real-
128 time data released to present day. NLDAS-2 is not a purely observational dataset, in that it is
129 derived from a combination of models and observations, but we refer to it as the observational
130 dataset here because of its role in the downscaling process. Nevertheless, its long record, high
131 resolution, and extensive history of evaluation and successful use in applications make it an
132 appealing dataset to use as a high-resolution training dataset for our downscaling purposes. For
133 our downscaling technique, we need 6-hourly, daily averaged, and climatological averaged daily

resolution datasets; here we take hourly observations sampled at 00, 06, 12, and 18 hours to make our 6-hourly dataset, and average this data to generate a daily dataset for the period 1982-2017 (to match the temporal availability of our forecasting dataset). We then create a long term daily averaged climatological data set by averaging each day of the year over the entire period.

2.2. Forecasting Dataset

We use the NASA Goddard Earth Observing System subseasonal to seasonal forecast ensemble (GEOS-S2S) for our forecasting and training dataset (**Molod et al. 2012, Borovikov et al. 2017**). Here we use the original version of the system ("version 1") in order to study performance over a long hindcast record. This dataset operates ~11 ensembles (4 unperturbed lagged forecasts initialized at 5 day intervals, and 7 perturbed simultaneous forecasts) per forecast month, and has a ~1° global spatial grid. The available ensembles are generated from perturbed and unperturbed initial states with start dates closest to the start of the month. GEOS has recently been updated, with some modifications to the model and ensemble strategy.

Downscaling an ensemble forecasting product allows us the opportunity to apply the downscaling technique to multiple forecast realizations using a single observational dataset. Each ensemble member is treated as a unique dataset, downscaled individually, and is only averaged post analysis for purposes of plotting and presentation. For this study, we focus on 3-month hindcasts initialized on May 1st of each year. GEOS forecasts run for nine months, but we focus on the first three months because meteorological forecast skill degrades significantly over time. The surface meteorological forcing fields taken from this dataset have a daily temporal resolution and match those of the observation-based dataset and span the period 1982-2017. For consistency

with the overall project (Getirana et al. 2020), we chose to limit the results shown here to the 2003-2017 period.

3. METHODS

3.1. Generalized Analog Regression Downscaling

We utilize the National Center for Atmospheric Research (NCAR) Generalized Analog Regression Downscaling (GARD) algorithm for our downscaling purposes. As its name suggests, GARD has capabilities of downscaling using an analog approach, regression approach, or a combination of the two, which we utilize in our study. It requires three input datasets for downscaling: an observational dataset taken to be "truth" (in our case NLDAS-2), a training dataset that is compared with our observational dataset (historical GEOS hindcasts), and a prediction dataset that we wish to downscale (GEOS for the time period to be downscaled).

The observational and training datasets must have identical time steps in order for the algorithm to create a link between the coarse and fine resolution data points; this is necessary for finding analog values for the predictand variable. The algorithm examines a single data point in the prediction dataset for analog values of that grid cell in the training dataset (and its paired corresponding value in the observational dataset) and performs a regression on this subset to determine its downscaled value. Importantly, the regression applied in this process can be multi-variable, such that information from multiple aspects of the atmospheric forecast can inform the estimate of the meteorological value of interest. This results in a regression only being applied using observation and training data pairs that are analogous to the value we want to downscale; for example, if we want to downscale a temperature value where weather conditions are cold and have low winds, the downscaled value is only predicted using days that share these conditions. This

process is repeated on a grid-cell by grid-cell basis for each data point in the prediction dataset, resulting in gridded predicted values with a resolution of the observational dataset.

The results GARD produces are fundamentally different than only correcting biases in the mean of the target variable. While part of the technique inherently does accomplish a bias correction, the analog and regression approach used in GARD can also inform estimates of temporal variability (anomalies) in the downscaled variable. This also allows for better representation of extreme event statistics.

3.2. GARD Configuration

There are various options that can be set in the downscaling algorithm to change its function. A full review of GARD configuration options and sensitivities is beyond the scope of this paper (**see Gutmann et al. 2021**), but we will address the options most relevant to our implementation of this technique. While GARD has the ability to downscale using a pure regression or pure analog approach, we use the hybrid analog regression to fully utilize both features in the code.

The number of analogs chosen will affect the output of the regression: choosing too few analogs will yield a regression from a biased sample, while using too many reduces the value of using analog-regression over regression alone. Therefore, we ran a number of experiments, varying only the number of analogs to find a suitable number, and found that 300 (~10% of the training dataset) worked for our purposes. In addition, we set the option to weigh each analog by how close its value (inverse-square distance) is to the current predictand.

Selecting the variables to include in the regression is also consequential. Theoretically, the more information you have on the large-scale circulation of your domain, the more accurate the

resultant regression would be. Ideally this would include variables such as geopotential height or upper level winds; however, a major tenet of our project is to create a self-contained system that can be used operationally for downscaled seasonal drought/flood forecasting without placing overwhelming data burdens on the forecaster. Thus, we limit the variables used in multi-variate regressions to the set of 9 variables (**Table 1**) we are downscaling. While a formal study on the choice of predictors (particularly by region) would be an interesting research topic, it is beyond the scope of this paper and we use the variables found most important in our testing.

Total and convective precipitation need to be handled in a distinct way, due to zero inflation in records of six-hourly precipitation. GARD offers an option to apply a transform to the input data, perform the downscaling, and then apply the inverse transform on the output data; for precipitation, we use a cube root transformation. Furthermore, a logistic regression is applied to compute the probability of occurrence of precipitation, which we can then use in post-processing.

Output precipitation-based variables have a conditional bias owing to the distribution of precipitation being non-Gaussian. GARD optionally outputs the probability of exceeding set threshold values (in the case of precipitation this value is 0, indicating no precipitation); in this scenario, this is analogous to the probability of precipitation (PoP). Using this in conjunction with the residuals of the regression used to compute the mean precipitation amounts, we can then use a spatial-temporal autocorrelation technique (**Clark and Slater 2006, Gutmann et al. 2021**) to post-process precipitation.

In addition, the temporal resolution of our downscaled forecast dataset is daily, and needs to be temporally disaggregated to 6-hourly as an additional step prior to input to LIS. We achieve this by first creating a 6-hourly long term mean of our observational dataset, followed by a transformation of this data into 6-hourly anomalies of the long term daily mean (thus capturing the

diurnal cycle). We then apply these anomalies to our output forecast datasets to generate temporally disaggregated 6-hourly forecast data. The type of anomalies used can be either additive (shifting the values) or multiplicative (applying a ratio to the values), and each is chosen depending on the type of variable to be disaggregated; for this study, we use multiplicative anomalies for precipitation variables (convective and total), and use additive anomalies for other variables.

3.3. Land Surface Modeling

We use NASA's Land Information System (LIS; **Kumar et al. 2006**) as a framework for our land surface modeling, which can easily be used with each of our datasets and has been used extensively in a number of studies (**e.g. Seneviratne et al. 2010, Zaitchik et al. 2010**, **Clark et al. 2015, Kumar et al. 2016, McNally et al. 2017**). Within LIS, we use the Catchment Land Surface Model (CLSM; **Koster et al. 2000**) with our downscaled meteorological variables as a forcing dataset. This experiment could be repeated with a number of different land surface models within LIS, however an extensive study comparing the differences between these is beyond the scope of this paper. CLSM was originally chosen for the drought/wetness monitoring application because it simulates groundwater storage changes, which is essential for GRACE data assimilation. It is used in drought/wetness forecasting application, and hence in this study, for the same reason.

Although GEOS is a global product, our model domain strictly covers the Contiguous United States (CONUS; to the extents of the NLDAS-2 product). As we are conducting a comparative study of three experiments using distinct meteorological forcing fields, it is important that the initial conditions of each are similar. While the focus of this research is on the most recent decades, in order to properly create drought/wetness indicators (based on climatological soil moisture percentiles), we require a longer record than our datasets provide. Because of this, we

248 opt to use the Princeton meteorological reanalysis dataset (**Sheffield et al. 2006**), which
249 encompasses 1948 – 2014, as the meteorological forcing to spin-up the CLSM states. We initialize
250 our NLDAS-2 meteorological forcing based runs using the resulting initial conditions to generate
251 our "observational" hydrological data. However, because the climatologies differ between the
252 datasets, we rescale the NLDAS-2 meteorological fields to match Princeton's monthly climatology
253 for our run period using simple scaling factors (Houborg et al. 2012). We then use these same
254 initial conditions to initialize both the original GEOS and downscaled GEOS hindcasts. There are
255 a number of output hydrological variables from this experiment; however, we will focus on only
256 three of them here: soil moisture, evapotranspiration, and runoff.
257
258 *3.4. Drought Indicators*
259 The U.S. Drought Monitor (**USDM; Svoboda et al., 2002**) is widely available to the public
260 via its website (**https://droughtmonitor.unl.edu/**) hosted by the National Drought Mitigation
261 Center (NDMC) at the University of Nebraska, Lincoln. It is used in a number of different sectors
262 for decision making including but not limited to declaring drought disaster situations, prescribing
263 federal aid for farmers, performing state and local level drought assessments, and informing farmer
264 planting and management decisions. The USDM provides a weekly operational product that is
265 generated by a team of authors including one lead (the position rotates). The team subjectively
266 combines information from a number of sources, including modeled climatological conditions,
267 accumulated precipitation and other compiled indices, satellite based measurements, local
268 observations, and input from state climatologists, other experts, and a network of volunteer
269 observers. The observers receive a draft of the map ahead of time and provide feedback and
270 suggestions based on the knowledge they have on the conditions at their location.

271 The different categories of drought can be determined using a variety of different variables
272 and indices, such as the Palmer Drought Severity Index (PDSI), the Standard Precipitation Index
273 (SPI), and modeled soil moisture or streamflow percentiles. For the research presented here, we
274 use root zone soil moisture percentiles on account of their relevance to agriculture, a primary sector
275 of interest for drought monitoring. These percentiles are created by comparing the local soil
276 moisture to its climatology; lower percentiles indicate drier than normal conditions for that location
277 and time of year, while higher values indicate wetter than normal conditions. The 5 categories the
278 NDMC uses to identify drought are D0 (Abnormally Dry; 21% – 30%), D1 (Moderate Drought;
279 11% – 20%), D2 (Severe Drought; 6% – 10%), D3 (Extreme Drought; 3% –5%), and D4
280 (Exceptional Drought; 0% – 2%).

281 To compile the spatial drought indicators and simplify the analysis, we aggregate the
282 results to distinct regions of the United States. While we could have regionalized by state, USGS
283 hydrologic unit code (HUC), or by applying a statistical regionalization technique, we chose to use
284 Bukovsky Regions (**Bukovsky 2011**). These are a collection of 29 regions (with 13 larger
285 compound groupings) that have been delineated based on eco-regions; that is, they respond
286 similarly to variations in temperature and precipitation, or share an important regional climate
287 feature. In order to use this resource with our datasets, we interpolated the masks (defined on a
288 0.5° latitude/longitude grid) to the resolution of NLDAS-2, 0.125°. For the purposes of our
289 research, we focused on the five compound regions that cover the majority of the CONUS: Central
290 U.S., Eastern U.S., Southern U.S., Desert, and Mountain West. While this does exclude some
291 regions of the CONUS (namely parts of California and Michigan), we believe that using only
292 compound regions simplified the presentation of the analysis and was sufficient for our research.
293 The boundaries and locations of each of these compound regions can be seen in **Figure 1**.

4. RESULTS

4.1. Downscaled Meteorological Fields

We begin with an examination of the downscaled meteorological fields from a climatological perspective (i.e., averaged over 2003-2017 for each month individually). While each of the 9 variables provided in **Table 1** is used as input to the land surface model, here we focus on temperature and precipitation, given their central importance to local climate and hydrology. In viewing Figures 3-9, it is helpful to remember that blues indicate more skill in GARD relative to GEOS while reds indicate the opposite. In addition, each of these figures include an inset table showing the long term monthly regional averages of the variables shown.

Figure 2 shows an example of monthly averaged 2-meter temperature and total precipitation for NLDAS, GEOS, and GARD. These plots were created by averaging the initial 31 days of the forecast for May in the year 2004. This figure showcases the resolution differences between observations and the raw downscaled data. This emphasizes a key characteristic of downscaling the original GEOS forecasts: coarser resolution datasets have difficulty in resolving topography at high elevations. This is particularly important for hydrological forecasts in regions with significant snowpack, and it can also influence water balance simulations via temperature impacts on potential evaporation and snowmelt.

Downscaled temperature shows greater detail over the mountains, resolving elevation contrasts, and generally captures many of the topographic features found in observations. Precipitation is a much more difficult meteorological phenomenon to forecast, such that GARD shows significant differences from NLDAS in a number of regions. The overall distribution of precipitation statistics in GARD does, however, match NLDAS more closely than GEOS does,

reflecting the ability of the downscaling algorithm to replicate higher intensity precipitation at sub-GEOS resolution. Notably, because we use multivariate regression through downscaling, GARD precipitation is a result of information from many variables (total precipitation, convective precipitation, temperature and surface pressure; see **Table 1**). This fact, in addition to the decrease in accuracy of precipitation forecasts overtime (e.g. Koster et al. 2004), results with some regions having different precipitation patterns between GARD and GEOS.

Figure 3 shows the climatological difference between downscaled forecasts (which we will refer to as "F_{GARD}") and raw forecasts (which we will refer to as "F_{GEOS}") in absolute bias of 2-meter temperature for May (top panel), June (middle panel), and July (bottom panel), where bias is calculated for each dataset relative to NLDAS-2. This figure highlights where the greatest departure from NLDAS-2 between the two datasets occurs, regardless of whether the value is under or over predicted. Warmer colors are indicative of areas where F_{GARD} has a larger absolute bias (greater departure from the true value), whereas cooler colors highlight areas of larger absolute bias in F_{GEOS}. Note that F_{GEOS} is linearly interpolated to match the spatial resolution of F_{GARD} and NLDAS-2. One of the more noticeable features in this figure is the dominance of large F_{GEOS} absolute bias in temperature when compared with F_{GARD}, particularly in the mountains and the Great Lakes region. In each month, there are select regions where F_{GARD} has a larger absolute bias than F_{GEOS}, but these regions are in the minority and the difference in biases is of smaller magnitude.

Figure 4 is similar to **Figure 3** but for long-term averaged total precipitation (values shown in *mm*). In contrast to temperature, absolute precipitation bias varies greatly between the two datasets. There are many regions where F_{GARD} has a larger absolute bias, and in May F_{GEOS} appears less biased overall. The spatial distribution of signed bias of F_{GEOS} (not shown) shows that F_{GEOS}

generally under-predicts precipitation in the Central and Southern U.S., while over-predicting elsewhere. However, these regions of over/under-predicting the observed value are not necessarily distributed similarly in F_{GARD}, leading to a less clear result than was seen with temperature. While there are some consistencies in absolute bias difference over time, there does not appear to be a general consensus where one forecast has better results regionally than the other.

While the majority of the results to be presented in this research use ensemble means for brevity, we also include a measure for probabilistic forecasts using the individual ensemble members. The Brier Score (BS) is a measure of the error in a probabilistic forecast of a binary event occurring. For our purposes, we wanted to capture the probability of a low precipitation event (<= 10% percentile; a strong indicator for drought). The Brier Skill Score (BSS) measures the accuracy of a probabilistic forecast relative to a reference forecast (climatology). A perfect accuracy forecast has a BSS of 1, the reference forecast has a value of 0, and values lower than that (to $-\infty$) represent forecasts worse than the reference.

Table 2 shows the BSS for each month and for the overall forecasting period. Bolded values indicate which forecast (F_{GARD} or F_{GEOS}) has a higher BSS for a given time period. While both forecasts have BSS near zero, neither shows capabilities higher than the reference forecast at the monthly time scale (as indicated by BSS < 0.0). F_{GARD} shows improved BSS relative to F_{GARD} in June, July, and over the forecasting period, however these changes are small and are still below the reference forecast.

Comparing the differences in absolute bias in the two models is helpful for determining how well they can approximate the mean of the observed dataset. In forecast applications, however, predicting the anomaly is often more important than matching climatology. GARD is a particularly appealing downscaling method when considering forecasts of anomalies, given its

363 ability to use multiple predictor variables to inform the prediction of each value (in contrast to
364 simple disaggregation methods). This means that a GARD-informed forecast can produce anomaly
365 time series that have systematically different variability from the model being downscaled, rather
366 than simply scaling the model anomalies to match the statistical distribution of the local record.

367 **Figure 5** shows the long term mean F_{GARD} - F_{GEOS} difference in normalized root mean
368 square error (NRMSE) of 2-meter temperature anomaly relative to NLDAS-2. Root mean square
369 error (RMSE) is a measure of the spread in the prediction errors when compared to the
370 observational dataset, while NRMSE is RMSE divided by the mean of the observed dataset. We
371 choose to express this value as a percentage by multiplying by 100, so in this context it is a measure
372 of the percentage error in the variation of temperature over time. Since **Figure 5** is the difference
373 of this metric between the raw and downscaled datasets, positive values (warmer colors) indicate
374 regions where F_{GARD} has a larger NRMSE and negative values (cooler colors) are indicative of
375 larger a NRMSE in F_{GEOS}. This figure shows that the 2-meter temperature anomaly error is larger
376 in F_{GARD} in the first month of the forecast, and larger in F_{GEOS} in the following 2 months.
377 Furthermore, while there are some regions with consistently lower or higher NRMSE values
378 comparatively, even the largest differences are no greater than ~0.5℃, showing that skill in
379 reproducing temperature anomalies is largely similar.

380 Likewise, **Figure 6** shows a similar metric as in **Figure 5** but for the anomaly in total
381 precipitation. As was seen for the mean value in absolute bias in **Figure 4**, the spatial distribution
382 in precipitation is much noisier when compared with temperature; however, the regions where
383 each dataset has favorably smaller differences are clustered. During May, F_{GARD} has a larger
384 NRMSE temperature anomaly throughout much of the Southern-Central United States. In contrast,
385 F_{GARD} has consistently smaller error throughout the Eastern and Southern U.S. in June. It is

386 important to note that the largest NRMSE difference values shown here are an order of magnitude
387 smaller than the typical RMSE errors found when F_{GEOS} or F_{GARD} is compared to NLDAS-2, so
388 downscaling seems to have relatively little impact on this metric when compared to general
389 forecast error. That being said, **Figure 6** shows many favorable regions for F_{GARD}, particularly in
390 June and July, over many Eastern and Mountain West regions.
391 We emphasize that GARD is a downscaling framework that allows for substantial user
392 flexibility in the choice of predictor variables and number of analogs. As such, the results obtained
393 here are not the only results that could be obtained when applying GARD to these datasets. It is
394 possible that better results for some regions could be obtained with more extensive testing or with
395 inclusion of predictors from outside our set of nine surface meteorological variables. Our
396 assessment was that these results are, however, adequate for our purposes when compared to other
397 combinations of GARD settings, and we proceed with them for the remainder of the paper.
398
399 *4.2. Downscaled Hydrological Fields*
400 Using the downscaled forecast meteorological variables in conjunction with the CLSM
401 within the LIS framework, we are able to generate high resolution hydrological forecasts. The
402 F_{GEOS} results shown here are linearly interpolated (as previously), but also bias corrected within
403 LIS using MicroMet (see Kumar et al. 2006 & Liston and Elder 2006 for more information). We
404 examine three key hydrological variables: root zone soil moisture, surface runoff, and
405 evapotranspiration, which are heavily influenced by many of the meteorological variables
406 discussed previously (**Kato et al. 2007**).
407 **Figure 7** shows the climatological $F_{GARD} - F_{GEOS}$ difference in normalized root mean
408 square error of root zone soil moisture anomaly, calculated relative to a retrospective Catchment

LSM simulation that uses NLDAS-2 meteorological forcing (hereafter, referred to as "retrospective"). The scale of the differences depicted here are ~1 order of magnitude less than the individual RMSE values, showing that on a climatological scale the errors are spatially similar between the two datasets (with some differences). FGARD shows lower NRMSE values in much of the Southeastern United States and desert regions during each month, and becomes dominantly favorable during the 3rd month of the forecast nationally. However, there do exist regions where the FGEOS has a smaller relative NRMSE, particularly near the Texas/Oklahoma border and parts of the upper Central United States. As soil moisture percentiles (discussed in more depth in the next section) are a key indicator of drought, regional biases in the skill of determining long term mean soil moisture anomalies are important in drought forecasting.

A large portion of the precipitation that does not seep into the ground and contribute to the terrestrial water storage will flow downhill along the surface as runoff. Higher resolution data should be able to better capture these fluxes, which are integral to the water cycle. **Figure 8** is similar to **Figure 7**, but for the anomaly of surface runoff. Compared with root zone soil moisture, there appear to be greater localized differences in the regions showing greater NRMSE in FGEOS and FGARD, however most of the region shows only small differences (less than or equal to 15%) between the two. In contrast, many areas (particularly in desert regions in June, and Mountain West in July) show much lower NRMSE runoff anomalies in FGARD.

Evapotranspiration (ET) is an important process to evaluate when considering drought; it is a measure of the flux of water from the surface to the atmosphere through soil moisture evaporation and plant transpiration. Being able to properly identify and predict changes in ET are important for the agricultural sector in decision making on irrigation. Furthermore, it has been reported that ET is important in detecting flash droughts (**Koster et al. 2019**) which can be

expensive and agriculturally and ecologically devastating events. **Figure 9** is similar to **Figures 7** and **8**, but for the anomaly of evapotranspiration. Much of the United States shows smaller NRMSE in F_{GARD}; however, this is not true in some regions, such as Texas and the Southern Plains. Much of the South-Eastern United States and the West have favorable values for F_{GARD} in each of the three months.

The evolution of relative performance of F_{GARD} and F_{GEOS} for evapotranspiration exemplifies a tendency towards relatively smaller error in F_{GARD} in the latter half of the forecast (shown here particularly in July) which is also seen in other variables. This tendency can be explained in part by changes in long term mean temperature (**Figures 3 & 5**), precipitation (**Figures 4 & 6**), and radiation fluxes (not shown). It is also likely that this reflects the dominant control of precipitation over soil moisture early in the forecast, transitioning to influence of surface energy balance terms later in the forecast, as precipitation forecast skill degrades. We will explore this phenomenon further as we examine drought indicators in the following section.

4.3. Monthly Averaged Drought Indicators

Here we evaluate how well F_{GARD} (compared to F_{GEOS}) forecasts each of the USDM drought categories individually and cumulatively over the 2003-2017 period relative to retrospective. There are a few ways to represent this, of which we will be exploring two here: monthly mean and RMSE percentage of area in drought. Using the regions shown in **Figure 1**, we can examine the percentage of each region that is experiencing drought. **Figure 10** shows the monthly average of this measure for each USDM drought category in each month, year, and dataset over the CONUS region. Each month and year in the plot includes three bars: retrospective, F_{GEOS}, and F_{GARD}.

454 We will examine recent notable drought years specifically in the following section. It is
455 evident from **Figure 10**, however, that both F_{GEOS} and F_{GARD} are capable of capturing interannual
456 variations in the percentage of area in total drought (expressed as total bar height), with some
457 exceptions. These results reflect the combined contribution of forecast skill and, importantly, skill
458 in initial surface conditions. We note that while not all regions experienced extreme or exceptional
459 drought based on observations during the period of record, there is interest to examine the ability
460 to forecast weak drought in off years. Further, there does not appear to be a general dry or wet bias
461 in either dataset when compared to retrospective. One promising feature is how well F_{GARD} does
462 in correctly determining the percentage of area in total drought when F_{GEOS} over-predicts
463 (frequently seen in June between 2003 – 2010) and under-predicts (frequently seen in July 2010 –
464 2017). Moreover, this aspect is especially true not only for the total drought percentage, but for
465 the exceptional drought category (D4), meaning that F_{GARD} is more capable in resolving drought
466 severity.

467 There are some months where F_{GARD} degrades skill relative to F_{GEOS} in predicting the
468 percentage of area in all drought classes. These periods are often when F_{GEOS} is not skilled in
469 capturing drought either, but F_{GARD} does even worse. This underscores an important tenet of our
470 research: while downscaling can and does improve our results overall, we shouldn't expect an
471 unskilled forecast to be saved solely through downscaling. In addition, we have included this figure
472 for different regions in the appendix (Figures A1 & A2) and will address some of the regional
473 differences later in this section.

474 Figure 11 shows a scatter plot of the data shown in Figure 10 (monthly average percentage
475 of the Contiguous U.S. in drought) by drought category and month. Retrospective values are shown
476 along the x-axis and the associated forecast values are shown on the y-axis for F_{GEOS} (red) and

477 F_{GARD} (blue). Values along the 1:1 diagonal represent strong correlation, whereas values above
478 (below) this line represent over (under) prediction. The coefficients of determination (r^2) between
479 the forecast and retrospective values are also included on the bottom right of each panel. As lower
480 drought severity is more common, the top-right section of each panel commonly has lower drought
481 categories, while the bottom-left section has more severe drought categories. The spread of the
482 data in both forecasts is low in May, and becomes increasingly wider (chiefly in lower drought
483 categories) at longer lead times. Overall, F_{GARD} shows stronger correlations with observation when
484 compared with F_{GEOS} (though both are high), particularly in the most severe (D3 and D4) drought
485 categories.

486 **Table 3** shows r^2 for monthly average percentage of regions in drought determined by root
487 zone soil moisture percentiles for 2003-2017 by (a) category and (b) month. Bolded values indicate
488 which forecast (F_{GARD} or F_{GEOS}) has a larger correlation (relative to retrospective) for a given
489 region. This table summarizes the results found in Figure 11, but includes each of the sub-regions
490 of our domain. The correlations by drought category show a higher correspondence in F_{GARD} over
491 all regions, with the exception of the Central U.S. where F_{GEOS} is higher (though the differences
492 between them in this region are generally small; < 0.1). In addition, some regions (CONUS and
493 Desert) have a slightly higher relative correlation in F_{GEOS} at lower drought categories (D0 and
494 D1). Examining the correlations by month (panel B), shows stronger correlations in F_{GARD} in all
495 regions and months, except for the Central U.S. in June. Overall, this table highlights the Eastern
496 U.S. (Central U.S) as one of the most (least) favorable regions in F_{GARD}.

497 We examine the regional contrasts in **Figure 12**, which summarizes the results discussed
498 previously using RMSE of the percentage of area in drought for the CONUS (top panel), Central
499 (middle panel), and Eastern (bottom panel) U.S. regions. Here we calculate the difference (F_{GARD}

500 – F_{GEOS}) in RMSE from daily values for each category, forecast year, and climatological month.
501 We have also included the cumulative drought category, which corresponds to < 30% in root zone
502 soil moisture percentiles (bottom row of each table). Blue cells (positive values) are indicative of
503 lower F_{GARD} error, and red cells (negative values) indicate lower F_{GEOS} error. This depiction
504 reinforces some of the key results of the previous figures, namely the varying skill F_{GARD} has in
505 the Central vs. Eastern United States, doing exceedingly well at capturing the most severe droughts
506 in CONUS and Eastern regions, and more often than not showing improvement in its ability to
507 capture the percentage of area experiencing drought.

508 Looking at the right side of the figure, we can see how well F_{GARD} does in each
509 climatological month. Overall it shows improvement for each category and month, with the
510 exception of May where for a few categories F_{GEOS} performed better in each region. This follows
511 with our general finding that precipitation controls the soil moisture forecast in the first month, but
512 evapotranspiration (driven by temperature and solar radiation) controls more in months 2 and 3.
513 Downscaling can't help much in overcoming a poor precipitation forecast, but because GARD is
514 more skilled in downscaling temperature, F_{GARD} begins to outperform F_{GEOS} in months 2 and 3.

515 In addition to comparing the percentage of each region experiencing drought among the
516 datasets, we also want to present an evaluation of the detection skills of F_{GARD} and F_{GEOS}. We
517 examined a number of metrics such as probability of detection and false alarm rate, and threat
518 score (also called the critical success index) which measures a forecast's ability to detect
519 observational features on a grid-cell by grid-cell basis. **Figure 13** shows the threat score difference
520 for drought severity categories by week and region. Threat score (TS) is a ratio between hits
521 (number of grid cells correctly detecting an observed drought category) and the sum of hits, misses
522 (number of grid cells not detecting a drought category that was observed), and false alarms

(number of grid cells detecting a drought category that was not observed). The values for the threat score can range from 0 – 1, with 1 being the best score. As we are depicting the difference in TS, positive (negative) values are indicative of better F_{GARD} (F_{GEOS}) performance.

The data shown in **Figure 13** are weekly averaged (rather than monthly) to better depict changes in skill as the forecast progresses. Over the CONUS, F_{GARD} TS is better than F_{GEOS} at the very beginning and tail end of the forecasts. The Central U.S has one of the weakest F_{GARD} results for both cumulative and D3-D4 joint categories (values are rarely positive for F_{GARD}), which is consistent with our previous findings. A majority of the regions have favorable F_{GARD} scores in the last few weeks of the forecast overall, and in the Eastern U.S. F_{GARD} almost always outperforms F_{GEOS} for D3-D4 scores.

4.4. Spatial Drought Indicators in Recent Years

As mentioned previously, the USDM incorporates data from a variety of sources. One source is the ground water and soil moisture drought indicator product derived from LIS/CLSM driven by NLDAS-2 forcing with GRACE data assimilation (GRACE-DA). This product is generated and released weekly by scientists at NASA Goddard Space Flight Center. It is directly comparable to our results (**Figure 14**) because both sets of simulations use LIS/CLSM; the differences being our usage of forecast data (as opposed to NLDAS-2 meteorology) for forcing and our lack of data assimilation. The main difference between the LIS/CLSM soil moisture drought indicators used here for evaluation and the GRACE-DA drought indicators is that the former are derived from an open loop (no data assimilation) LIS/CLSM simulation. While we have run similar experiments using data assimilation, we've elected not to include that in this study, as **Getirana et al. (2020)** have already evaluated the effects of GRACE data assimilation on drought/wetness forecasting and it would complicate the analysis of our results.

There are a number of notable U.S. summer droughts within our experiment period (2003 – 2017), which can be used as benchmarks for our downscaled forecasted drought indicators. In this section, we focus on two such events, the summer droughts of 2006 and 2012. The drought in 2006 was quite costly, especially for those in the agricultural sector. Soil moisture exhaustion lead to severely damaged pastures, loss of crops, a number of wildfires, and other devastating effects (**Kogan & Guo 2015**). **Figure 14** depicts spatial drought indicator (root zone soil moisture percentile) forecasts at 30, 60, and 90 days after the forecast initialization for F_{GEOS} and F_{GARD} compared with drought indicators derived from the retrospective (open-loop, NLDAS-2 forced LIS/CLSM) simulation for the summer of 2006. The color scale follows the USDM categories for drought but also includes the upper percentiles (typically used to diagnose floods/water rich locations).

The retrospective simulation shows a growing drought throughout much of the U.S. (particularly in the Central region), with a large number of areas experiencing exceptional drought in the North and Northwest U.S. by the end of July. At 30 days into the forecast, F_{GEOS} has over-predicted the drought in the Eastern and parts of the Southern U.S. and under predicted drought in the Central regions. At 60 days after, the plots are largely similar with a few regional exceptions (e.g. Indiana, Ohio, and Louisiana). However, at 90 days, while F_{GEOS} is able to capture much of the areas experiencing drought, the severities are not captured well. F_{GARD} seems to under predict the drought at 30 days, but generally improves by day 60. At 90 days after initialization, F_{GARD} also captures much of the area experiencing drought, as well as the severity in parts of the Northern Central U.S. Like F_{GEOS}, F_{GARD} predicts an overly extreme drought in Louisiana, Mississippi, and Alabama, but it corrects that problem in Florida.

568 The summer U.S. drought of 2012 was one of the worst droughts, in terms of both severity
569 and extent, since the early 1930s. It is estimated that the total cost of this drought was roughly
570 $33.9 billion dollars (NCEI **https://www.ncdc.noaa.gov/billions/),** and resulted in widespread
571 harvest failure for a number of crops. There are a number of characteristics that led to this event,
572 including one of the largest summer rainfall deficits ever recorded, and summer temperatures
573 among the highest on record to date. However, one of the most notable aspects of this event was
574 the difficulty that many climate models had in predicting it, due in part to its sudden onset
575 (**PaiMazumder & Done 2016**).

576 **Figure 15** is similar to **Figure 14** but for the year 2012. The NLDAS-2-forced LIS/CLSM
577 plots show widespread drought throughout the U.S. at the beginning of June and increasing
578 severity in the following two months. At 30 days after initialization, F_{GEOS} captured some of the
579 drought in the western and eastern United States, but over-predicted its severity; moreover, it
580 missed the drought completely in the central United States. At 60 days after initialization, it is
581 unable to capture the drought, forecasting wet conditions in some of the most severe locations. At
582 90 days after, it does capture much of the drought through the central U.S. but still does not capture
583 the severity well in many regions. Using the downscaled data, not much is improved at day 30,
584 however at 60 days after initialization F_{GARD} captures much more of the drought than F_{GEOS} does
585 (though both significantly under-predict extent). At 90 days after initialization, F_{GARD} shows a
586 higher skill in detecting the extent of the drought, while still under-predicting the severity in the
587 central U.S. Thus, while we can see improvement through downscaling, there is still substantial
588 room for progress.

589

590 **5. SUMMARY/CONCLUSIONS**

591 While seasonal forecasting has seen great advances in recent decades, many seasonal
592 ensemble forecast products have relatively coarse resolutions. Downscaling has the potential to
593 improve these forecasts by increasing their resolution and, in the case of GARD, by leveraging
594 forecast information from multiple fields to inform the estimate of each variable of interest. In this
595 application, we find that GARD downscaling provides clear advantages over raw GEOS forecasts
596 in bias correcting local temperature. Using GARD shows mixed results in downscaling
597 precipitation. The fact that GARD can significantly improve or degrade a forecast at regional
598 scales stems from the fact that it attempts to inform estimates with information from multiple
599 variables. Given the degrees of freedom available to optimize GARD for each variable, it is likely
600 that better performance could be obtained by customizing GARD settings on a regional basis.
601 Results might also be improved by including upper-air meteorological variables that can provide
602 information on weather patterns. Furthermore, the regression approach in GARD might perform
603 better with a different training dataset. Because there is no expectation for a seasonal forecast to
604 match daily precipitation values, it is likely to revert towards climatology of the analog days
605 selected. This approach was used here to keep the training data as consistent with the forecast data
606 as possible.

607 From the perspective of seasonal forecasting, capturing anomalies is generally more
608 important than replicating climatological mean values. For this reason, our evaluation of F_{GARD}
609 largely focused on skill in capturing meteorological anomalies. Results are mixed. Overall,
610 downscaling did improve representation of meteorological anomalies, but that was not the case in
611 all regions—especially for precipitation—or for all times, and differences between F_{GARD} and
612 F_{GEOS} were generally small relative to the difference between both and the NLDAS-2 reference

dataset. F_{GARD} tended to outperform F_{GEOS} in the eastern United States and for the final month of the forecast period, but in the central United States downscaling with GARD degraded some fields.

Our evaluation of the resulting drought indicators showed that F_{GARD} increased skill relative to F_{GEOS} in predicting drought across the CONUS and in many sub-regions. Moreover, F_{GARD} generally better detected and resolved extreme (D3) and exceptional (D4) drought. This is an encouraging finding, as if it can be replicated using a wider variety of datasets and in regions outside the United States, it would suggest that downscaling meteorological forecasts can be advantageous when forecasting major droughts at S2S timescales. Performance in forecasting soil moisture drought generally followed performance in forecasting meteorology: for regions and months where the downscaled forecast improved precipitation and/or temperature estimates, it improved soil moisture drought predictions as well.

We conclude that over the CONUS and during the period of this study, downscaling with GARD improved forecasts more often and in more regions than it degraded them, and downscaling appears to have particular promise for capturing the most severe droughts. For the majority of sub-regions in our study, it appears to be preferable to implement a drought forecast system that uses F_{GARD} over one that simply uses F_{GEOS}. Further, the value of downscaling with an advanced, customizable tool like GARD can vary with location, climate, and degree of customization. At the same time, we recognize that this implementation of GARD did not improve performance consistently for all regions, and that the improvement that downscaling offers can overcome only a small portion of forecast error.

6. ACKNOWLEDGMENTS

This study was funded by NASA's Applied Sciences Program–Water Resources Award 13-WATER13-0030, and the GRACE and GRACE Follow On Science Team. The NLDAS-2 meteorological dataset was acquired as part of the activities of NASA's Science Mission Directorate and is archived and distributed by the Goddard Earth Sciences (GES) Data and Information Services Center (DISC) on https://earthdata.nasa.gov/about/daacs/daac-ges-disc. LIS is open source and available on https://lis.gsfc.nasa.gov.

7. REFERENCES

Arsenault, K. R., and Coauthors, 2020: The NASA Hydrological Forecast System for Food and Water Security Applications. *Bull. Amer. Meteor. Soc.*, **101**, E1007 - E1025, https://doi.org/10.1175/BAMS-D-18-0264.1.

Benedetti, A., and F. Vitart, 2018: Can the Direct Effect of Aerosols Improve Subseasonal Predictability?. *Mon. Wea. Rev.*, **146**, 3481–3498, https://doi.org/10.1175/MWR-D-17-0282.1.

Berne, A., G. Delrieu, J.-D. Creutin, C. Obled, 2004: Temporal and spatial resolution of rainfall measurements required for urban hydrology, *J. Hydrol.*, **299**, Issues 3–4, 166-179, ISSN 0022-1694, https://doi.org/10.1016/j.jhydrol.2004.08.002.

Borovikov, A., R. Cullather, R. Kovach, J. Marshak, G. Vernieres, Y. Vikhliaev, B. Zhao, and Z. Li, 2019: GEOS-5 seasonal forecast system. *Clim. Dyn.*, **53,** 7335–7361. https://doi.org/10.1007/s00382-017-3835-2.

Bukovsky, M.S., 2011: Masks for the Bukovsky regionalization of North America, Regional Integrated Sciences Collective, Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, Boulder, CO. http://www.narccap.ucar.edu/contrib/bukovsky/

Clark, M. P., and A. G. Slater, 2006: Probabilistic quantitative precipitation estimation in complex terrain. *J. Hydrometeor.*, **7**, 3–22, https://doi.org/10.1175/JHM474.1.

Clark, M. P., and Coauthors, 2015: A unified approach for process-based hydrologic modeling: 1. Modeling concept, *Water Resour. Res.*, **51**, 2498– 2514, doi:10.1002/2015WR017198.

662 Gutmann, E.D., J. J. Hamman, M. P. Clark, T. Eidhammer, A. W. Wood , J. R. Arnold, K.
663 Nowak, 2021: Evaluating the effect of regional climate inference methodologies in a
664 common framework (in submission).
665 Getirana, A., M. Rodell, S. Kumar, H. K. Beaudoing, K. Arsenault, B. Zaitchik, H. Save, and S.
666 Bettadpur, 2020: GRACE Improves Seasonal Groundwater Forecast Initialization over
667 the United States. *J. Hydrometeor.*, **21**, 59–71, https://doi.org/10.1175/JHM-D-19-
668 0096.1.
669 Houborg, R., M. Rodell, B. Li, R. Reichle, and B. F. Zaitchik, 2012: Drought indicators based on
670 model-assimilated Gravity Recovery and Climate Experiment (GRACE) terrestrial water
671 storage observations, *Water Resour. Res.*, **48**, W07525, doi:10.1029/2011WR011291.
672 Jenney, A. M., K. M. Nardi, E. A. Barnes, and D. A. Randall, 2019: The seasonality and
673 regionality of MJO impacts on North American temperature. *Geophysical Research*
674 *Letters*, **46**, 9193– 9202. https://doi.org/10.1029/2019GL083950
675 Kato, H., M. Rodell, F. Beyrich, H. Cleugh, E. V. Gorsel, H. Liu, T. P. Meyers, 2007: Sensitivity
676 of Land Surface Simulations to Model Physics, Land Characteristics, and Forcings, at Four
677 CEOP Sites, Journal of the Meteorological Society of Japan. Ser. II, **85A**, 187-204, Online
678 ISSN 2186-9057, Print ISSN 0026-1165, https://doi.org/10.2151/jmsj.85A.187.
679 Kim, H., M. A. Janiga, and K. Pegion, 2019: MJO propagation processes and mean biases in the
680 SubX and S2S reforecasts. *J. Geophys. Res. Atmos*, **124**, 9314–
681 9331. https://doi.org/10.1029/2019JD031139
682 Kim, H., J. H. Richter, and Z. Martin, 2019: Insignificant QBO-MJO prediction skill relationship
683 in the SubX and S2S subseasonal reforecasts. *J. Geophys. Res. Atmos.*, **124**, 12655–
684 12666. https://doi.org/10.1029/2019JD031416

685 Kirtman, B. P., and Coauthors, 2014: The North American Multimodel Ensemble: Phase-1
686 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal
687 Prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, https://doi.org/10.1175/BAMS-D-12-
688 00050.1.
689 Kogan, F. and W. Guo, 2015: 2006–2015 mega-drought in the western USA and its monitoring
690 from space data, *Geomatics, Natural Hazards and Risk*, **6:8**, 651-
691 668, DOI: 10.1080/19475705.2015.1079265
692 Koster, R. D., M. J. Suarez, A. Ducharne, M. Stieglitz, and P. Kumar, 2000: A catchment-based
693 approach to modeling land surface processes in a general circulation model: 1. Model
694 structure, *J. Geophys. Res.*, **105(D20)**, 24809– 24822, doi:10.1029/2000JD900327.
695 Koster, R. D., M. J. Suarez, P. Liu, et al. 2004. "Realistic Initialization of Land Surface States:
696 Impacts on Subseasonal Forecast Skill." *Journal of Hydrometeorology*, **5 (6):** 1049-1063
697 [10.1175/JHM-387.1]
698 Koster, R. D., S. D. Schubert, H. Wang, S. P. Mahanama, A. M. DeAngelis, 2019: Flash Drought
699 as Captured by Reanalysis Data: Disentangling the Contributions of Precipitation Deficit
700 and Excess Evapotranspiration; *J. Hydrometeor.*, **20 (6)**, 1241–1258.
701 https://doi.org/10.1175/JHM-D-18-0242.1
702 Kumar, S. V., and Coauthors, 2016: Assimilation of Gridded GRACE Terrestrial Water Storage
703 Estimates in the North American Land Data Assimilation System, *J. Hydrometeor.*, **17**
704 **(7)**, 1951–1972.
705 Kumar, S.V., C.D. Peters-Lidard, Y. Tian, P.R. Houser, J. Geiger, S. Olden, L. Lighty, J.L.
706 Eastman, B. Doty, P. Dirmeyer, J. Adams, K. Mitchell, E.F. Wood, and J. Sheffield,
707 2006: Land information system: An interoperable framework for high resolution land

surface modeling, *Environmental Modelling & Software*, **21**, Issue 10, 1402-1415, ISSN 1364-8152, https://doi.org/10.1016/j.envsoft.2005.07.004.

Li, B., M. Rodell, S.V. Kumar, H.K. Beaudoing, A. Getirana, B.F. Zaitchik, L.G. Goncalves, C. Cossetin, S. Bhanja, A. Mukherjee, S. Tian, N. Tangdamrongsub, D. Long, J. Nanteza, J. Lee, F. Policelli, I. B. Goni, D. Daira, M. Bila, G. de Lannoy, D. Mocko, and S. C. Steele-Dunne, 2019: Global GRACE data assimilation for groundwater and drought monitoring: advances and challenges, *Water Resour. Res.*, **55**, 7564-7586, doi:10.1029/2018WR024618.

Liston, G. E., and K. Elder, 2006: A Meteorological Distribution System for High-Resolution Terrestrial Modeling (MicroMet), *J. Hydrometeor.*, ***7*(2)**, 217-234.

Luo, L., and E. F. Wood, 2008: Use of Bayesian merging techniques in a multimodel seasonal hydrologic ensemble prediction system for the eastern United States. *J. Hydrometeor.*, **9**, 866–884, https://doi.org/10.1175/2008JHM980.1.

McNally, A., K. Arsenault, S. Kumar, S. Shukla, P. Peterson, S. Wang, C. Funk, C. D. Peters-Lidard and J. P. Verdin, 2017: A land data assimilation system for sub-Saharan Africa food and water security applications, *Sci. Data,* **4**, 170012, https://doi.org/10.1038/sdata.2017.12.

Molod, A., L. Takacs, M. Suarez, J. Bacmeister, I. S. Song, and A. Eichmann, 2012: The GEOS-5 atmospheric general circulation model: Mean climate and development from MERRA to Fortuna, *NASA Technical Reports Server.*

Mueller, B., and S. I. Seneviratne, 2012: Hot days induced by precipitation deficits at the global scale. Proceedings of the National Academy of Sciences, **109**, 12398-12403,

doi:10.1073/pnas.1204330109.

NCEI, 2020: Billion-Dollar Weather and Climate Disasters. Accessed 15 June 2020, https://www.ncdc.noaa.gov/billions/.

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, doi:10.1002/2016JD025322.

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. Clim. Chang.* **10,** 191–199, https://doi.org/10.1038/s41558-020-0709-0.

Robertson, A. W., F. Vitart, and S. J. Camargo, 2020: Subseasonal to seasonal prediction of weather to climate with application to tropical cyclones. *J. Geophys. Res. Atmos.*, **125**, e2018JD029375. https://doi.org/10.1029/2018JD029375

S. I. Seneviratne, T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, A. J. T., 2010: Investigating soil moisture–climate interactions in a changing climate: A review, *Earth-Science Reviews,* **99**, Issues 3–4, 125-161, ISSN 0012-8252,https://doi.org/10.1016/j.earscirev.2010.02.004.

Sheffield, J., E. F. Wood, and M. L. Roderick, 2012: Little change in global drought over the past 60 years. *Nature*, **491**, 435-438, doi:10.1038/nature11575.

752 Sheffield, J., G. Goteti, and E. F. Wood, 2006: Development of a 50-Year High-Resolution
753 Global Dataset of Meteorological Forcings for Land Surface Modeling. *J. Climate*, **19**,
754 3088–3111, https://doi.org/10.1175/JCLI3790.1.
755 Shukla, S., A. McNally, G. Husak, and C. Funk, 2014: A seasonal agricultural drought forecast
756 system for food-insecure regions of East Africa. *Hydrol. Earth Syst. Sci.*, **18**, 3907–3921,
757 https://doi.org/10.5194/hess-18-3907-2014.
758 Sossa, A., B. Liebmann, I. Bladé, D. Allured, H.H. Hendon, P. Peterson and A. Hoell, 2017:
759 Statistical Connection between the Madden–Julian Oscillation and Large Daily
760 Precipitation Events in West Africa, J. Clim., **30**, 1999-2010, doi:10.1175/JCLI-D-16-
761 0144.1
762 Svoboda, M., and Coauthors, 2002: The Drought Monitor, *Bull. Amer. Meteor. Soc.*, **83**, 1181-
763 1190, https://doi.org/10.1175/1520-0477-83.8.1181.
764 Thober, S., R. Kumar, J. Sheffield, J. Mai, D. Schäfer, and L. Samaniego, 2015: Seasonal Soil
765 Moisture Drought Prediction over Europe Using the North American Multi-Model
766 Ensemble (NMME), *J. Hydrometeor.*, **16 (6)**, 2329–2344. https://doi.org/10.1175/JHM-
767 D-15-0053.1
768 USDM, 2020: United States Drought Monitor. Accessed 15 June 2020,
769 https://droughtmonitor.unl.edu/.
770 Vitart, F., 2017: Madden-Julian Oscillation Prediction and Teleconnections in the S2S. Quarterly
771 *Journal of the Royal Meteorological Society*, **143**, 2210-2220, doi:10.1002/qj.3079.
772 Vitart, F., and Coauthors, 2017: The Subseasonal to Seasonal (S2S) Prediction Project
773 Database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-
774 0017.1.

775 Wanders, N. and E. F. Wood, 2016: Improved sub-seasonal meteorological forecast skill using
776 weighted multi-model ensemble simulations. *Environ. Res. Lett.*, **11**, 094007,
777 https://doi.org/10.1088/1748-9326/11/9/094007.
778 Wang, J., H. Kim, D. Kim, S. A. Henderson, C. Stan, and E. D. Maloney, 2020: MJO
779 Teleconnections over the PNA Region in Climate Models. Part II: Impacts of the MJO
780 and Basic State. *J. Climate*, **33**, 5081–5101, https://doi.org/10.1175/JCLI-D-19-0865.1.
781 Wilhite, D. A., & Svoboda, M. D. (2000). Drought early warning systems in the context of
782 drought preparedness and mitigation. Early warning systems for drought preparedness
783 and drought management, 1-21.
784 Wood, A. W., E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-range experimental
785 hydrological forecasting for the eastern United States, *J. Geophys. Res.*, **107**, 4429,
786 https://doi.org/10.1029/2001JD000659.
787 Xia, Y., and Coauthors, 2012: Continental-scale water and energy flux analysis and validation
788 for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1.
789 Intercomparison and application of model products, *J. Geophys. Res.*, **117**, D03109,
790 doi:10.1029/2011JD016048.
791 Zaitchik, B. F., M. Rodell, and F. Olivera, 2010: Evaluation of the Global Land Data
792 Assimilation System using global river discharge data and a source-to-sink routing
793 scheme, *Water Resour. Res.*, **46**, W06507, doi:10.1029/2009WR007811.

8. TABLES

Variable Name	Abbreviation	Input Predictors
Convective Precipitation	*CNPRCP*	*CNPRCP, PRECTOT, T2M, PS*
Surface Downward Longwave Radiation	*LWS*	*LWS, T2M, PS*
Total Precipitation	*PRECTOT*	*PRECTOT, CNPRCP, T2M, PS*
Surface Pressure	*PS*	*PS, U10M, V10M*
Specific Humidity	*Q2M*	*Q2M, T2M, PS*
Surface Downward Shortwave Radiation	*SLRSF*	*SLRSF, T2M, PS*
Air Temperature	*T2M*	*T2M, U10M, V10M, PS*
U-Wind Component	*U10M*	*U10M, PS*
V-Wind Component	*V10M*	*V10M, PS*

TABLE 1: The nine downscaled variables and their associated input predictors used for multi-variate regression. The level of all variables is at or near the surface.

	May	June	July	Overall
GARD	-0.074	**-0.032**	**-0.089**	**-0.063**
GEOS	**-0.065**	-0.087	-0.090	-0.080

799

800 **TABLE 2**: Brier Skill Scores (BSS) for capturing the probability of a low precipitation event (<=
801 10% percentile) for each month and for the overall forecasting period. Bolded values indicate
802 which forecast (GARD or GEOS) has a better BSS for a given time period. A BSS of 1 indicates
803 perfect skill for a probabilistic forecast.

804

A		CONUS	Central U.S.	Desert	Eastern U.S.	Mountain West	Southern U.S.
D0	F_{GARD}	0.251	0.200	0.212	**0.349**	**0.418**	**0.204**
	F_{GEOS}	**0.278**	**0.231**	**0.350**	0.107	0.189	0.041
D1	F_{GARD}	0.178	0.086	0.380	**0.288**	**0.364**	**0.423**
	F_{GEOS}	**0.231**	**0.165**	**0.408**	0.107	0.192	0.417
D2	F_{GARD}	**0.305**	0.015	**0.741**	**0.334**	**0.382**	**0.371**
	F_{GEOS}	0.302	**0.107**	0.672	0.075	0.329	0.349
D3	F_{GARD}	**0.501**	0.011	**0.914**	**0.435**	**0.453**	**0.343**
	F_{GEOS}	0.355	**0.034**	0.862	0.056	0.266	0.318
D4	F_{GARD}	**0.696**	0.057	**0.911**	**0.262**	**0.671**	**0.253**
	F_{GEOS}	0.316	**0.092**	0.895	0.125	0.275	0.128

B		CONUS	Central U.S.	Desert	Eastern U.S.	Mountain West	Southern U.S.
May	F_{GARD}	**0.864**	**0.607**	**0.603**	**0.756**	**0.768**	**0.732**
	F_{GEOS}	0.718	0.604	0.424	0.278	0.604	0.420
June	F_{GARD}	**0.826**	0.545	**0.852**	**0.746**	**0.556**	**0.567**
	F_{GEOS}	0.816	**0.634**	0.709	0.467	0.426	0.465
July	F_{GARD}	**0.813**	**0.562**	**0.709**	**0.551**	**0.710**	**0.658**
	F_{GEOS}	0.691	0.477	0.493	0.351	0.623	0.319

805

806 **TABLE 3:** Coefficients of determination (r^2) for monthly average percentage of regions in drought
807 determined by root zone soil moisture percentiles for 2003-2017 by (a) category and (b) month.
808 Bolded values indicate which forecast (F_{GARD} or F_{GEOS}) has a stronger correlation (relative to
809 retrospective) for a given region.

810

9. FIGURES

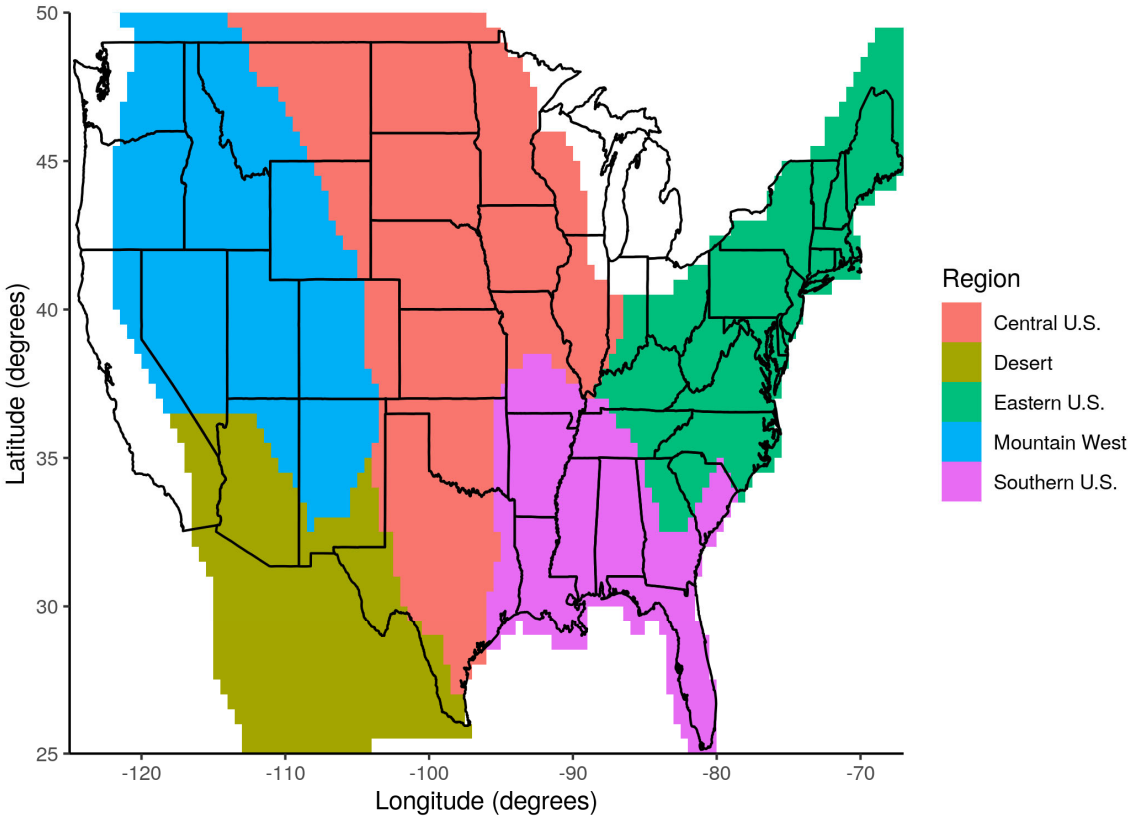
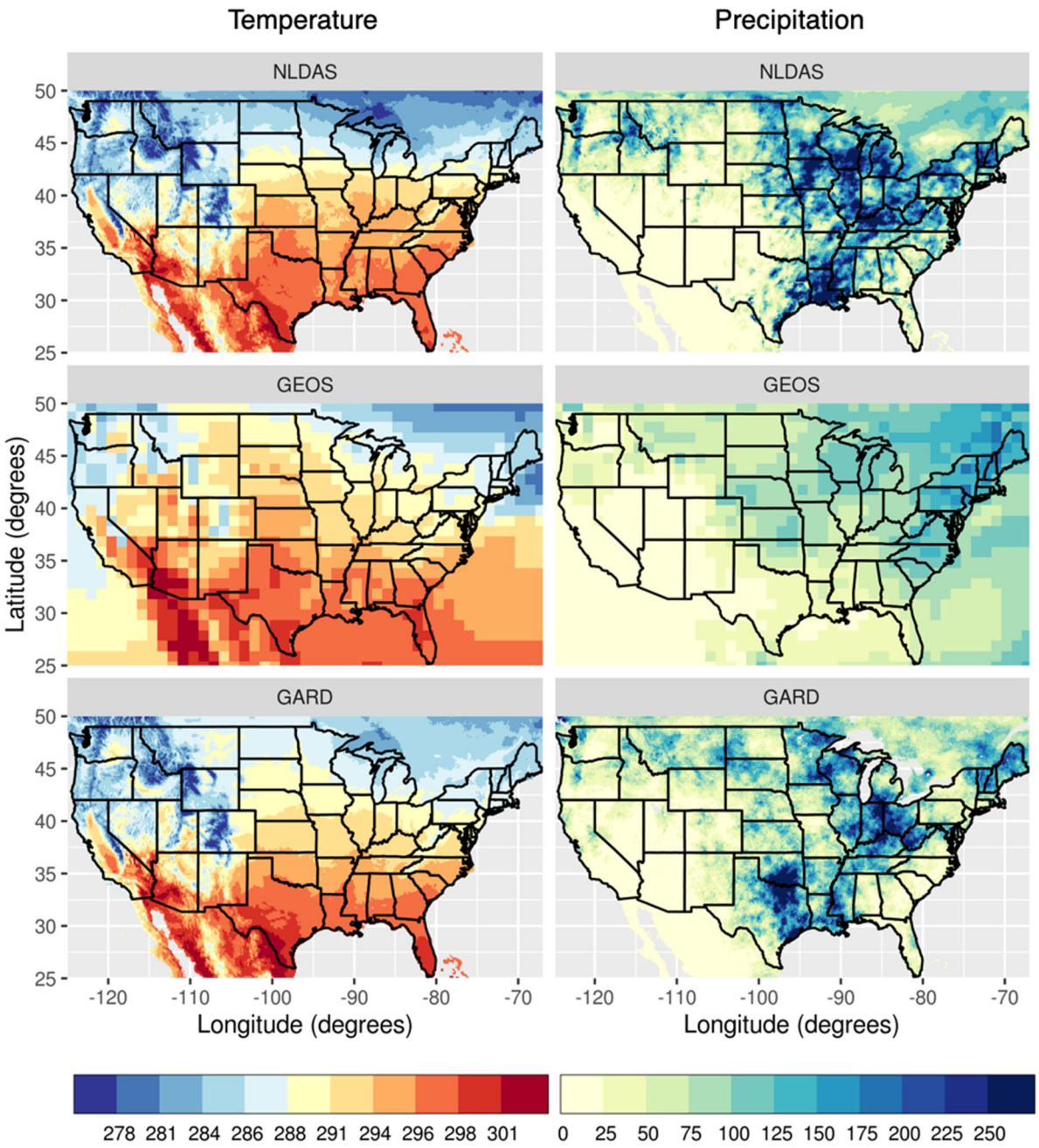


FIGURE 1: Composite Bukovsky regions used to examine changes among similar ecological regions.

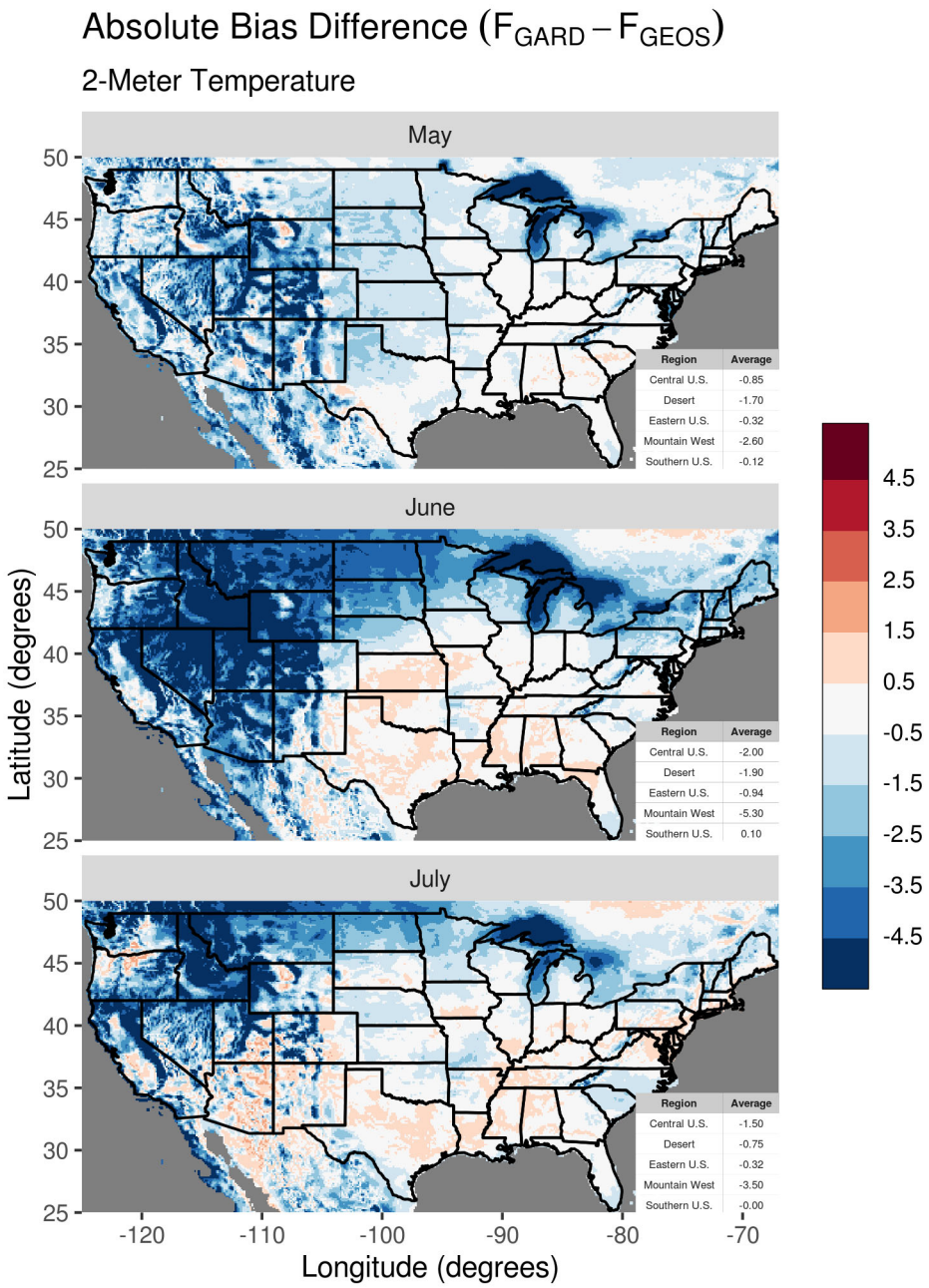


817

818 **FIGURE 2**: Monthly averaged 2-meter temperature (left) and total precipitation (right) from

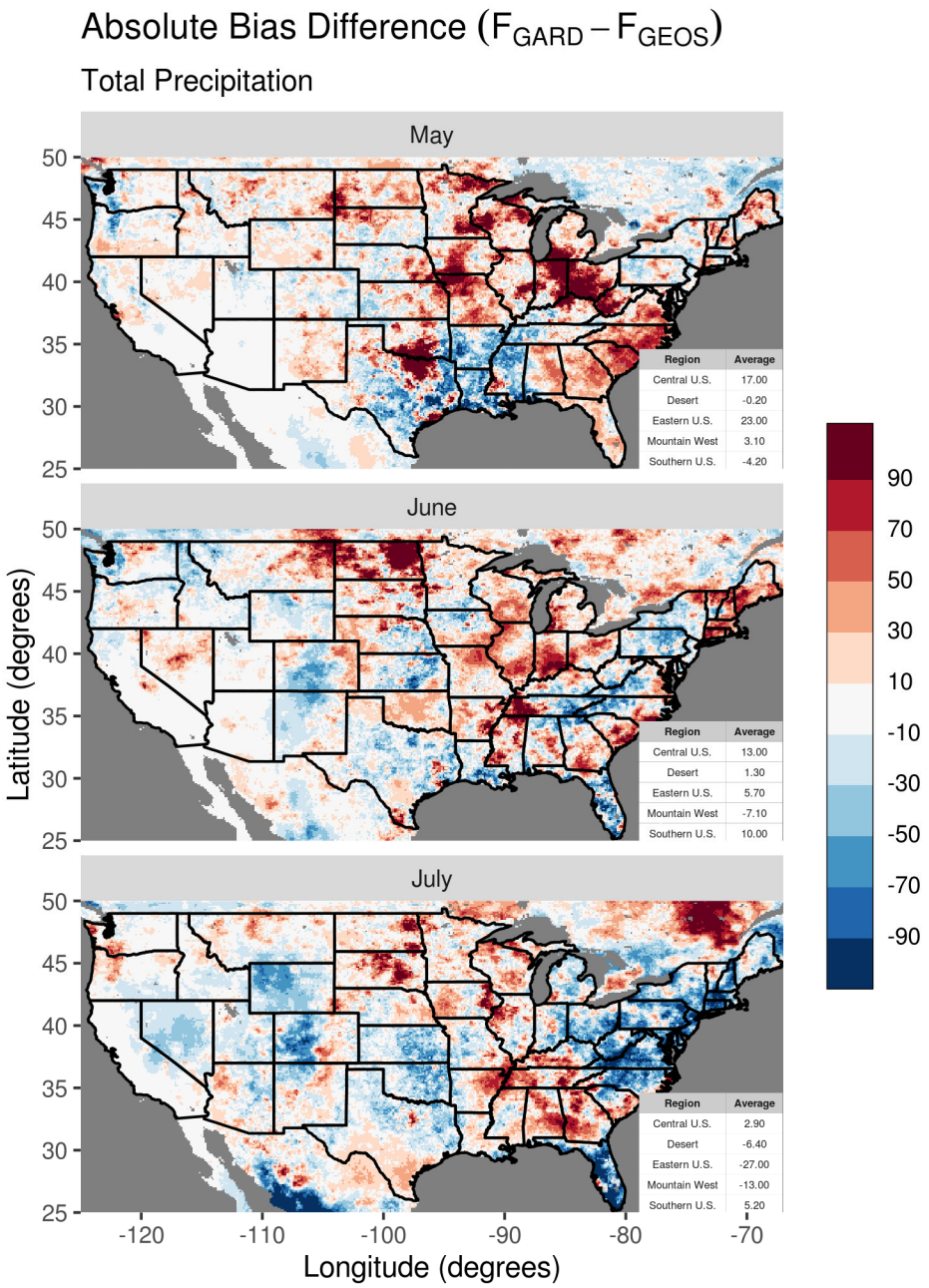
819 NLDAS (top), GEOS (middle), and GARD (bottom) for an example month (May 2004). Units of

820 temperature are shown are in K, and precipitation are shown in mm.



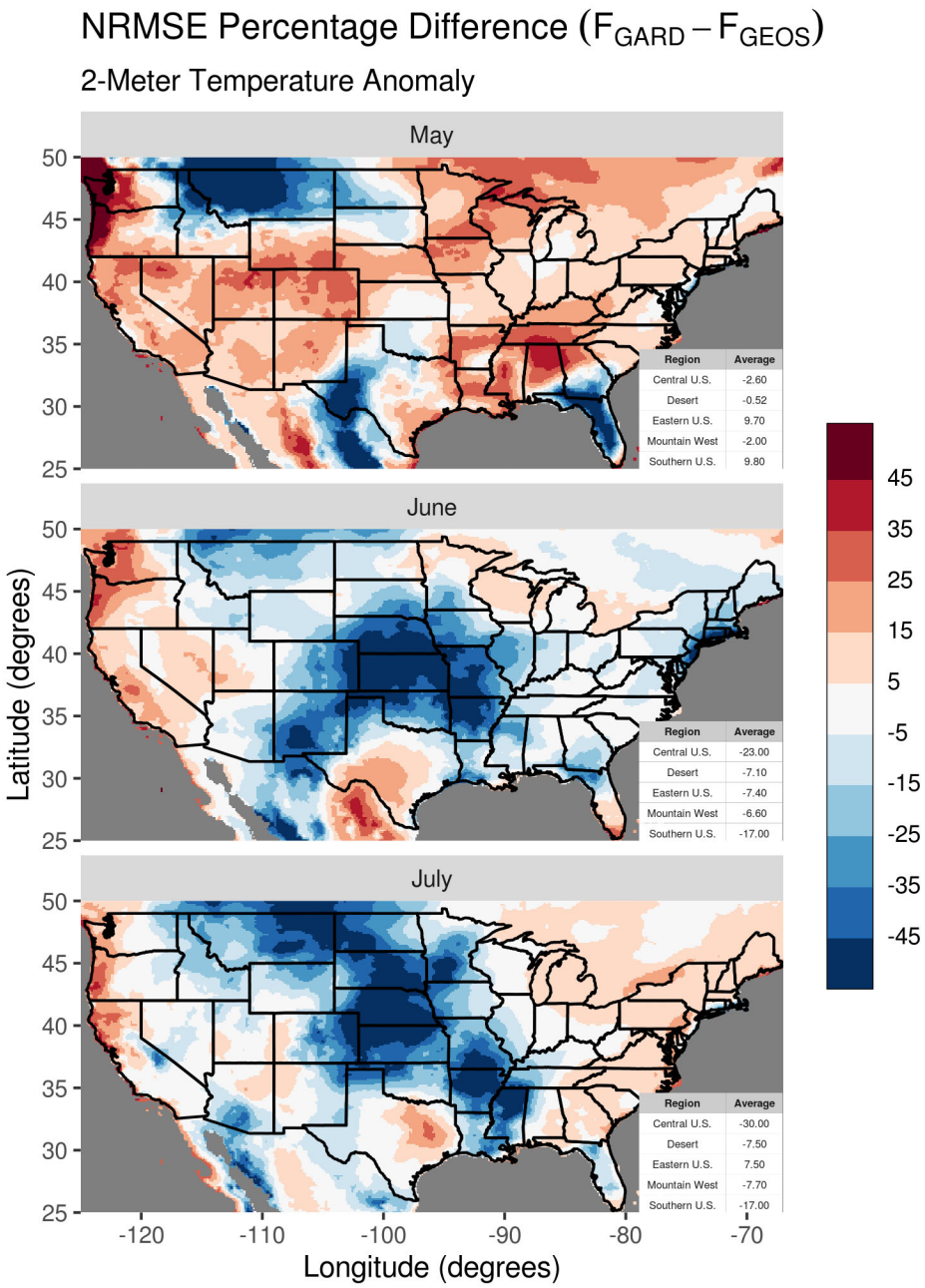
821

822 **FIGURE 3**: Long term mean (2003 – 2017) difference between F_{GARD} and F_{GEOS} in absolute bias
823 of 2-meter temperature for May, June, and July. Bias is calculated for each dataset relative to
824 NLDAS-2. Units of values shown are in K. Inset table shows region averages.



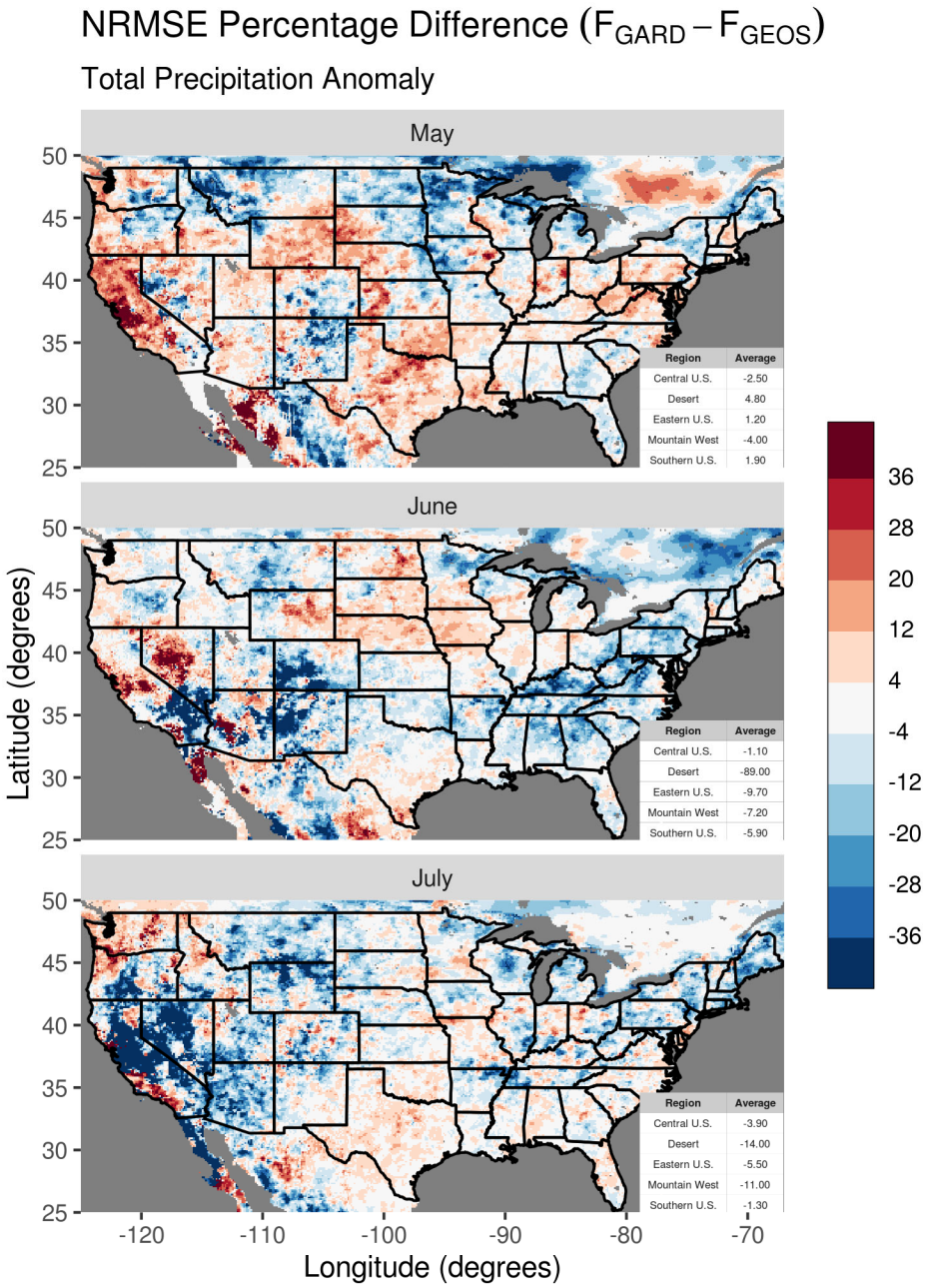
825

826 **FIGURE 4**: Long term mean (2003 – 2017) difference between FGARD and FGEOS in absolute bias
827 of total precipitation for May, June, and July. Bias is calculated for each dataset relative to
828 NLDAS-2. Units of values shown are in mm. Inset table shows region averages.



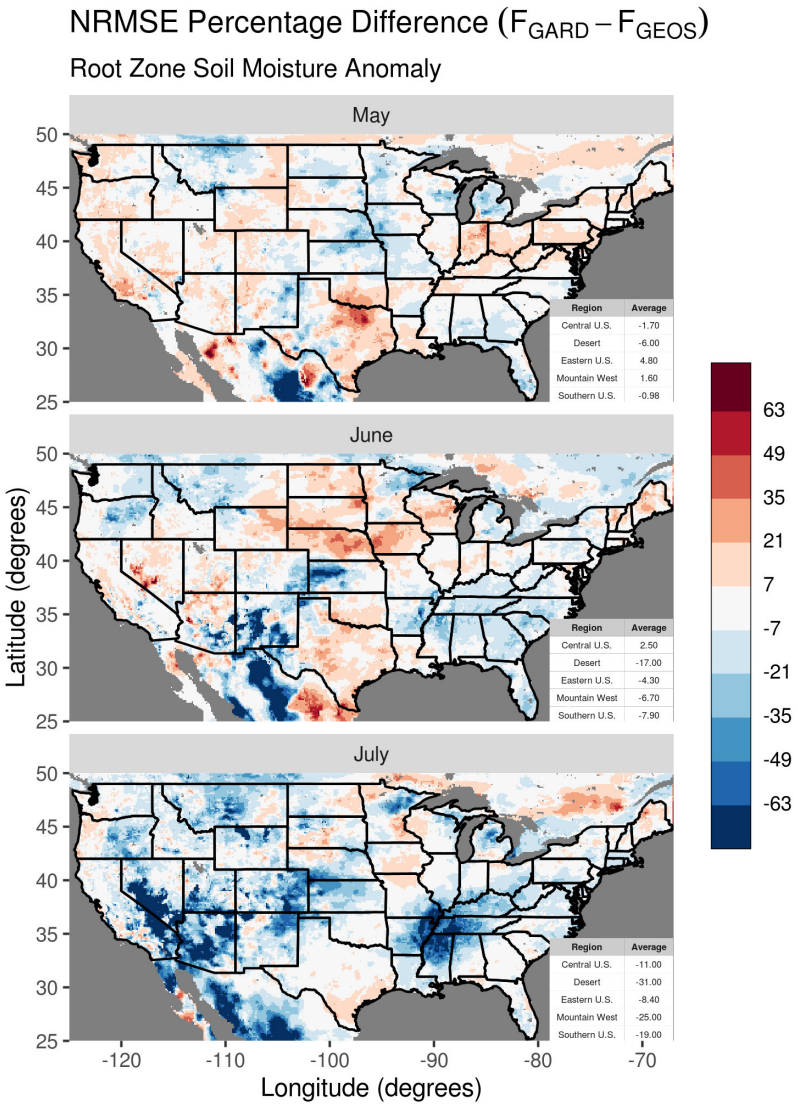
829

830 **FIGURE 5**: Long term mean (2003 – 2017) difference between FGARD and FGEOS in normalized
831 root mean square error (NRMSE) of 2-meter temperature anomaly for May, June, and July.
832 NRMSE is calculated for each dataset relative to NLDAS-2. Units of values shown are in
833 percentage. Inset table shows region averages.



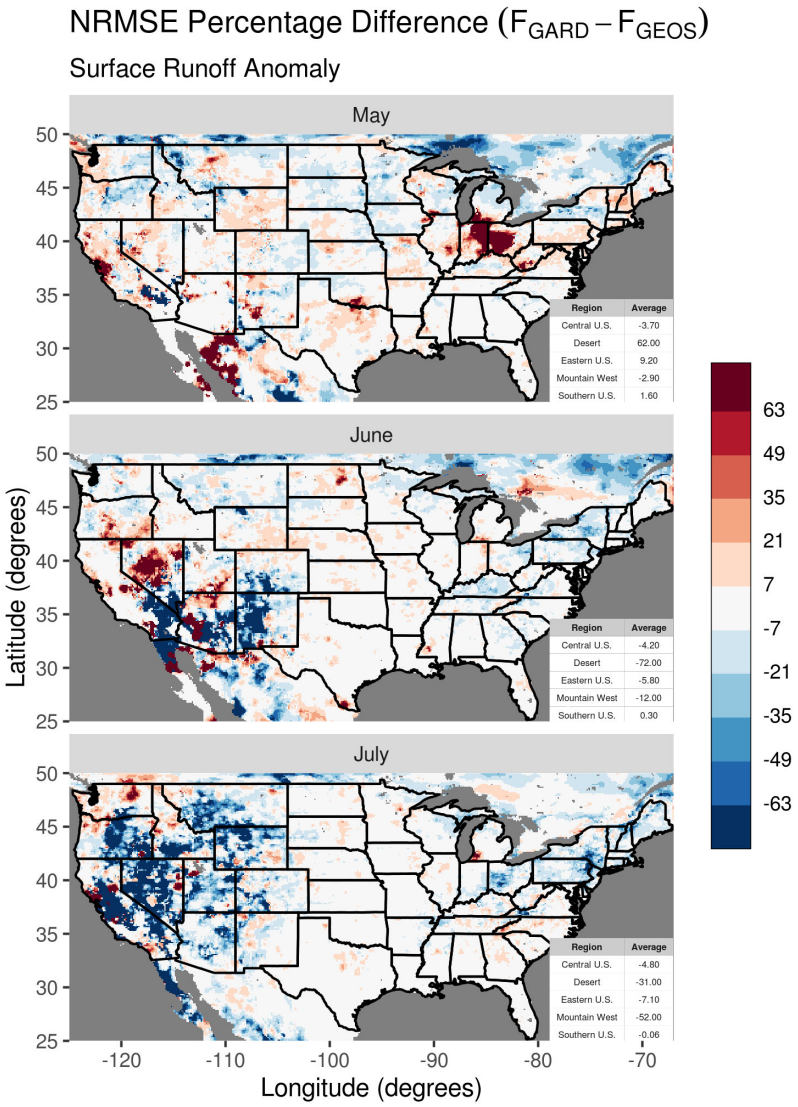
834

835 **FIGURE 6**: Same description as Figure 5 but for total precipitation anomaly.



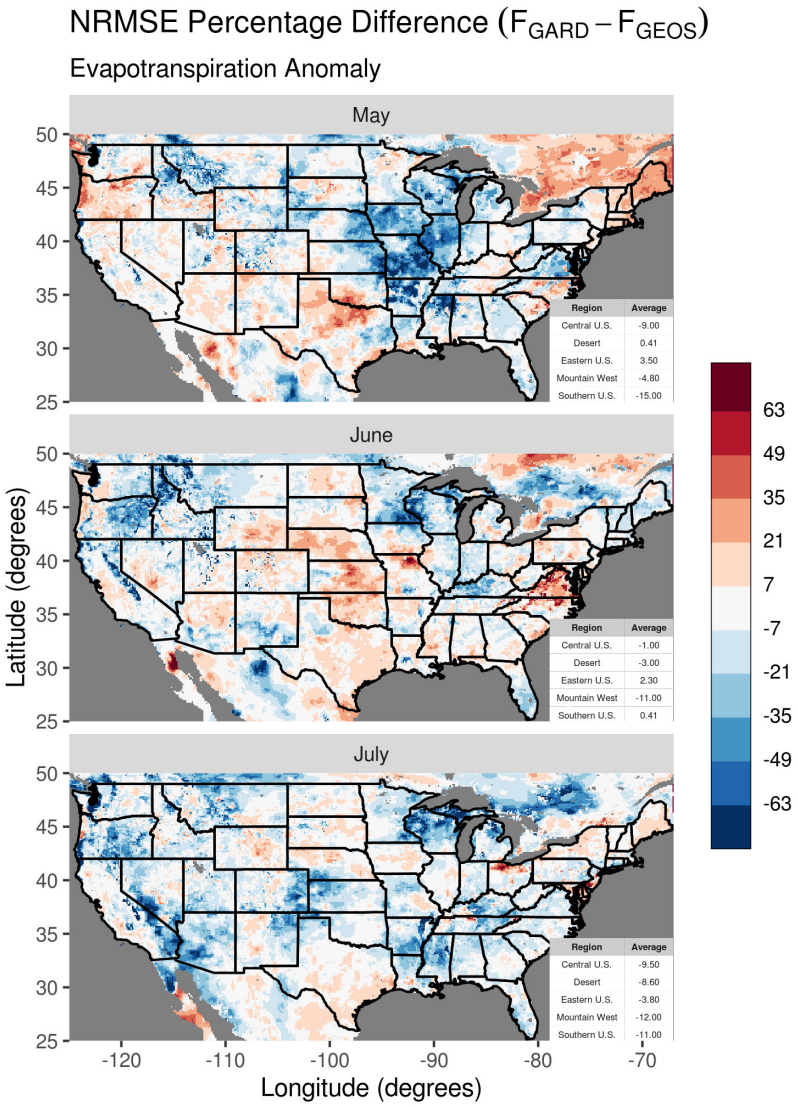
836

837 **FIGURE 7**: Long term mean (2003 – 2017) difference between F_{GARD} and F_{GEOS} in normalized
838 root mean square error (NRMSE) of root zone soil moisture anomaly for May, June, and July.
839 NRMSE is calculated for each dataset relative to retrospective (NLDAS-2 forced LIS/CLSM).
840 Units of values shown are in percentage. Inset table shows region averages.



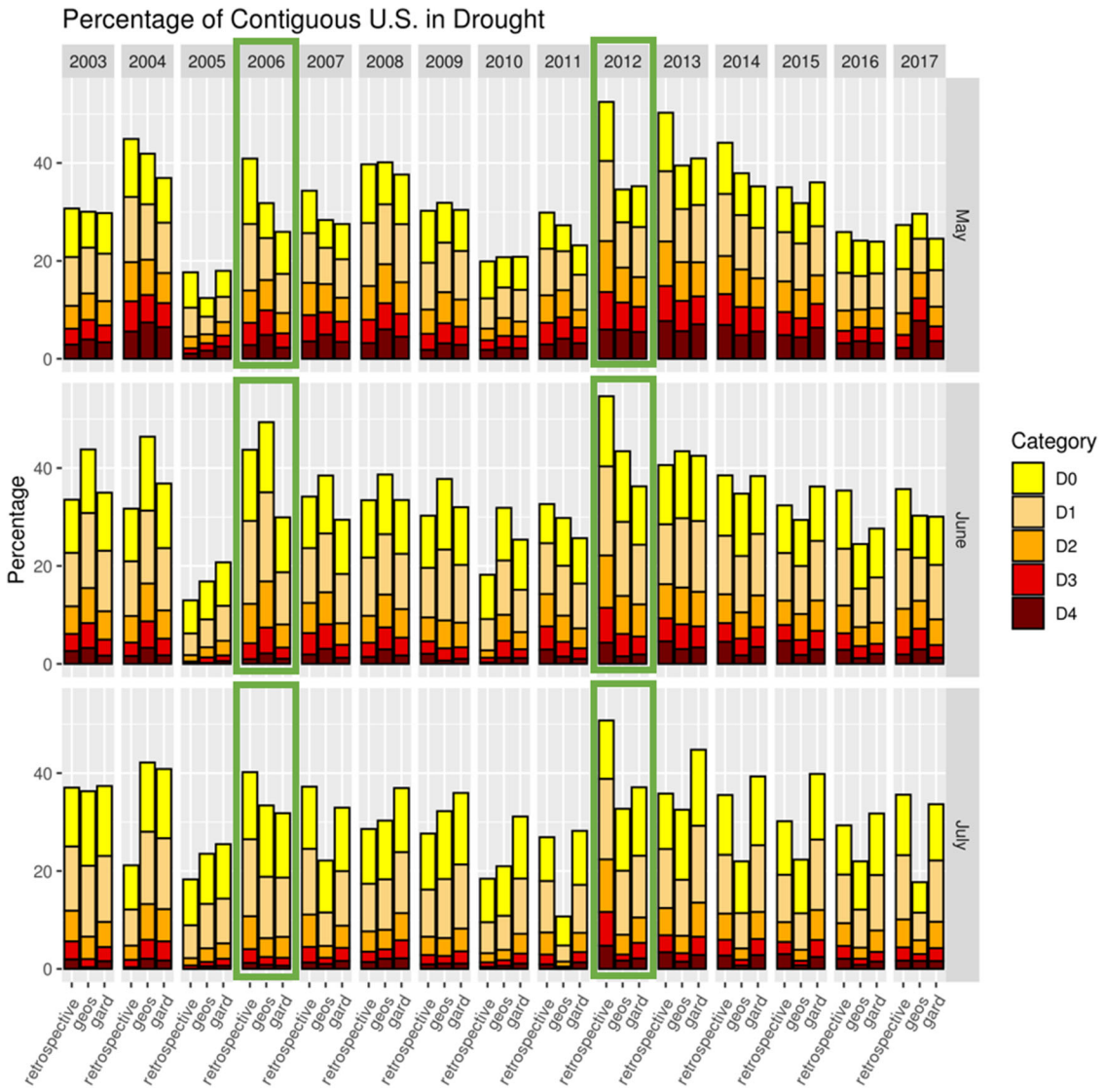
841

842 **FIGURE 8**: Same description as Figure 7 but for surface runoff anomaly.



843

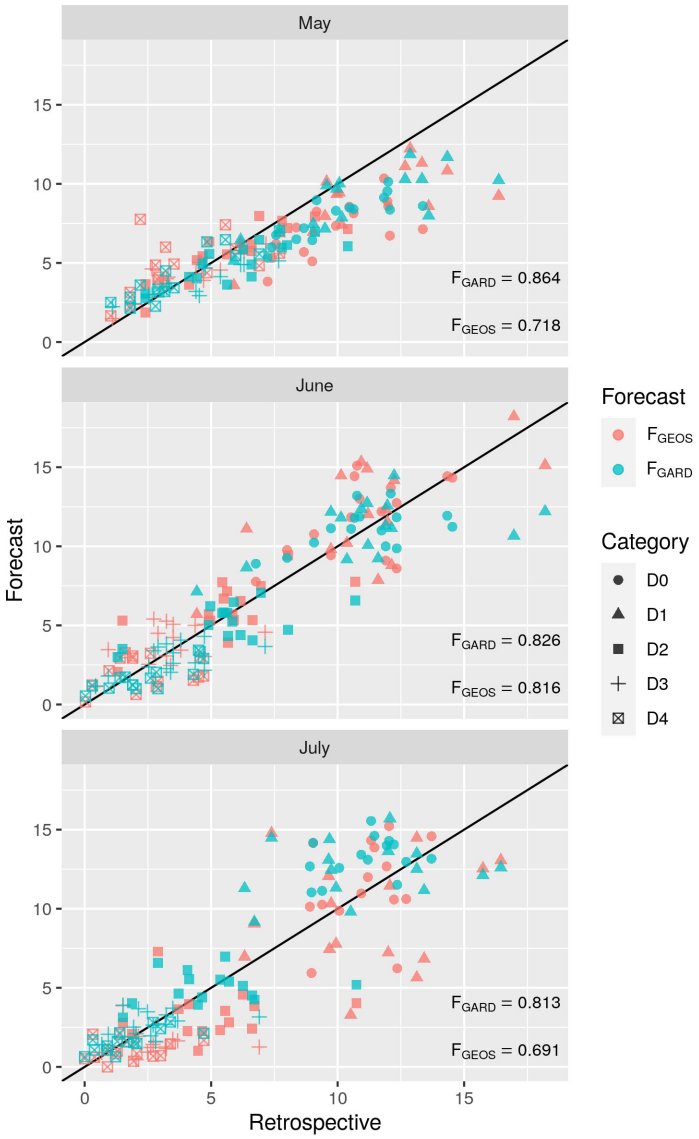
844 **FIGURE 9**: Same description as Figure 7 but for evapotranspiration anomaly.



845

846 **FIGURE 10**: Monthly average percentage of the Contiguous U.S. in drought determined by root
847 zone soil moisture percentiles for 2003-2017 by month (May, June, July). For each year and month,
848 retrospective is the left bar, the FGEOS hindcasts are the middle bar, and FGARD hindcasts are
849 represented by the right bar. Data for 2006 and 2012 is emphasized (green outline).

850



851

852 **FIGURE 11**: Scatter plot of monthly average percentage of the Contiguous U.S. in drought
853 determined by root zone soil moisture percentiles for 2003-2017 by drought category and month.
854 Retrospective values are shown along the x-axis and the associated forecast values are shown on
855 the y-axis for F_{GEOS} (red) and F_{GARD} (blue). Values along the 1:1 diagonal represent strong
856 correlation, and overlaid text shows coefficients of determination (r^2) for each forecast.

857

Category	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	May	June	July
D0	-0.82	-0.56	0.24	-0.29	-0.68	-0.43	-0.47	0.32	-0.69	-0.05	-0.37	-0.81	-0.48	0.14	-2.56	-0.62	-0.66	-0.29
D1	-1.37	-0.47	-0.36	1.45	-1.59	0.01	-0.80	0.39	-1.95	0.32	-0.06	-1.39	1.10	-0.93	-3.46	-0.06	-0.35	-0.89
D2	-0.12	-0.66	-0.01	1.28	-0.63	-0.35	-0.07	-0.48	-0.04	0.04	0.30	-0.65	0.48	-0.94	-1.19	0.54	0.00	-0.68
D3	-0.78	-0.69	0.33	0.30	0.13	-0.18	-0.01	-0.56	0.25	-0.26	-0.08	-0.38	-0.12	-0.23	-0.88	0.24	-0.26	-0.51
D4	-0.37	-0.46	0.49	-0.29	-0.08	-0.67	-0.28	-0.03	0.12	-0.15	-0.71	-0.80	-0.58	-0.49	-2.47	-0.56	-0.51	-0.43
Cumulative	-2.68	-2.26	0.78	5.66	-3.18	0.58	-0.27	0.86	-2.39	0.96	-0.24	-3.36	0.20	-2.88	-7.52	1.42	-1.04	-2.26

858

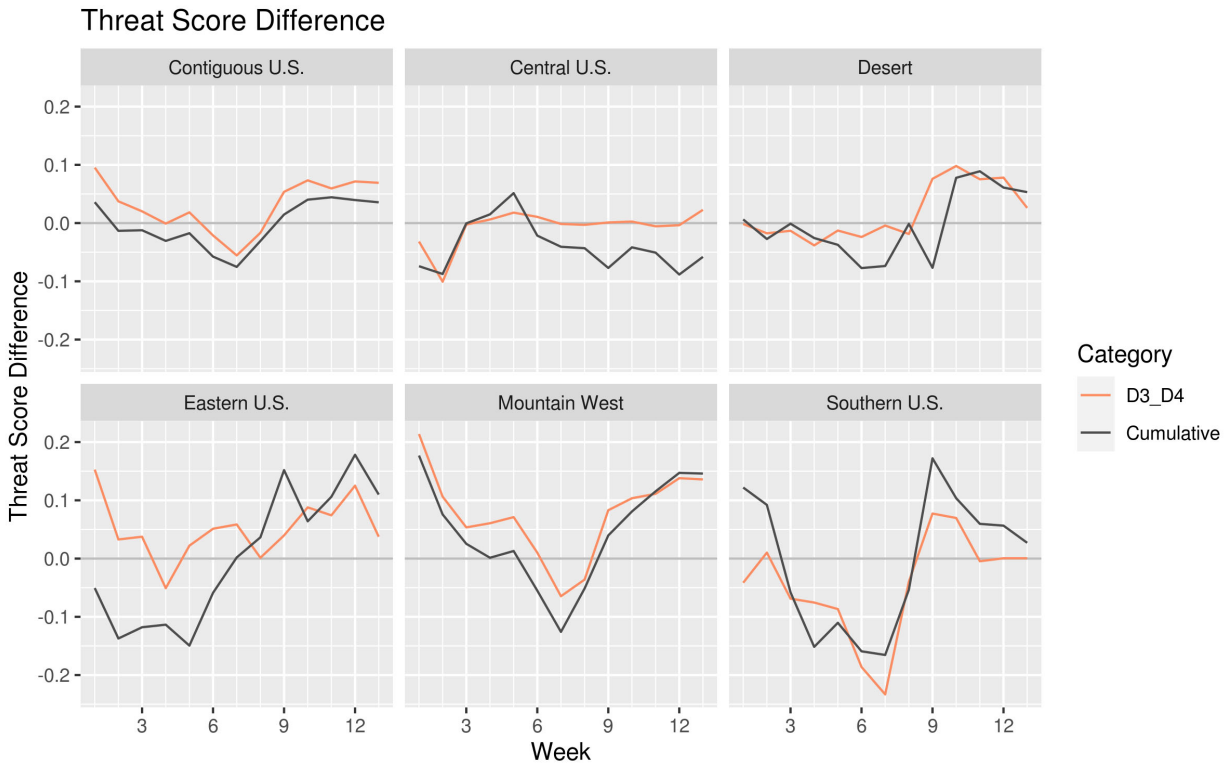
Category	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	May	June	July
D0	-2.80	-2.02	-0.93	-0.86	-0.58	-2.81	-1.35	-0.34	0.04	-3.06	-3.86	-2.05	-3.40	-1.37	-5.12	-0.74	-1.90	-3.02
D1	-3.33	-3.75	-2.78	1.90	-0.88	-2.39	-3.25	1.60	-2.58	-2.02	-1.74	1.05	0.22	-2.25	-7.37	-0.92	-1.18	-3.32
D2	0.50	-4.25	-1.11	1.27	-0.58	-0.56	-0.35	1.77	-0.64	1.32	-0.39	1.63	1.14	-1.12	-2.63	-0.27	-0.35	-0.59
D3	-0.08	-2.60	-0.22	0.56	-0.41	-0.03	-0.12	0.57	-0.43	0.74	-0.14	0.75	0.34	-0.37	-1.14	-0.01	-0.03	-0.41
D4	-0.40	0.34	0.94	-0.07	0.08	0.62	0.10	0.39	-0.12	0.17	-0.08	0.15	1.15	-0.21	-0.25	0.44	0.17	-0.09
Cumulative	-5.05	-11.06	-3.57	4.00	-1.43	-6.18	-4.79	4.14	-5.40	4.43	-3.11	3.35	-0.22	-4.49	-17.80	-2.54	-1.93	-5.58

859

Category	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	May	June	July
D0	0.89	0.59	-0.70	-0.20	-2.25	-2.59	-0.96	-3.20	-3.53	-1.28	1.65	-1.38	0.00	1.36	-3.23	-0.28	-1.31	-1.48
D1	-0.21	0.68	-0.16	0.11	-2.35	-1.66	0.65	-2.90	-3.61	-2.64	3.19	-1.83	3.44	-3.60	2.47	0.08	0.34	-1.51
D2	-1.12	0.27	-0.01	-5.37	-0.37	-1.04	-3.30	-0.41	-2.04	-2.18	0.49	-0.69	1.69	-2.51	-2.08	-1.03	-1.86	-1.27
D3	-1.14	-0.23	-0.31	-5.41	0.21	-2.09	-3.06	-0.65	-1.57	-1.52	-1.91	-0.34	0.66	-1.26	-6.93	-2.13	-2.94	-0.82
D4	0.32	-0.87	-0.83	-6.49	-1.34	-7.39	-3.57	-0.87	-2.73	-3.56	-3.82	-0.75	-2.25	-4.43	-16.03	-8.22	-3.65	-0.03
Cumulative	-1.29	1.45	-0.68	-8.66	-5.07	-9.25	-6.52	-6.48	-11.08	-9.92	2.47	-4.16	10.09	-16.50	-12.86	-4.00	-4.68	-7.20

860

861 **FIGURE 12:** Difference in root mean square error values (F_{GARD} – F_{GEOS} derived daily) of the
862 percentage of area in drought by category for each forecast year and climatological month for the
863 Contiguous U.S. (top panel), Central U.S. (middle panel), and Eastern U.S. (bottom panel). Units
864 of values shown are in root zone soil moisture percentiles. Blue cells are indicative of smaller
865 F_{GARD} RMSE, while red cells indicate smaller F_{GEOS} RMSE.

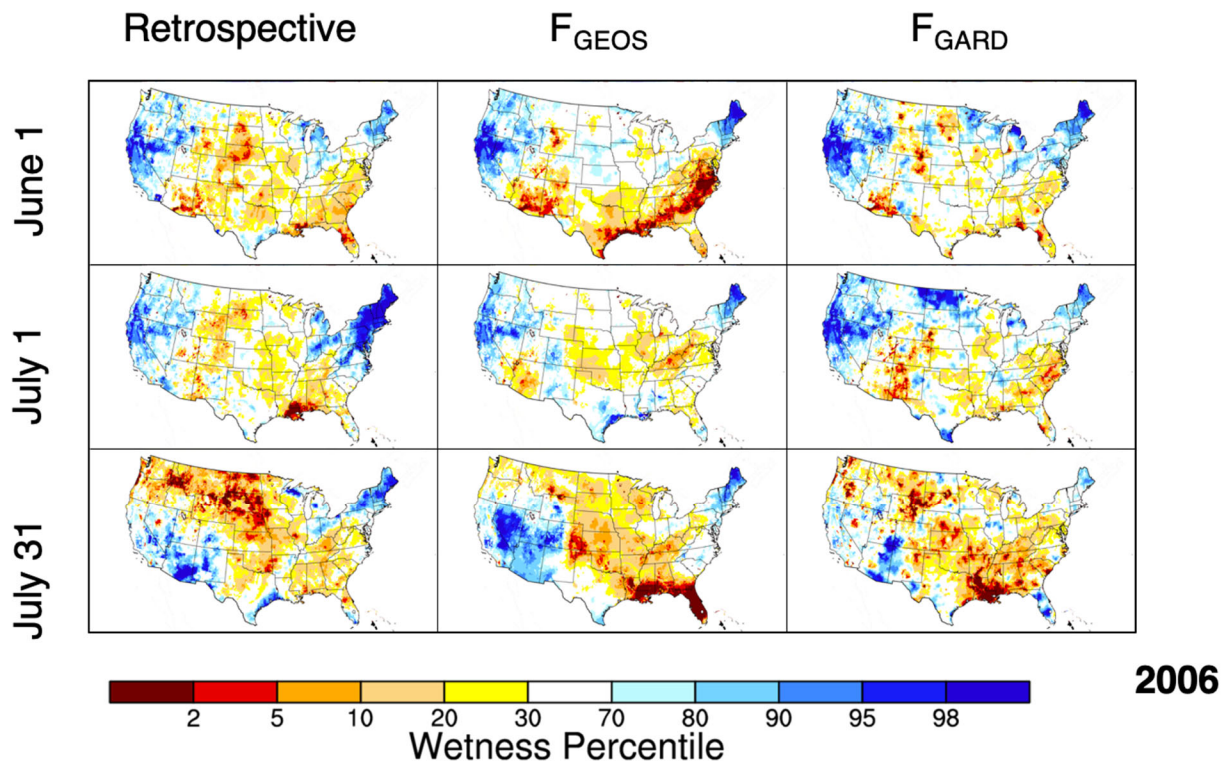


866

867 **FIGURE 13:** Threat Score difference ($F_{GARD} - F_{GEOS}$) in drought severity indices for long-term

868 (2003-2017) root zone soil moisture by week and region. Categories shown are for cumulative (all

869 drought categories) and D3_D4 (extreme and exceptional categories only).



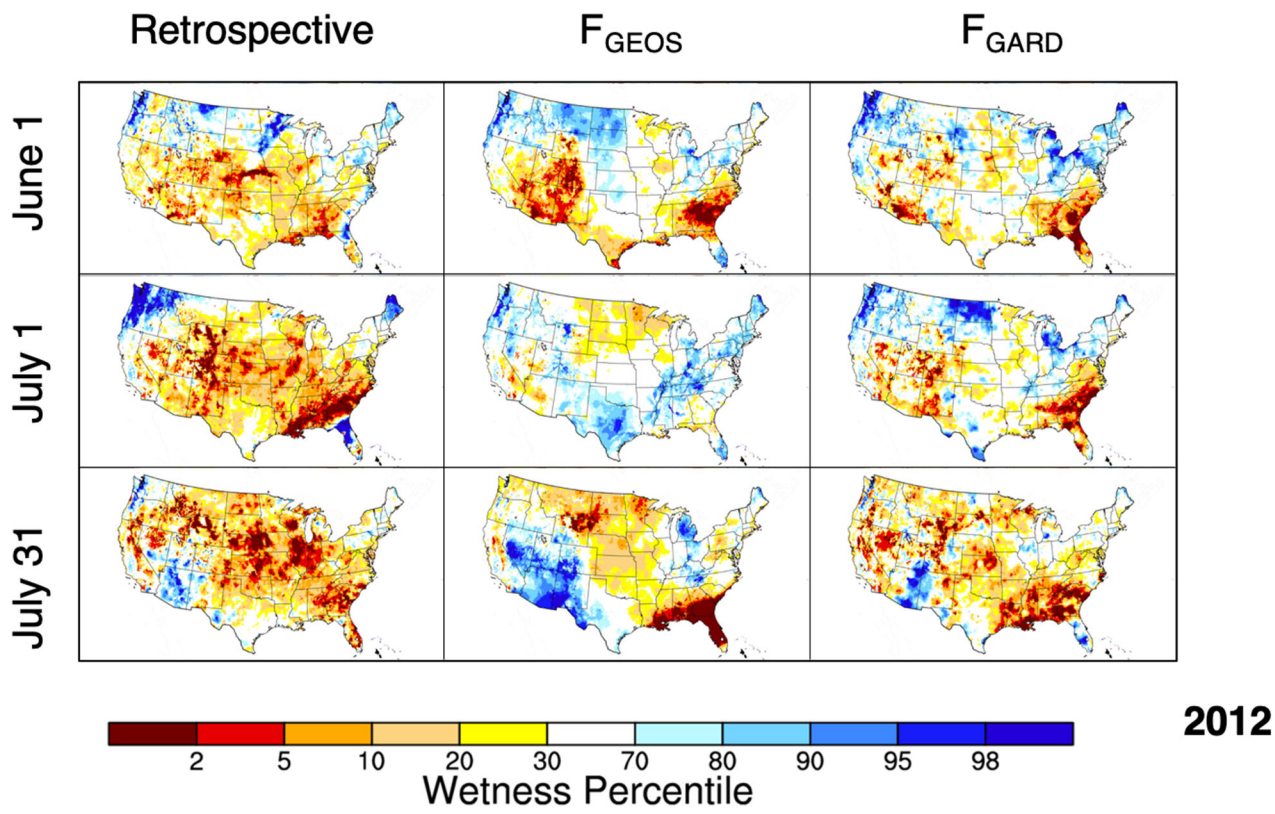
870

871

872 **FIGURE 14:** Spatial drought indicators (root zone soil moisture percentiles) for retrospective,

873 FGEOS, and FGARD datasets. Plots shown for 30, 60, and 90 days after initial May 1st forecast for

874 2006.

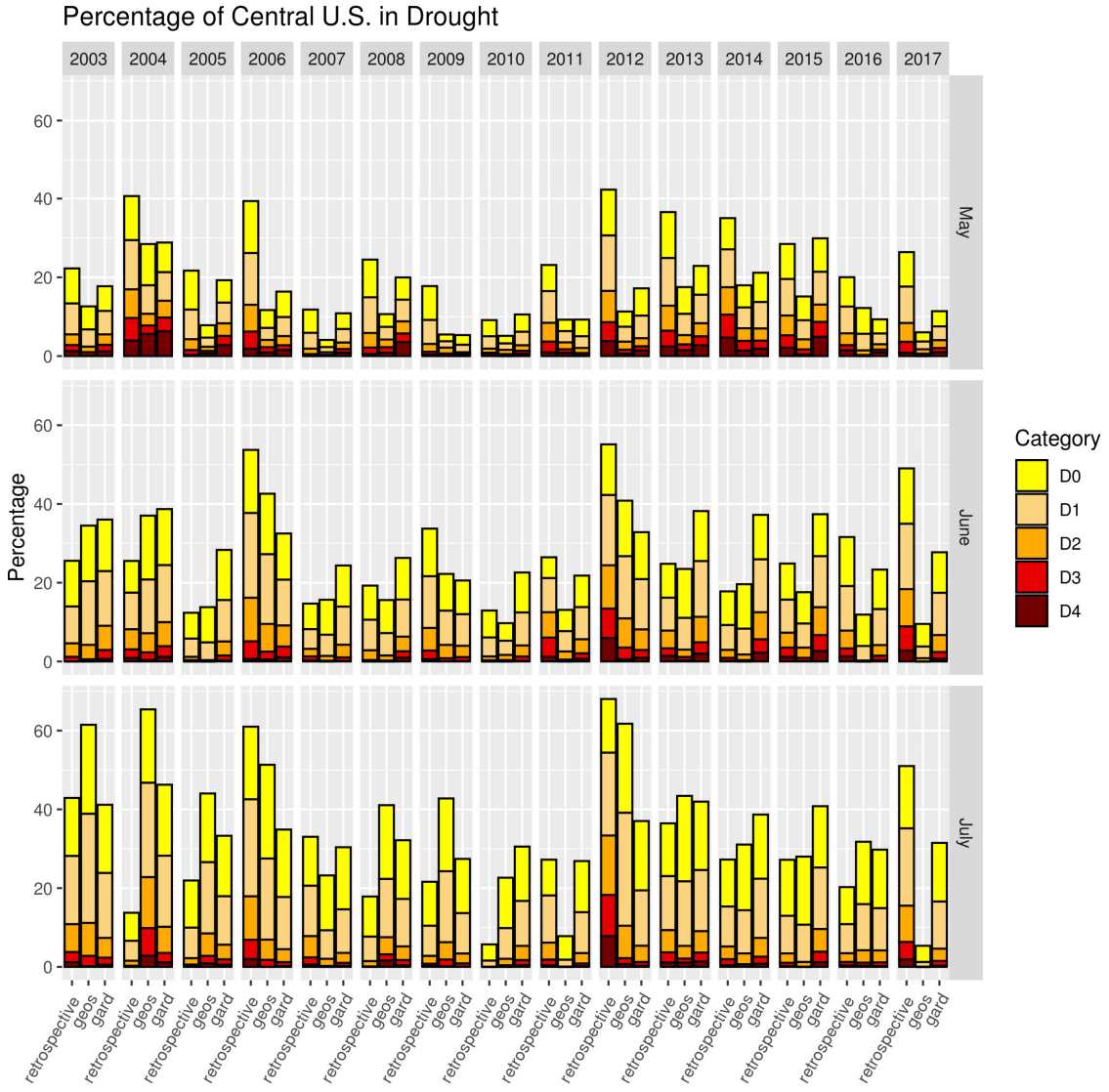


875

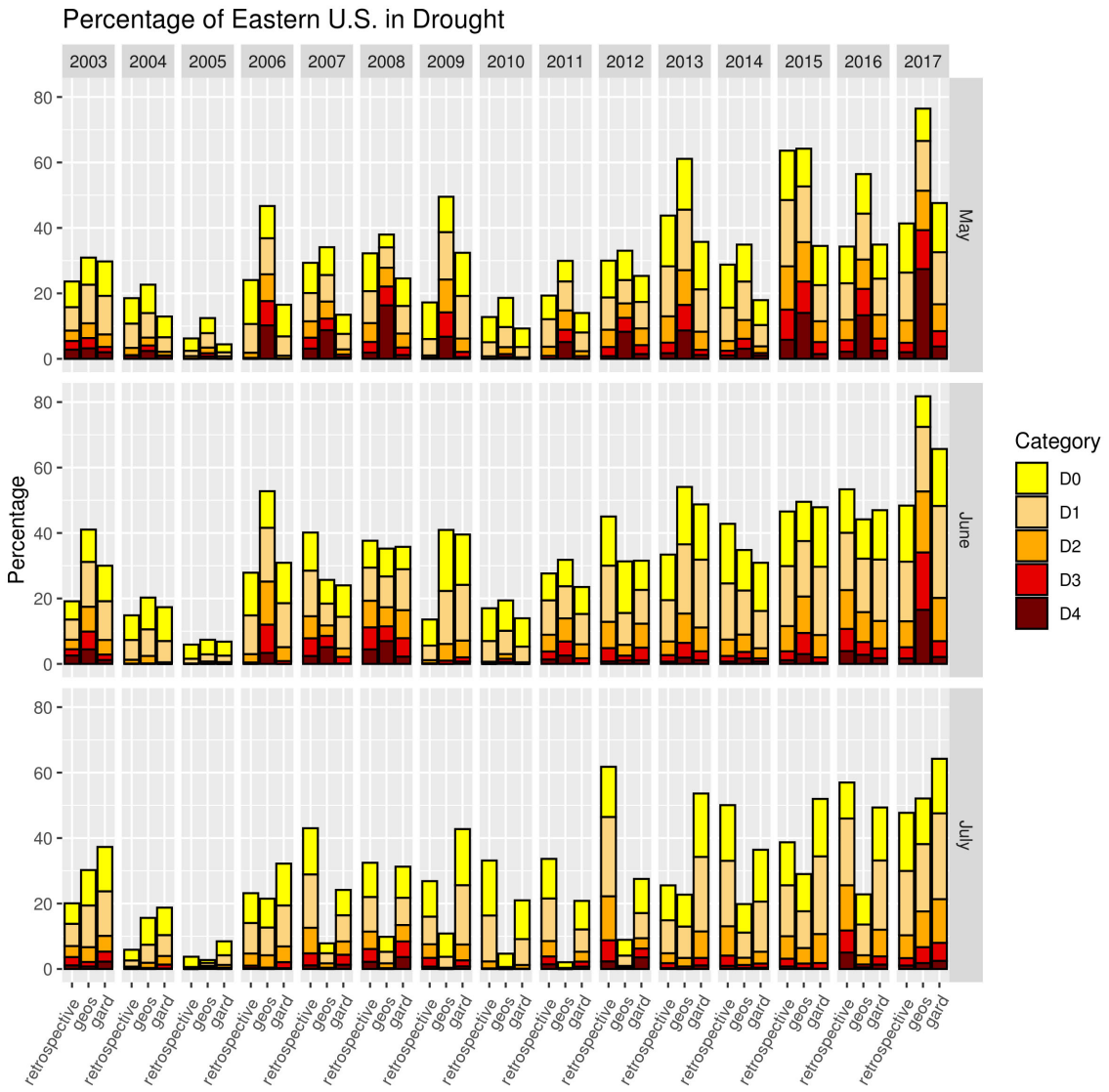
876 **FIGURE 15:** Same description as Figure 15 but for 2012.

877

10. APPENDIX A



APPENDIX A1: Monthly average percentage of the Central U.S. in drought determined by root zone soil moisture percentiles for 2003-2017 by month (May, June, July). For each year and month, retrospective is the left bar, the FGEOS hindcasts are the middle bar, and FGARD hindcasts are represented by the right bar.



884

885 **APPENDIX A2**: Same description as Appendix A2 but for the Eastern U.S.