A Causal Inference Model Based on Random Forests to Identify the Effect of Soil Moisture on Precipitation

LU LI,^a WEI SHANGGUAN,^a YI DENG,^b JIAFU MAO,^c JINJING PAN,^a NAN WEI,^a HUA YUAN,^a
SHUPENG ZHANG,^a YONGGEN ZHANG,^d AND YONGJIU DAI^a

^a *Southern Marine Science and Engineering Guangdong Laboratory, Zhuhai, and Guangdong Province Key Laboratory for Climate Change and Natural Disaster Studies, School of Atmospheric Sciences, Sun Yat-sen University, Guangzhou, Guangdong, China*

^b *School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, Georgia*

^c *Environmental Sciences Division and Climate Change Science Institute, Oak Ridge National Laboratory, Oak Ridge, Tennessee*

^d *Institute of Surface-Earth System Science, Tianjin University, Tianjin, China*

(Manuscript received 11 September 2019, in final form 15 April 2020)

ABSTRACT

Soil moisture influences precipitation mainly through its impact on land–atmosphere interactions. Understanding and correctly modeling soil moisture–precipitation (SM– P) coupling is crucial for improving weather forecasting and subseasonal to seasonal climate predictions, especially when predicting the persistence and magnitude of drought. However, the sign and spatial structure of SM– P feedback are still being debated in the climate research community, mainly due to the difficulty in establishing causal relationships and the high degree of nonlinearity in land–atmosphere processes. To this end, we developed a causal inference model based on the Granger causality analysis and a nonlinear machine learning model. This model includes three steps: nonlinear anomaly decomposition, nonlinear Granger causality analysis, and evaluation of the quality of SM– P feedback, which eliminates the nonlinear response of interannual and seasonal variability and the memory effects of climatic factors and isolates the causal relationship of local SM– P feedback. We applied this model by using National Climate Assessment–Land Data Assimilation System (NCA-LDAS) datasets over the United States. The results highlight the importance of nonlinear atmosphere responses in land–atmosphere interactions. In addition, the strong feedback over the southwestern United States and the Great Plains both highlight the impacts of topographic factors rather than only the sensitivity of evapotranspiration to soil moisture. Furthermore, the SM– P index defined by our framework is used to benchmark Earth system models (ESMs), which provides a new metric for efficiently identifying potential model biases in modeling local land–atmosphere interactions and may help the development of ESMs in improving simulations of water cycle variability and extremes.

1. Introduction

Soil moisture (SM) plays a fundamental role in regulating mass and energy exchange processes in climatological, hydrological, and ecological systems. SM is a particularly important source of water for the atmosphere

through the evapotranspiration (ET) process and, as a result, substantially impacts surface energy processes by influencing the latent and sensible heat flux at the land surface (Seneviratne et al. 2010). SM also controls soil respiration and photosynthesis and therefore affects the soil carbon balance (Davidson et al. 1998, 2000). These effects and feedbacks from SM can have impacts not only on the local scale but also on regional and global scales; thus, SM plays a significant role in climate change projections (Hirschi et al. 2011). SM anomalies can persist for months and are indicators of extreme climate events (Trenberth and Guillemot 1996; Fischer et al. 2007), moreover, this relationship suggests that assessing the impact of SM variation on the climate, especially on precipitation P , which is the most relevant aspects in

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JHM-D-19-0209.s1>.

Corresponding author: Wei Shangguan, shgwei@mail.sysu.edu.cn

DOI: 10.1175/JHM-D-19-0209.1

© 2020 American Meteorological Society. For information regarding reuse of this content and general copyright information, consult the [AMS Copyright Policy \(www.ametsoc.org/PUBSReuseLicenses\)](#).

climate systems, is very important for predicting climate change and extreme events (Santanello et al. 2018), which may reduce the loss caused by disasters and enhance our ability to make climate projections.

The sign and pattern of the soil moisture–precipitation (SM– P) feedback are still under debating. Previous studies have indicated that this feedback varies by location. However, a consistent determination of the sign and pattern of the impact at the local scale is hard to obtain, even by analyzing the same data (Findell and Eltahir 1997; Salvucci et al. 2002). Most studies have indicated that the SM– P feedback has a positive sign (Pal et al. 2000; Koster et al. 2006), though a few studies have suggested the opposite (Cook et al. 2006; Wei et al. 2008; Taylor et al. 2011). For the “hot spot” regions (i.e., locations of strong feedback), most studies have found strong feedbacks are located in transitional zones, however, these studies have also disagreed on the magnitude of the SM– P coupling strength (Koster et al. 2004; Lawrence and Slingo 2005; Guo et al. 2006). On the contrary, by using atmospheric reanalysis and offline hydrologic model outputs for the most recent 36-yr data, Song et al. (2016) indicated that there was no significant feedback in the southern Great Plains, which is a major transition zone in the United States. Tuttle and Salvucci (2016) reached similar conclusions by carrying out Granger causality (GC) analysis.

Physical and statistical models are commonly used to investigate SM– P feedback. Physical modeling is used to simulate SM– P feedback according physical process. However, it has been suggested that the land–atmosphere feedback is not yet accurately represented in physical models (Guo et al. 2006; Hohenegger et al. 2009). In addition, convective parameterization schemes and spatial scale could significantly influence both the sign and the patterns of the feedback. We, therefore, do not know whether the simulated responsiveness of the atmosphere reflects reality, or it is just an artifact of the models (Koster et al. 2003). Thus, physical models used to analyze SM– P feedback should be considered carefully, and benchmark analysis is necessary.

In static modeling, the SM– P feedback is usually defined as the lagged or simultaneous correlation between SM and P (Zeng and Yuan 2018). However, linear correlation essentially includes the effects of the long-term periodic terms (e.g., sea surface temperature) and P autocorrelation in addition to the local SM– P feedback itself, and it is difficult to identify the causal SM– P relationship in the highly interconnected climate system (Wei and Dirmeyer 2012). This difficulty is partly because SM has the ability to influence land–atmosphere processes over time, and the persistence of SM anomalies could affect the coupled feedback (Shukla and

Mintz 1982; Rowell and Blondin 1990; Douville 2002). Additionally, soil accumulates and retains current and past P , which makes it challenging to identify the causal relationship between SM and P . It is also difficult to determine whether the investigated feedback is the real causal relationship or only a result of the temporal autocorrelation in the P . Tuttle and Salvucci (2016) developed a linear GC framework to quantify the surface SM effect on the next-day P , which eliminates the issue of temporal autocorrelation in P . By constructing a regression relationship between the effects beyond land–atmosphere coupling (i.e., annual cycles, seasonal cycles, lagged P , and lagged atmosphere pressure) and precipitation occurrence (POCC) using an empirical statistical model (i.e., generalized linear model, GLM), the authors eliminated the long-term and persistent effects from POCC, and the GC analysis was used to estimate the local causal relationship between SM and POCC. Contrary to physical models, this statistical model can directly construct the relationship between past data (e.g., SM in this study) and target variables (e.g., next-day P), making it a feasible and convenient approach to investigate the causal relationship.

However, the statistical method used in Tuttle and Salvucci (2016) was too simplistic to represent the complex nonlinear land–atmosphere feedback due to the assumptions of linearity. Additionally, the memory effect is spatially different (Wei and Dirmeyer 2012); thus, the proper lag time used to represent the persistence effect is different in each pixel. In addition, the traditional feature selection method used to avoid overfitting is computationally expensive and is not appropriate for nonlinear models.

Based on this information, we highlighted the five important aspects involved in exploring the causal relationship between local SM– P feedback.

a. The importance of the nonlinear atmosphere response to periodic terms

Local climate variables are time dependent and exhibit different patterns in their trends and periodicity. Several decomposition methods have been proposed in previous studies, including the simple two-step decomposition method, multiplicative models, and breakpoint method (Cleveland et al. 1990; Grieser et al. 2002; Papagiannopoulou et al. 2017). However, these methods have focused only on eliminating seasonality (i.e., 12-month cycle) and have ignored longer time scale oscillations (e.g., quasi-biennial oscillation), which can significantly influence the short-term analysis of P (Stuecker et al. 2013).

Moreover, these methods considered only the linear response to periodic terms rather than the nonlinear

response. The importance of the nonlinear atmosphere response to periodic terms must be emphasized. [Krakauer et al. \(2010\)](#) indicated that P variance has a nonlinear response to the combined interannual variability of sea surface temperature and SM. [Stuecker et al. \(2013\)](#) used observational and climate model data to highlight the importance of the nonlinear atmospheric response to the annual cycle of sea surface temperature. They found that the quadratic nonlinear process allowed the El Niño–Southern Oscillation (ENSO) to interact with the seasonal and interannual cycles of sea surface temperature, which affected precipitation.

b. The memory effect of the atmosphere terms

The memory effect is a critical aspect to consider in the SM– P feedback and has been difficult to define in previous studies. Some other studies have accounted for persistence indirectly by using past P data to separate the persistence ([Guillod et al. 2014](#)) or by examining the covariance of possible confounding variables ([Seneviratne et al. 2010](#)). Few works have directly accounted for the memory effects using statistical causal identification methods. [Tuttle and Salvucci \(2016\)](#) developed a method to address this issue by applying the lagged variables to represent the atmosphere and P persistence from synoptic weather systems and constructed vector autoregression (VAR) to separate it from daily P by using a GLM. However, this method considers only the linear relationship between persistence and P , which is an obviously nonlinear process.

c. The optimal lag window length for persistence terms

Environmental dynamics reveal their effects on P at differing temporal scales (e.g., seasonal and subseasonal scales), and the memory of soil and atmosphere is spatially different. Previous studies have indicated that dry regions may have longer memory than other regions ([Dirmeyer et al. 2009](#)). These studies also indicated that lag times depend on both the specific climatic controlling variable and the characteristics of the system. Previous studies using the GC analysis usually empirically set the length of the lag window, which may lead to the generation of artificial signs and patterns of the land–atmosphere feedback ([Tuttle and Salvucci 2016](#); [Papagiannopoulou et al. 2017](#)).

d. The feature selection method to avoid overfitting

In the machine learning model, redundant independent variables have a higher possibility of overfitting, which requires the selection of the “best” combination of nonredundant features from all explanatory variables.

Some studies avoid this issue, which may lead to unreliable results because of the lower generalization ability in the overfit model. Other studies have utilized all possible linear regression methods and stepwise methods to address this issue; however, it is infeasible and unrealistic to use a nonlinear model due to the much higher computational cost.

e. Endogeneity of the regression

Endogeneity will arise if independent variables are correlated with error terms. It represents a bias in the regression model ([Tuttle and Salvucci 2016](#)). It comes from several aspects, including that confounding variables is omitted, the independent variable is measured with errors, or the dependent variable is jointly influenced with the independent variable ([Tuttle and Salvucci 2016](#)).

Endogeneity could raise bias in the regression model and also change the statistical significance of the coefficients ([Tuttle and Salvucci 2016](#)). A previous study applied the bootstrapping method to bias-correct the coefficient of SM in ordinary least squares (OLS) regression ([Tuttle and Salvucci 2016](#)). They bootstrapped samples and generated lots of regressions with these samples, and got the mean of coefficient of independent variables in sample regressions, then used it to bias correct the coefficient of original regression. However, the nonlinear method in machine learning cannot provide coefficient of each feature, and thus the method to bias correct the nonlinear regression is still open for discussing. Recently, [Ghosal and Hooker \(2018\)](#) developed the one-step boost random forest for bias correction.

The main objective of this study is to integrate and improve some existing causal inference methods ([Tuttle and Salvucci 2016](#); [Papagiannopoulou et al. 2017](#)) and get a complete causal inference framework, which will provide a useful method for estimating the causal relationship between SM and P . From this, we could identify the accurate sign and pattern of SM– P feedback. The hot spot spatial distribution defined by our model was used to benchmark Earth system models (ESMs), which helps to improve the parameterization scheme over the land surface processes. This model includes the following three main features:

- 1) Instead of employing linear methods, a nonlinear machine learning algorithm [i.e., random forest (RF)] is introduced to represent the complex and nonlinear relationship between the explaining (i.e., periodicity, trending, and persistence terms) and explained (i.e., POCC) variables.
- 2) The out-of-bag error derived from RF regression is used to select the proper memory time (i.e., lag

window length) and to evaluate the impact of P and pressure memory.

- 3) A hybrid feature selection method is proposed to improve the computational efficiency and avoid the overfitting issue.

This paper is outlined as follows. In the second section, we first introduced the dataset used in this study, and then we showed the theory of the GC analysis and demonstrated the explaining power of nonlinear models over linear models using a simple nonlinear system. Furthermore, the causal inference model based on the nonlinear GC analysis was introduced. The results are shown in relation to the following four aspects in the third section:

- 1) We found evidence that the nonlinear framework could significantly increase the explanatory power compared to the linear framework.
- 2) The sign and hot spot of SM–POCC feedback over the United States were introduced.
- 3) We reinforced our results by applying different datasets of SM and P .
- 4) We applied the model output into the proposed framework and used the SM– P index to benchmark the models in CMIP6. Concluding remarks are presented in the final section.

2. Data and methodology

a. Data description

The National Climate Assessment–Land Data Assimilation System (NCA-LDAS; Kumar et al. 2019) dataset, which contains the surface volumetric SM content (surface 0–5 cm), P , and surface pressure data, was used to identify the short-term SM– P feedback over the contiguous United States in this study. NCA-LDAS is derived from land surface hydrologic modeling with multivariate assimilation of satellite environmental data records (EDRs), including the Scanning Multichannel Microwave Radiometer (SMMR), the Special Sensor Microwave Imager (SSM/I), the Advanced Microwave Scanning Radiometer (AMSR-E and AMSR-2), the Advanced Scatterometer (ASCAT), and the Soil Moisture Ocean Salinity (SMOS). The core of NCA-LDAS is the multivariate assimilation of past and current satellite-based data records within the Noah version 3.3 land surface model (LSM) at 0.125° (approximately 10 km) resolution using NASA’s Land Information System (LIS) software framework during the Earth observation satellite era. This dataset covers 42 variables, including land surface fluxes, stores, states, and routing variables, which are driven by atmospheric forcing data from the

North American Land Data Assimilation System Phase 2 (NLDAS-2). The daily version of this dataset was used in this study. Previous studies indicated that the surface SM in these datasets showed significant improvements over regions such as the Great Plains and Arkansas Red and Lower Mississippi basins, with degradations over the Southwest, where the land–atmosphere feedback is still intensively being debating (Kumar et al. 2019). Thus, this dataset is suitable for our study to investigate the causal SM–POCC feedback.

The investigated domain of this study was restricted to the contiguous United States (25° – 53° N, 67° – 150° W). All variables were spatially remapped at a $0.25^\circ \times 0.25^\circ$ grid size using bilinear interpolation, with a total number of 25 984 grids. All datasets used in this study were averaged to a daily time step, spanning from 19 June 2002 to 19 June 2011. A total of 3287 daily data points was obtained for each grid cell, which we believe is enough to identify the short-term land–atmosphere dynamics at the climatology scale (Tuttle and Salvucci 2016).

b. Granger causality overview

1) LINEAR GRANGER CAUSALITY ANALYSIS PROCEDURE

GC is used in this study to develop the causal relationship between SM and P , which is an operational definition of causality (Granger 1969), and the method has been widely used in fields such as econometrics, medicine, and meteorology (Roebroek et al. 2005; Brovelli et al. 2004). GC is a measure of the directional influence of two variables and is based on the assumption that if a variable causally affects another, then the past values of the former should be helpful in predicting the future values of the latter (Granger 1969). In the following, we summarized the steps of the GC for the case of two stationary and scalar-valued time series, that is, $\mathbf{X} = [x_1, x_2, \dots, x_N]$ and $\mathbf{Y} = [y_1, y_2, \dots, y_N]$, with the index N denoting the length of the series. We assumed that $\mathbf{Z} = [z_1, z_2, \dots, z_N]$ represents variables that could influence the causal relationship between $\mathbf{X}$ and $\mathbf{Y}$.

The linear GC analysis with multivariables includes the following three steps. First, the past i information of $[x_t, y_t, z_t]$ is defined as $\mathbf{M}_{t-i} = [x_{t-i}, y_{t-i}, z_{t-i}]^T$, where T means transpose. The p th-order polynomials of the linear VAR model were constructed to predict y_t from $\mathbf{M}_{t-i}$ as follows:

$$y_t = \sum_{i=1}^p \Phi_F(i) \mathbf{M}_{t-i} + \varepsilon_t, \quad (1)$$

where t is the time step and $\Phi_F(i)$ is the coefficient matrix of past information of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$. Equation (1)

incorporates all available information ($\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$), and we call Eq. (1) the “full” model.

Second, we exclude the past information of $\mathbf{X}$, and past i time step information of $[y_t, z_t]$ is defined as $\mathbf{N}_{t-i} = [y_{t-i}, z_{t-i}]^T$. Additionally, the p th order of the VAR model is calculated as

$$y_t = \sum_{i=1}^p \Phi_B(i) \mathbf{N}_{t-i} + \epsilon_t, \quad (2)$$

where Φ_B is the coefficient matrix of past information of $\mathbf{Y}$ and $\mathbf{Z}$. Equation (2) considers only the past information of $\mathbf{Y}$ and $\mathbf{Z}$; thus, we call Eq. (2) the “baseline” model.

Finally, we define the null hypothesis H_0 of the GC analysis as follows: $\mathbf{X}$ has no influence on $\mathbf{Y}$ conditioned on $\mathbf{Z}$, that is, the full model has equal or less predictive power as that of the baseline model. Alternatively, H_0 is rejected only if the full model predicts the target variable (e.g., P) better than the baseline model, which means that the past information of $\mathbf{X}$ is helpful in predicting the future value of $\mathbf{Y}$, that is, $\mathbf{X}$ Granger causes $\mathbf{Y}$, which is denoted as $\mathbf{X} \rightarrow \mathbf{Y}$.

2) NONLINEAR GRANGER CAUSALITY ANALYSIS PROCEDURE

Because a linear relationship might not be unsuitable for highly nonlinear land–atmosphere feedback processes, we used a machine learning algorithm, that is, RF, to investigate the feedback. RF is capable of handling a mix of different types of independent variables and is not vulnerable to model parameters (Breiman 2001). By building many regression trees with randomly selected input variables, RF is more sensitive and robust to noise and outliers than when an individual regression tree is used. The output is based on the entire set of regression trees, and the results of all trees are averaged or weighted (Van Looy et al. 2017). Thus, RF is superior over regression trees because of its ability to overcome the problem of overfitting. The proposed nonlinear GC analysis includes the following three steps.

First, the nonlinear VAR model of order p was constructed as

$$y_t = \Psi(\mathbf{H}_{t,p}) + \mu_t, \quad (3)$$

where t is the time step; Ψ represents the RF model; μ_t is the noise term; and $\mathbf{H}_{t,p}$ represents the past p time step information of $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{Z}$ and is the $3 \times p$ input-dependent matrix of RF, which is calculated as

$$\mathbf{H}_{t,p} = \begin{bmatrix} x_{t-1} & \cdots & x_{t-p+1} & x_{t-p} \\ y_{t-1} & \cdots & y_{t-p+1} & y_{t-p} \\ z_{t-1} & \cdots & z_{t-p+1} & z_{t-p} \end{bmatrix}. \quad (4)$$

We also called Eq. (3) the full model.

Second, the past information of $\mathbf{X}$ was excluded, and the nonlinear VAR model of order p was calculated as

$$y_t = \Psi(\mathbf{l}_{t,p}) + \tau_t, \quad (5)$$

where t is the time step; Ψ represents the RF; τ_t is the noise term; and $\mathbf{l}_{t,p}$ represents the past p information of $\mathbf{Y}$, $\mathbf{Z}$ and is the $2 \times p$ input-dependent matrix of RF, which is calculated as follows:

$$\mathbf{l}_{t,p} = \begin{bmatrix} y_{t-1} & \cdots & y_{t-p+1} & y_{t-p} \\ z_{t-1} & \cdots & z_{t-p+1} & z_{t-p} \end{bmatrix}. \quad (6)$$

Equation (5) is also designated the baseline model.

Finally, the corresponding out-of-bag error from RF regression is obtained from Eqs. (3) and (5). If the error of Eq. (5) is greater than that of Eq. (3), this means that the past information of $\mathbf{X}$ is helpful for predicting the future value of $\mathbf{Y}$, that is, $\mathbf{X}$ Granger causes $\mathbf{Y}$.

We applied three main machine learning methods (e.g., support vector machine regression, artificial neural network, RF) to contrast the ability of different nonlinear regression models in our study. We selected RF because of its better performance and lower calculation cost applying on large volume of data. In this study, we applied one-step boost random forest (Ghosal and Hooker 2018) to address the endogeneity problem. We run two RF regressions for each regression model, that is, we established an original RF and extracted the residuals and then fit another RF to these residuals. We used the RF regression with 100 trees, and the maximum number of predictor variables per node was equal to one-third of the total number of predictor variables. This setting is based on the observations that any changes in these parameters do not cause substantial changes in the results. We selected lag window length using method in nonlinear Granger causality analysis described in section 2c.

3) IDENTIFYING THE NONLINEAR CAUSAL RELATIONSHIP USING DIFFERENT APPROACHES

The land–atmosphere system is a highly complex nonlinear system, and it is difficult to define the system using several precise equations. Thus, in this section, we constructed a simple nonlinear system to illustrate the ability of the nonlinear GC model compared with the linear GC model. Two stationary time series were used, that is, $\mathbf{Q} = [q_1, q_2, \dots, q_M]$ and $\mathbf{R} = [r_1, r_2, \dots, r_M]$, with M denoting the length of the time series. The nonlinear causal relationship between $\mathbf{Q}$ and $\mathbf{R}$ are constructed as

FIG. 1. The variation of nonlinear Granger causality (GC) value of $\mathbf{R} \rightarrow \mathbf{Q}$ with nonlinear index α increased from 0 to 1. Nonlinear GC value is defined by machine learning models and Granger causality analysis. Parameter α is a nonlinear impact index that indicates the magnitude of the nonlinear impact of $\mathbf{R}$ on $\mathbf{Q}$. GK, GLM, BP, and RF denote Gaussian kernel model, generalized linear model, back propagation neural network, and random forest, respectively. Two dashed lines are the results of the two linear models (i.e., GK, GLM), and the solid lines indicate the results of two nonlinear models (i.e., BP, RF). We added different error terms that obey a Gaussian distribution with different means (μ) and variations (σ) shown in each title.

$$q_{t+1} = (1 - \alpha)(1 - cq_t^2) + \alpha(1 - cr_t^2) + \tau_t, \quad (7)$$

$$r_{t+1} = (1 - cr_t^2) + \mu_t, \quad (8)$$

where $t = 1, 2, \dots, 1000$; $c = 1.8$; and μ_t and τ_t are the error terms that obey Gaussian distribution. Further, we conducted multiple experiments using different Gaussian noise terms to assess the sensitivity of the proposed approach on model noise. Before applying GC analysis, we normalized these two time series to create stationary time series. From Eq. (8), it is clear that r_{t+1} depends only upon the past information of $\mathbf{R}$ (i.e., r_t), which indicates that time series $\mathbf{R}$ has a Markov property; furthermore, from Eq. (7), q_{t+1} depends on the past information of both $\mathbf{Q}$ (i.e., q_t) and $\mathbf{R}$ (i.e., r_t). Parameter α is defined as a nonlinear impact index that indicates the magnitude of the nonlinear impact of $\mathbf{R}$

on $\mathbf{Q}$. The nonlinear causal effect of $\mathbf{R}$ on $\mathbf{Q}$ become larger as the nonlinear impact index α increases from 0 to 1.

To investigate the ability of identifying the nonlinear causal relationship of utilizing linear and nonlinear GC models, we used four machine learning models and applied the GC analysis (see section 2b) to identify the causal effect of $\mathbf{R}$ on $\mathbf{Q}$ ($\mathbf{R} \rightarrow \mathbf{Q}$). These four models included two linear models (i.e., Gaussian Kernel (GK) model and probit model of GLM) and two nonlinear models [i.e., back-propagation (BP) neural network and RF].

Figure 1 shows how the GC values of $\mathbf{R} \rightarrow \mathbf{Q}$ vary with α from 0 to 1 with different error terms. As one might expect, the GC values increase with the nonlinear impact index α for the RF and BP models (solid line) in all four experiments, which coincide with the theoretical

FIG. 2. Diagram of the nonlinear Granger causality framework. Parameter f is the bias-corrected random forest regression method, “lag” is lag window length of regression, and c is a constant. The augmented Dickey–Fuller test in step 1 is used to test whether the time series is stationary. SM–POCC impact in step 3 is soil moisture–precipitation occurrence feedback.

settings of the nonlinear system; however, this result is not observed for the two linear models, that is, GK and GLM (dashed lines). As variance and mean value increases, curve of RF presents larger increase than neural networks, which indicated that RF is more sensitive to the causal relationship. The result of RF shows a more stable increasing trend than ANN generally. And ANN has values around 0 or even negative values when the feedback is week, which indicated that ANN cannot capture the relative week nonlinear feedback.

This test suggests that the linear models cannot simulate the true causal nonlinear relationship of simple systems. Thus, linear models are not appropriate for complex systems, such as the Earth system. Moreover, RF performs better in detecting nonlinear GC value than ANN. Therefore, it is preferable to use RF in the GC analysis to estimate the causal relationships among variables.

c. Methodology

The methodology of the proposed causal inference model based on the nonlinear GC analysis and RF is shown in Fig. 2, and the method includes the following three steps: nonlinear anomaly decomposition, nonlinear GC analysis, and evaluation of the quality of SM–POCC feedback. The nonlinear impact of the long periodic terms, memory on formatting P were eliminated (see

steps 1, 2 in Fig. 2); therefore, the short-term local causal relationships between SM and POCC were isolated (Tuttle and Salvucci 2016). From this, we qualified the causal effect of SM on POCC from this causal inference model (see step 3 in Fig. 2).

1) NONLINEAR ANOMALY DECOMPOSITION

Time-dependent series data are nonstationary and unsuitable for direct application in a GC analysis. Therefore, we resorted to anomaly analysis by removing trends and periodicity from raw series before analyzing the causal relationship of SM– P feedback.

Here, we adopted the Fourier method proposed by Tuttle and Salvucci (2016), with continuous sinusoid variables that vary on interannual (18–1.8-yr periods) and seasonal (12–2.4-month periods) temporal scales to represent periodicity and trends, which were used to account for internal atmospheric and climatic influences on P (e.g., sea surface temperature and seasonality). Both the sine and the cosine terms were included for each given oscillation period. The continuous sinusoid variables of each period (sine and cosine terms) were expressed as $[\sin(2\pi t/l), \cos(2\pi t/l)]$, with l denoting the length of the period and t denoting the time step. This method considers both interannual and seasonal cycles, and the given period oscillations have been proven to be sufficient to characterize the variability in the precipitation data.

Additionally, regarding the nonlinear impact of periodic terms, we applied RF regression to fit POCC with continuous variables (i.e., interannual and seasonal terms), as stated above (see step 1 in Fig. 2). The dependent variable was POCC, and we used a binary variable to represent it, with the rainy day setting as 1 and the no rain day as 0. This step eliminated the nonlinear response to long periodic terms (e.g., quadratic nonlinear) in precipitation, allowing us to emphasize the short-term local land–atmosphere dynamics and applying Granger causality analysis further.

Following the anomaly decomposition, the augmented Dickey–Fuller (ADF) test was used to check whether the residual time series was stationary. Pixels that did not pass the stationary test were omitted from the following analysis.

2) NONLINEAR GRANGER CAUSALITY ANALYSIS

After the nonlinear anomaly decomposition, we carried out the following four procedures (shown as step 2 in Fig. 2) before performing the nonlinear GC analysis. These four procedures included temporal memory variable construction, lag window selection, a hybrid method of feature selection, and nonlinear Granger causality analysis, which are explained in detail below:

(i) *We constructed the independent variables that represent the memory effect of precipitation and pressure*

In this study, we considered the lagged P and lagged pressure as the persistence of P and atmosphere, respectively, and RF regression was employed to eliminate the persistence effects from the P , which represented nonlinear autocorrelation in the synoptic weather system.

Furthermore, we also incorporated the spatial autocorrelations at a given pixel by extending the explaining variables (i.e., lagged P and lagged surface pressure) in our models (Papagiannopoulou et al. 2017), but found the performance of the regression models changed a little. From this, we assumed that spatial effect of the precipitation and pressure is little, thus we ignored it in the following study.

(ii) *The optimal length of the lag window was selected by the following step*

After constructing the temporal memory variables, we built a regression between deperiodicity POCC and these variables to eliminate the nonlinear impact of memory terms. However, the lag windows length in these variables could significantly influence the performance of the regression, thus we selected the optimal

length of the lag window in this step. First, several different lag window lengths were employed, and the corresponding nonlinear regression models (i.e., RF) were constructed. Model performance was evaluated by utilizing the out-of-bag error of RF. The out-of-bag error is an unbiased estimate of the generalization error of RF by testing MSE in out-of-samples in each subtrees. It approximates k -fold cross validation, which requires a lot of calculations in nonlinear regressions. Second, the optimal lag window length was selected based on the minimum out-of-bag error, and the RF regression model with optimal lag window length was further employed to evaluate the effect of memory terms.

Our preliminary results indicated that lags greater than 35 days did not add extra predictive power for short-term feedback; thus, we sliced the lag window length from 5 to 35 days, with a spacing of 5 days.

(iii) *We developed a new hybrid method of feature selection to avoid overfitting*

In this study, we developed a hybrid feature engineering method that combined the optimal features by considering both linear and nonlinear relationships between the target and explaining variables. First, we calculated Pearson coefficient between each single explaining variable with target variable on whole datasets rather than the train sets (see the supplemental material for the explanation), which consider the linear relationships. Only the explaining variables that were most linearly related to the target variable (i.e., high coefficient and pass the significant test) were considered. Second, all explaining variables were then used to fit the target variables using RF regression, and a resulting vector was obtained, which contained the importance of the explaining variables. The resulting vector is the average of the mean squared error (MSE) of all trees in the RF regression. Likewise, the highest nonlinear correlation between the target and explaining variables were obtained according to the preceding resulting vector. Third, we combined the predictors selected by the two steps above and constructed the “optimal” RF model using selected variables to fit target variables. Furthermore, to avoid overfitting and underfitting, the ratio between the number of selected explaining variables and the number of original explaining variables was empirically set to range from 20% to 80%, with a spacing of 10%. From this, we run several optimal RF regressions by changing the selected variables using different ratios, and the “best” model was selected based on the minimum out-of-bag error value.

After applying the steps in sections 2c(2)(i)–(iii), we got the best regression between memory terms with target variable, and we eliminated the impact of these variables by the best RF model.

(iv) Nonlinear Granger causality analysis

The nonlinear GC analysis was applied to identify the pixels with significant SM–POCC feedback. After eliminating the impact of periodicity, trends, and the persistence terms from the preceding steps, the residual SM and POCC were obtained. RF regression was employed to fit the relationship between residual SM and residual POCC and to estimate the causality between SM and POCC. Full and baseline models based on the nonlinear GC analysis procedure were then constructed. Specifically, the model based on the fitting of residual SM and residual POCC was called the “full model,” while the full model that excluded the SM from the independent variables was denoted as the “baseline model” (see [section 2c](#)).

In the GC analysis, the null hypothesis was that SM had no influence on future rainfall events. This hypothesis was rejected only when SM still improved the prediction of future POCC after the effects of atmospheric periodicity and trends; atmospheric persistence and P persistence on POCC and SM were removed in the previous steps. We further set the minimum error improvement threshold equal to 0.01 to remove the situation that the difference between these models is infinitesimally small. The performance of the baseline model relative to the full model was compared using the out-of-bag error from RF regression. If the difference between the out-of-bag error of the baseline model and the full model is over the threshold, the performance of the full model was superior after including the past information on SM. In this case, we reject the null hypothesis that SM has no significant influence on future rainfall. The pixels in which the null hypothesis was rejected were used for further analysis, and all the other pixels were omitted.

In our study, we did not apply the parametric or nonparametric methods because they are unsuited for the statistical test for random forest. In our study, we only focused on the sign and the hot spot of the feedback rather than a qualitative analysis.

3) QUANTIFICATION OF SM–POCC FEEDBACK IMPACT

The SM–POCC feedback was estimated for each pixel and quantified by examining the ratio of the POCC values predicted by the full model and the baseline model for each day in the 9-yr period (step 3 in [Fig. 2](#)). A seasonal anomaly of SM was used here, and we divided the seasonal anomaly into dry and wet days, which were determined by comparison with the seasonal average value of SM.

3. Results and discussion

a. Comparison between linear and nonlinear SM– P feedback

We first compared the linear and nonlinear regression to identify the POCC estimated from periodic terms (i.e., interannual and seasonal period) and persistence terms (i.e., lagged precipitation and pressure) (see [section 2c](#)) of atmospheric terms, that is, impact derived from interannual and seasonal variations, precipitation and atmospheric persistence. Both the frameworks were analyzed using the same datasets, that is, NCA-LDAS. For the linear framework, we employed a GLM, which was consistent with the approach used by [Tuttle and Salvucci \(2016\)](#), while for the nonlinear framework, the RF regression method was utilized.

Five-fold cross-validation was coupled with a linear and nonlinear regression framework for each pixel. The cross-validation can also be employed to avoid the overfitting issue of the regression. The corresponding spatial distribution of the root-mean-square error (RMSE) values of the validation dataset is shown in [Fig. 3](#). The average RMSE value of the RF model is 0.3919, while the value was 0.5346 for the GLM model, suggesting that RF has more predictive power in estimating the highly nonlinear atmospheric response (but not SM) in the formation of P . Additionally, both frameworks showed that the eastern part of the United States had larger RMSE values than did the western part of the United States.

Five cases were designed to investigate the impact of the predictive methods, spatial variables, and hybrid feature choosing method based on the previously mentioned 25 984 grids. The median values of R^2 of the five cases were 0.103, 0.646, 0.635, 0.696, and 0.689, respectively ([Table 1](#)). We found that the explanatory power was significantly improved when the nonlinear atmosphere response was considered (compared with the linear response), and the median R^2 value increased from 0.103 to 0.646 from the GLM to the RF models, suggesting the importance of including the nonlinear response of the atmospheric terms in the process of forming P . The predictive power was also improved by introducing the spatial impact of regional climatic factors, with the predictive power increasing by approximately 5% (cases 3 and 5). This indicated that the spatial effect of precipitation and atmospheric pressure (i.e., rapid changed variables) on the SM– P feedback is indeed too little. In addition, by using the hybrid method, R^2 was similar to the model that used all explanatory variables (cases 2 and 3; cases 4 and 5), suggesting that the best feature selected by our hybrid method was enough to represent the information of the

FIG. 3. The RMSE of testing set using (a) linear (GLM) and (b) nonlinear (RF) framework. The target variable is precipitation occurrence, and the explaining variables are periodic terms and persistence terms.

nonlinear atmospheric response process and would save significant computational time.

b. The spatial distribution of SM–P feedback

Figure 2 shows the nonlinear causal inference model we applied to the NCA-LDAS datasets to analyze the short-term SM–POCC feedback of the contiguous United States, and the results are separated into dry and wet days. The results of dominating positive feedback (i.e., impact value higher than 1 in wet SM conditions and lower than 1 in dry SM conditions, see Fig. 4) are consistent with the results of several previous studies that used different methods, such as the scaling analysis (Jones et al. 2009), the lagged covariance ratios study (Zhang et al. 2009), and the convergent cross-mapping study (Wang et al. 2018).

However, the hot spot regions of the SM–POCC feedback across the United States are still heavily debated, especially over the Great Plains region. Koster et al. (2004) showed that the Great Plains is a strong feedback region based on the Global Land Atmosphere Coupling Experiments (GLACE), which use highly controlled numerical experiments and the same model settings. However, individual models performed significantly differently in GLACE. For example, the Geophysical Fluid Dynamics Laboratory (GFDL) model found strong

feedback over the Great Plains, while the Bureau of Meteorology Research Center (BMRC) model indicated very weak feedback, suggesting the dependence on the parameterization schemes of models. However, there were contradictory conclusions drawn from the observations found by Taylor et al. (2012), who suggested that there was no clear feedback of SM–POCC in the Great Plains. Similar conclusions were reached by Alfieri et al. (2008), who found that mostly insignificant feedbacks of SM–P occurred in the upper region of the Great Plains by utilizing the atmosphere profiles.

Figure 4 also shows that the hot spot regions over dry and wet days are significantly different. For example, the strength of SM–P feedback on dry days is much stronger than that on wet days over the entire contiguous United States (deeper and wider cold colors in the dry days than warmer colors in the wet days). This pattern may result from the much larger variation in the sensitivity of ET to SM on dry days than the variation in wet days (Wei and Dirmeyer 2012). Additionally, the southeastern United States is an interesting region to investigate, as it was found that the SM–P feedback was strong on dry days but had less impact on wet days. It is known that P on wet days in the southeastern United States is generally caused by cyclones, mesoscale convective systems, and

TABLE 1. Regression, spatial variable, and selection method used in five cases, and corresponding median values of R^2 . GLM and RF denote generalized linear model and random forest, respectively.

Case No.	Regression methods		Spatial variable	Selection method			Median of R^2
	GLM	RF		None	Hybrid method	All possible regression	
1	✓					✓	0.103
2		✓		✓			0.646
3		✓			✓		0.635
4		✓	✓	✓			0.696
5		✓	✓		✓		0.689

FIG. 4. The impact of soil moisture on the next-day precipitation probability (SM-POCC) over the United States for (a) dry and (b) wet days suggesting drier or wetter than seasonal median conditions. Cold colors indicate the inclusion of soil moisture in the model reduced the predicted precipitation probability, while warm colors indicate the opposite. A white color indicates insignificant feedback regions and areas outside of the study area.

even large-scale weather systems (Henderson and Robinson 1994; Silvestri and Vera 2003; Bombardi et al. 2014), rather than by local land-atmosphere interactions. However, P on dry days is probably strongly controlled by local ET, leading to much stronger SM- P feedback on dry days (Wei and Dirmeyer 2012). Positive SM- P feedback in dry days means less SM will decrease P probability, and further bring less SM to the ground, which may enhance the drought over the strong feedback region. Therefore, we suspect that the strong and positive feedback on dry days may gradually amplify the drought over some critical regions, such as the southwestern United States and the Great Plains.

The hot spots found in this study were mainly located around the southwestern coastal area and the Great Plains region. Sanford and Selnick (2013) estimated the fraction of P lost to ET in these regions and indicated that nearly 90% of the P was transferred by ET, suggesting a typical regional control pattern in relation to ET. Moreover, the hot spot in the southwestern region was located at the leeward slope of the mountains (i.e., Rocky Mountains), where the water vapor was blocked, and therefore, the ET was controlled mainly by SM rather than by the horizontal transport of water vapor (Lam et al. 2007). Regarding the other regions that were not identified as hot spot regions, such as the East and West Coasts of the United States, ET was limited by atmospheric factors and not by the availability of SM, suggesting the atmosphere-control pattern identified by Koster et al. (2003).

To investigate the SM- P impact across different climatic regions, the definition of the arid index derived from the Consultative Group for International Agriculture Research (CGIAR) (Trabucco and Zomer 2018) was used to divide the contiguous United States into four climatic regions (see Fig. 5a). Figure 5b shows a summary

of the probability density curves (PDCs) of the SM-POCC impact on dry, semidry, semiwet, and wet regions in wet days, as identified by the nonlinear GC model.

It is clear that the regions with strong SM- P feedback (i.e., high SM-POCC impact values in Fig. 5b) are located over dry and transition zones (semidry and semiwet regions), which is consistent with the spatial distribution of the SM-POCC impact identified in Fig. 4. In regions with the strongest SM- P feedback (i.e., impact over 1.05), the SM-POCC impact decreased as the climate of the region changed from dry to wet. Two peaks were found in the curves for the semidry and semiwet regions (red and orange curves), suggesting that the dynamics of the land-atmosphere interaction over transition zones are somewhat different and complex. The northwestern area and the Great Plains regions are both located in transition zones but show opposite SM- P feedback patterns. The possible reason is that these regions are both influenced by the eastward moist air transported by the westerlies, but the mountains (e.g., Rocky Mountains) block the moist air, keeping it to the west of the Rocky Mountains (i.e., the northwestern area). Therefore, moisture in the northwestern area is controlled by the atmosphere rather than by SM. Furthermore, the Rocky Mountains decreased the impact of large-scale circulation over the Great Plains and, therefore, ET controlled the P in these regions. Furthermore, the local SM-POCC feedback in these areas may be stronger than that in the northwestern area. In addition, the Rocky Mountains may not completely block the transport of water vapor into the northern Great Plains; thus, the feedback in these areas is not as strong as that in the southern Great Plains.

In general, we identified that the hot spot regions were mostly located over the basin and the plains around the mountains, where the transport of moisture is blocked.

FIG. 5. (a) Four climate regions over the United States. The red color is dry regions, the orange color is semidry regions, the yellow color is semiwet regions, and the green color is the wet regions. (b) The probability density curves of SM-POCC impact in wet days over different climate regions.

The strong feedback over the southwestern area and the Great Plains both highlight the impact of topographic factors rather than only the sensitivity of ET to SM.

c. The impact of different datasets

To demonstrate the reliability of our findings, we used different datasets to compare the results based on the NCA-LDAS dataset. We added two additional experiments based on different datasets. In experiment I, the P and surface pressure from NLDAS and the surface SM from GLDAS were analyzed (Figs. 6a,b). In experiment II, we adopted P , SM, and surface pressure from ERA5 (Figs. 6c,d). The results showed that the overall pattern of the SM-POCC feedback over the contiguous United States was in good agreement among the three experiments (Figs. 4 and 6). For example, the hot spot regions were mainly located on the southwestern coast and the

Great Plains, and the signs of SM-POCC feedback were all positive. However, the variability among different experiments was modest, with significant differences in the strength of the feedback. We attributed this disagreement to the uncertainty and variability of the input dataset. Kumar et al. (2019) suggested that global datasets (e.g., GLDAS and ERA5) are less accurate than the NCA-LDAS dataset, and we therefore highlight the importance of the quality of the input datasets in identifying the SM-POCC feedback.

The dataset used by Tuttle and Salvucci (2016) (i.e., P of NLDAS and SM of AMSR-E) was also analyzed by applying the proposed nonlinear framework, but we did not find consistent SM-POCC feedback compared with the NCA-LDAS dataset over the United States (Fig. 7). This result may be caused by the uncertainty of the remote sensing SM data. Previous studies have shown that satellite datasets may have significant uncertainty over regions with extensively distributed vegetation (Wei et al. 2012). Tuttle and Salvucci (2016) further indicated that the AMSR-E SM may be inaccurate due to the dense vegetation in the eastern United States. Therefore, the leaf water content and dew formation, which are anticorrelated to SM, may lead to the negative SM- P feedback in the eastern United States, contradicting the preceding results.

Although the signs of the SM-POCC feedback were different using the different input datasets, the agreement of the hot spot locations in these three different input datasets (i.e., NCA-LDAS, GLDAS, and ERA5) indicates the results are robust; additionally, this information provides an opportunity to benchmark the ESMs.

d. Benchmark for CMIP6

To investigate the ability of different ESMs to characterize the hot spot regions of the SM- P feedback, we adopted five coupled model outputs from World Climate Research Programme's Coupled Model Intercomparison Project phase 6 (CMIP6) and applied the proposed nonlinear GC causality framework into the output. Table 2 provides a summary of the CMIP6 output data used in this study. Figure 8 shows the impact of the SM-POCC feedback over the United States using different datasets from CMIP6, with the left and right columns indicating the results of the dry days and wet days, respectively. In the dry days, all five ESMs showed a strong feedback signal over the southern United States, which was consistent with our preceding results. However, the SM- P signal was significantly different over the northeastern United States. SAM0-UNICON (Figs. 8c,d) and EC-Earth3-Veg (Figs. 8e,f) present strong SM- P feedback over the northeastern United States, which was contradictory to the SM- P signal compared with the

FIG. 6. The impact of SM-POCC feedback over the United States of different datasets. (a),(b) The result of dry and wet days using ERA5 precipitation and surface soil moisture datasets. (c),(d) The result of dry and wet days using GLDAS precipitation and NLDAS surface soil moisture datasets.

other three ESMs and our preceding results (Fig. 4). This signal may be artificial because this region is a typical atmosphere-controlled region rather than an ET-controlled region (Koster et al. 2003). Thus, in these two models, the importance of SM on forming P may be overestimated in the northeastern United States on dry days. On wet days, MRI-ESM2.0 (Fig. 8b), CESM2

(Fig. 8h, and GFDL CM4 (Fig. 8j) showed nearly identical hot spot regions relative to the results shown in Fig. 4, where the hot spot regions were mostly located over the southern Great Plains and the southwestern United States. Furthermore, GFDL CM4 (Fig. 8j) presented a slightly weaker SM- P impact over the northern Great Plains. The strong feedback over the southwestern

FIG. 7. The impact of SM-POCC feedback over the United States using NLDAS precipitation and AMSR-E UMT soil moisture dataset. (a),(b) The result of dry and wet days using GLM as regression model. (c),(d) The result of dry and wet days result using RF as regression model.

TABLE 2. CMIP6 output used in this study, with time ranging from 19 Jun 2002 to 19 Jun 2011, corresponding to the time range of NCA-LDAS dataset. Precipitation and surface soil moisture were involved in this study. The model outputs are all from historical model run only, and the variant label is r1i1p1f1.

Center	Model version	Resolution
Meteorological Research Institute (MRI), Japan Meteorological Agency	MRI-ESM2.0	320 × 160
National Oceanic and Atmospheric Administration (NOAA)/Geophysical Fluid Dynamics Laboratory (GFDL)	GFDL CM4	288 × 180
EC-Earth Consortium	EC-Earth3-Veg	512 × 256
Seoul National University	SAM0-UNICON	288 × 192
National Center for Atmospheric Research (NCAR), Climate and Global Dynamics Laboratory	CESM2	288 × 192

coast can be approximately described by the SAM0-UNICON (Fig. 8d) and EC-Earth3-Veg (Fig. 8f) models, but strong feedback over the northeastern United States was also presented by these two models. The feedback presented by the SAM0-UNICON (Fig. 8c) and EC-Earth3-Veg (Fig. 8e) models was quite consistent on dry days, which further emphasized that the impact of SM in these areas may be overestimated. Thus, the parameterization schemes in these two models may need to be improved, especially over the northeastern United States.

In conclusion, all five ESMs indicate there is consistently strong SM–P feedback in the southwestern United States; however, the models show significant differences over the Great Plains and in the northeastern United States. These two regions should be focused on in further research on the land–atmosphere feedback using ESMs. Our model provides a land–atmosphere strength metric that can be used to measure and compare the performance skills among ESMs and provides a new perspective to improve the parameterization of ESMs.

e. Reasons for using our method in detecting SM–P feedback

Regression-based and information theory based methods could both be used for detecting nonlinear Granger causality in complex system. Both of them have their pros and cons. We provided several reasons to use our method to identify the SM–P feedbacks as follows. First, machine learning provides a good method for nonlinear regression, and can well explore the relationship between two time series. RF is a tree-based model, and it also provides a good performance of nonlinear fitting (Fig. 1), and the calculation cost is low. Thus, RF is adopted as the regression model in our causal inference model. Second, SM–P feedback is mostly local interaction (Santanello et al. 2018). Therefore, the aim of our study is to construct a grid-based causal inference model, and give the sign and pattern of SM–P feedback rather than explore the physical process of land–atmosphere feedback. So, we only used two variables, that is, SM and P , to isolate a

local SM–P feedback relationship by removing their periodicity, persistence effect. Third, our method provided a metric for identifying the SM–P feedback in dry and wet conditions. This feedback is sensitive to the dry and wet conditions of soil. For example, SM could not provide enough vapor to trigger precipitation in dry conditions (Tuttle and Salvucci 2016). Therefore, we divided the SM time series into dry days and wet days, and provide the feedback strength separately to identify the sign (i.e., positive and negative) of SM–P feedback. If we apply causal network learning, we can only get a relative intensity value in average state rather than the sign and intensity in both dry and wet soil situations.

4. Conclusions and future perspectives

SM–P feedback is a rapidly advancing field that is helpful for predicting climate change and extreme events and may improve parameterization schemes of ESMs. SM–P feedback is highly nonlinear, and it is difficult to identify the causal relationship. We combined and improved some existing useful causal inference methods, and got a more complete inference framework to identify the sign and pattern of SM–P feedback. We obtained the following conclusions and future perspectives.

- 1) We used a comprehensive causal inference framework to analyze the SM–P feedback, which includes predictive modeling by RF, time series decomposition techniques, the ADF test, lag window selection, a hybrid feature selection method, and the nonlinear GC analysis. Our framework significantly improved the linear framework of Tuttle and Salvucci (2016), and we highlighted the importance of considering the nonlinear atmospheric responses in the process of P formation and emphasized the nonlinear relationship in the complex SM–P feedback.
- 2) Our finding supports the hypothesis that strong feedback regions are mostly located in the humid–arid transition zone regions (i.e., the southern Great Plains and the southwestern coast of the United States)

FIG. 8. The impact of SM-POCC feedback over the United States using five different model output datasets from CMIP6. (left) Dry days and (right) wet days.

identified in previous studies. Our results may reveal that the local feedback may be not only caused by the sensitivity of ET to SM. Large-scale topographic features may play a role in blocking the transport of moisture, resulting in an ET-controlled P pattern over these regions.

- 3) The SM- P index from the proposed framework can be used as a metric to benchmark the ESMs, which provides a perspective to improve the parameterization schemes in the ESMs. We found that MRI-ESM2.0, CESM2, and GFDL CM4 performed well in

representing the short-term SM- P feedback. SAM0-UNICON and EC-Earth3-Veg could describe the strong feedback over the southwestern coast of the United States but could not describe the strong feedback over the Great Plains and might provide an unreliable strong signal over the northeastern United States. These two regions should be emphasized in future research on land-atmosphere feedbacks.

- 4) The proposed framework is expected to provide a new perspective to identify some other feedback loops in the Earth system, such as exploring the

dynamics of those key feedbacks (e.g., vegetation–atmosphere feedback). Further work should include the following three topics. First, the adaptability of this model at different temporal scales rather than only the synoptic scale will be investigated. For example, at the subseasonal scale, the impact of longer soil memory and P persistence must be considered. Second, although we introduced spatial effects of P and pressure to identified SM–POCC feedback, we considered only the spatial impacts for faster changed variables. Therefore, the spatial impact over the larger scale and slowly changed variables will be further investigated. Third, the adaptability of this model will be investigated for different variables participating in the SM– P feedback process, such as the ET and lift condensation level.

Acknowledgments. Lu Li thanks Samuel Tuttle for his detailed explanation of their method. This work was supported by the Natural Science Foundation of China under Grant U1811464, 41975122, 41575072, and 41807181, the National Key R&D Program of China under Grant 2017YFA0604303, and the Fundamental Research Funds for the Central Universities. Y. Deng is supported by the National Science Foundation Climate and Large-Scale Dynamics (CLD) program through Grant AGS-1445956. J. Mao is supported by the Reducing Uncertainty in Biogeochemical Interactions through Synthesis and Computation (BUBISCO) Scientific Focus Area (SFA) project funded through the Regional and Global Climate Modeling Program in Climate and Environmental Sciences Division (CESD) of the Biological and Environmental Research (BER) Program in the U.S. Department of Energy Office of Science. Oak Ridge National Laboratory is supported by the Office of Science of the U.S. Department of Energy under Contract DE-AC05-00OR22725.

REFERENCES

- Alfieri, L., P. Claps, P. D'Odorico, F. Laio, and T. M. Over, 2008: An analysis of the soil moisture feedback on convective and stratiform precipitation. *J. Hydrometeorol.*, **9**, 280–291, <https://doi.org/10.1175/2007JHM863.1>.
- Bombardi, R. J., L. M. Carvalho, C. Jones, and M. S. Reboita, 2014: Precipitation over eastern South America and the South Atlantic Sea surface temperature during neutral ENSO periods. *Climate Dyn.*, **42**, 1553–1568, <https://doi.org/10.1007/s00382-013-1832-7>.
- Breiman, L., 2001: Random forests. *Mach. Learn.*, **45**, 5–32, <https://doi.org/10.1023/A:1010933404324>.
- Brovelli, A., M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler, 2004: Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by Granger causality. *Proc. Natl. Acad. Sci. USA*, **101**, 9849–9854, <https://doi.org/10.1073/pnas.0308538101>.
- Cleveland, R. B., W. S. Cleveland, J. E. McRae, and I. Terpenning, 1990: STL: A seasonal-trend decomposition procedure based on loess. *J. Off. Stat.*, **6**, 3–33.
- Cook, B. I., G. B. Bonan, and S. Levis, 2006: Soil moisture feedbacks to precipitation in southern Africa. *J. Climate*, **19**, 4198–4206, <https://doi.org/10.1175/JCLI3856.1>.
- Davidson, E. A., E. Belk, and R. D. Boone, 1998: Soil water content and temperature as independent or confounded factors controlling soil respiration in a temperate mixed hardwood forest. *Global Change Biol.*, **4**, 217–227, <https://doi.org/10.1046/j.1365-2486.1998.00128.x>.
- , L. V. Verchot, J. H. Cattaneo, I. L. Ackerman, and J. E. M. Carvalho, 2000: Effects of soil water content on soil respiration in forests and cattle pastures of eastern Amazonia. *Biogeochemistry*, **48**, 53–69, <https://doi.org/10.1023/A:1006204113917>.
- Dirmeyer, P. A., C. A. Schlosser, and K. L. Brubaker, 2009: Precipitation, recycling, and land memory: An integrated analysis. *J. Hydrometeorol.*, **10**, 278–288, <https://doi.org/10.1175/2008JHM1016.1>.
- Douville, H., 2002: Influence of soil moisture on the Asian and African monsoons. Part II: Interannual variability. *J. Climate*, **15**, 701–720, [https://doi.org/10.1175/1520-0442\(2002\)015<0701:IOSMOT>2.0.CO;2](https://doi.org/10.1175/1520-0442(2002)015<0701:IOSMOT>2.0.CO;2).
- Findell, K. L., and E. A. Eltahir, 1997: An analysis of the soil moisture-rainfall feedback, based on direct observations from Illinois. *Water Resour. Res.*, **33**, 725–735, <https://doi.org/10.1029/96WR03756>.
- Fischer, E. M., S. I. Seneviratne, D. Lüthi, and C. Schär, 2007: Contribution of land-atmosphere coupling to recent European summer heat waves. *Geophys. Res. Lett.*, **34**, L06707, <https://doi.org/10.1029/2006GL029068>.
- Ghosal, I., and G. Hooker, 2018: Boosting random forests to reduce bias; One-step boosted forest and its variance estimate. arXiv, 39 pp., <https://arxiv.org/abs/1803.08000>.
- Granger, C. W., 1969: Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, **37**, 424–438, <https://doi.org/10.2307/1912791>.
- Grieser, J., S. Trömel, and C. D. Schönwiese, 2002: Statistical time series decomposition into significant components and application to European temperature. *Theor. Appl. Climatol.*, **71**, 171–183, <https://doi.org/10.1007/s007040200003>.
- Guilod, B. P., and Coauthors, 2014: Land-surface controls on afternoon precipitation diagnosed from observational data: Uncertainties and confounding factors. *Atmos. Chem. Phys.*, **14**, 8343–8367, <https://doi.org/10.5194/acp-14-8343-2014>.
- Guo, Z., and Coauthors, 2006: GLACE: The Global Land–Atmosphere Coupling Experiment. Part II: Analysis. *J. Hydrometeorol.*, **7**, 611–625, <https://doi.org/10.1175/JHM511.1>.
- Henderson, K. G., and P. J. Robinson, 1994: Relationships between the Pacific/North American teleconnection patterns and precipitation events in the south-eastern USA. *Int. J. Climatol.*, **14**, 307–323, <https://doi.org/10.1002/joc.3370140305>.
- Hirschi, M., and Coauthors, 2011: Observational evidence for soil-moisture impact on hot extremes in southeastern Europe. *Nat. Geosci.*, **4**, 17–21, <https://doi.org/10.1038/ngeo1032>.
- Hohenegger, C., P. Brockhaus, C. S. Bretherton, and C. Schär, 2009: The soil moisture–precipitation feedback in simulations with explicit and parameterized convection. *J. Climate*, **22**, 5003–5020, <https://doi.org/10.1175/2009JCLI2604.1>.

- Jones, L. A., J. S. Kimball, E. Podest, K. C. McDonald, S. K. Chan, and E. G. Njoku, 2009: A method for deriving land surface moisture, vegetation optical depth, and open water fraction from AMSR-E. *IEEE Int. Conf. on Geoscience and Remote Sensing Symp.*, Cape Town, South Africa, Institute of Electrical and Electronics Engineers, III-916–III-919, <https://doi.org/10.1109/IGARSS.2009.5417921>.
- Koster, R. D., M. J. Suarez, R. W. Higgins, and H. M. Van den Dool, 2003: Observational evidence that soil moisture variations affect precipitation. *Geophys. Res. Lett.*, **30**, 1241, <https://doi.org/10.1029/2002GL016571>.
- , and Coauthors, 2004: Regions of strong coupling between soil moisture and precipitation. *Science*, **305**, 1138–1140, <https://doi.org/10.1126/science.1100217>.
- , and Coauthors, 2006: GLACE: The global land–atmosphere coupling experiment. Part I: Overview. *J. Hydrometeorol.*, **7**, 590–610, <https://doi.org/10.1175/JHM510.1>.
- Krakauer, N. Y., B. I. Cook, and M. J. Puma, 2010: Contribution of soil moisture feedback to hydroclimatic variability. *Hydrol. Earth Syst. Sci.*, **14**, 505–520, <https://doi.org/10.5194/hess-14-505-2010>.
- Kumar, S. V., M. Jasinski, D. M. Mocko, M. Rodell, J. Borak, B. Li, H. K. Beaudoin, and C. D. Peters-Lidard, 2019: NCA-LDAS land analysis: Development and performance of a multisensor, multivariate land data assimilation system for the National Climate Assessment. *J. Hydrometeorol.*, **20**, 1571–1593, <https://doi.org/10.1175/JHM-D-17-0125.1>.
- Lam, A., M. F. Bierkens, and B. J. van den Hurk, 2007: Global patterns of relations between soil moisture and rainfall occurrence in ERA-40. *J. Geophys. Res.*, **112**, D17116, <https://doi.org/10.1029/2006JD008222>.
- Lawrence, D. M., and J. M. Slingo, 2005: Weak land–atmosphere coupling strength in HadAM3: The role of soil moisture variability. *J. Hydrometeorol.*, **6**, 670–680, <https://doi.org/10.1175/JHM445.1>.
- Pal, J. S., E. E. Small, and E. A. Eltahir, 2000: Simulation of regional-scale water and energy budgets: Representation of subgrid cloud and precipitation processes within RegCM. *J. Geophys. Res.*, **105**, 29 579–29 594, <https://doi.org/10.1029/2000JD900415>.
- Papagiannopoulou, C., D. Gonzalez Miralles, S. Decubber, M. Demuzere, N. Verhoest, W. A. Dorigo, and W. Waegeman, 2017: A non-linear Granger-causality framework to investigate climate-vegetation dynamics. *Geosci. Model Dev.*, **10**, 1945–1960, <https://doi.org/10.5194/gmd-10-1945-2017>.
- Roebroeck, A., E. Formisano, and R. Goebel, 2005: Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage*, **25**, 230–242, <https://doi.org/10.1016/j.neuroimage.2004.11.017>.
- Rowell, D. P., and C. Blondin, 1990: The influence of soil wetness distribution on short-range rainfall forecasting in the West African Sahel. *Quart. J. Roy. Meteor. Soc.*, **116**, 1471–1485, <https://doi.org/10.1002/qj.49711649611>.
- Salvucci, G. D., J. A. Saleem, and R. Kaufmann, 2002: Investigating soil moisture feedbacks on precipitation with tests of Granger causality. *Adv. Water Resour.*, **25**, 1305–1312, [https://doi.org/10.1016/S0309-1708\(02\)00057-X](https://doi.org/10.1016/S0309-1708(02)00057-X).
- Sanford, W. E., and D. L. Selnick, 2013: Estimation of evapotranspiration across the conterminous United States using a regression with climate and land-cover data. *J. Amer. Water Resour. Assoc.*, **49**, 217–230, <https://doi.org/10.1111/jawr.12010>.
- Santanello, J. A., Jr., and Coauthors, 2018: Land–atmosphere interactions: The LoCo perspective. *Bull. Amer. Meteor. Soc.*, **99**, 1253–1272, <https://doi.org/10.1175/BAMS-D-17-0001.1>.
- 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>.
- Shukla, J., and Y. Mintz, 1982: Influence of land-surface evapotranspiration on the earth's climate. *Science*, **215**, 1498–1501, <https://doi.org/10.1126/science.215.4539.1498>.
- Silvestri, G. E., and C. S. Vera, 2003: Antarctic Oscillation signal on precipitation anomalies over southeastern South America. *Geophys. Res. Lett.*, **30**, 2115, <https://doi.org/10.1029/2003GL018277>.
- Song, H. J., C. R. Ferguson, and J. K. Roundy, 2016: Land–atmosphere coupling at the Southern Great Plains Atmospheric Radiation Measurement (ARM) field site and its role in anomalous afternoon peak precipitation. *J. Hydrometeorol.*, **17**, 541–556, <https://doi.org/10.1175/JHM-D-15-0045.1>.
- Stuecker, M. F., A. Timmermann, F. F. Jin, S. McGregor, and H. L. Ren, 2013: A combination mode of the annual cycle and the El Niño/Southern Oscillation. *Nat. Geosci.*, **6**, 540–544, <https://doi.org/10.1038/ngeo1826>.
- Taylor, C. M., A. Gounou, F. Guichard, P. P. Harris, R. J. Ellis, F. Couvreur, and M. De Kauwe, 2011: Frequency of Sahelian storm initiation enhanced over mesoscale soil-moisture patterns. *Nat. Geosci.*, **4**, 430–433, <https://doi.org/10.1038/ngeo1173>.
- , R. A. de Jeu, F. Guichard, P. P. Harris, and W. A. Dorigo, 2012: Afternoon rain more likely over drier soils. *Nature*, **489**, 423–426, <https://doi.org/10.1038/nature11377>.
- Trabucco, A., and R. J. Zomer, 2018: Global aridity index and potential Evapo-Transpiration (ET0) climate database v2. CGIAR Consortium for Spatial Information (CGIAR-CSI), accessed 24 January 2019, <https://cgiarcsi.community>.
- Trenberth, K. E., and C. J. Guillemot, 1996: Physical processes involved in the 1988 drought and 1993 floods in North America. *J. Climate*, **9**, 1288–1298, [https://doi.org/10.1175/1520-0442\(1996\)009<1288:PPIITD>2.0.CO;2](https://doi.org/10.1175/1520-0442(1996)009<1288:PPIITD>2.0.CO;2).
- Tuttle, S., and G. Salvucci, 2016: Empirical evidence of contrasting soil moisture–precipitation feedbacks across the United States. *Science*, **352**, 825–828, <https://doi.org/10.1126/science.aaa7185>.
- Van Looy, K., and Coauthors, 2017: Pedotransfer functions in Earth system science: Challenges and perspectives. *Rev. Geophys.*, **55**, 1199–1256, <https://doi.org/10.1002/2017RG000581>.
- Wang, Y., J. Yang, Y. Chen, P. De Maeyer, Z. Li, and W. Duan, 2018: Detecting the causal effect of soil moisture on precipitation using convergent cross mapping. *Sci. Rep.*, **8**, 12171, <https://doi.org/10.1038/s41598-018-30669-2>.
- Wei, J., and P. A. Dirmeyer, 2012: Dissecting soil moisture–precipitation coupling. *Geophys. Res. Lett.*, **39**, L19711, <https://doi.org/10.1029/2012GL053038>.
- , R. E. Dickinson, and H. Chen, 2008: A negative soil moisture–precipitation relationship and its causes. *J. Hydrometeorol.*, **9**, 1364–1376, <https://doi.org/10.1175/2008JHM955.1>.
- Zeng, D., and X. Yuan, 2018: Multiscale land–atmosphere coupling and its application in assessing subseasonal forecasts over East Asia. *J. Hydrometeorol.*, **19**, 745–760, <https://doi.org/10.1175/JHM-D-17-0215.1>.
- Zhang, J., W. C. Wang, and L. Wu, 2009: Land–atmosphere coupling and diurnal temperature range over the contiguous United States. *Geophys. Res. Lett.*, **36**, L06706, <https://doi.org/10.1029/2009GL037505>.