Water Resources Research

RESEARCH ARTICLE
10.1029/2024WR039470

Special Collection:
Advancing Interpretable AI/ML Methods for Deeper Insights and Mechanistic Understanding in Earth Sciences: Beyond Predictive Capabilities

Incorporating Causality Into Deep Learning Architectures to Improve Flash Drought Forecasts

Sijie Tang[1,2], **Shuo Wang**[1,3], **Jiping Jiang**[2], **and Yi Zheng**[2]

[1]State Key Laboratory of Climate Resilience for Coastal Cities, Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong, China, [2]School of Environmental Science and Engineering, Southern University of Science and Technology, Shenzhen, China, [3]Research Institute for Land and Space, The Hong Kong Polytechnic University, Hong Kong, China

Abstract Soil moisture flash droughts present challenges to agriculture and ecosystems, leading to widespread socioeconomic impacts. Predicting and providing early warnings for these events remains difficult. We propose a novel deep learning framework, the ResAttCauRec model, which integrates an attention mechanism and additional causal information into a CNN-LSTM (convolutional neural network with long short-term memory) backbone to capture the dependence of soil moisture on spatial-temporal meteorological variables. Our results demonstrate that the causality module acts as a regularization technique, enhancing model generalization and performance. This enables effective forecasts of flash droughts, achieving an F1 score of 0.41 compared to 0.06 for the baseline model. Model interpretation analysis reveals that the causality degree significantly improves predictive performance for key drivers including daily maximum temperature, evaporation, and surface pressure, alongside soil temperature and moisture. While normal droughts are influenced by long-term temperature trends, flash droughts are more sensitive to rapid atmospheric changes. Our analysis also highlights a concerning trend of increasing drought complexity and intensification, complicating reliable predictions. This study offers valuable insights into flash drought onset mechanisms and advocates for enhanced predictive models that better support agricultural and ecological practices. Additionally, we introduce an effective approach to enhance data-driven models by incorporating additional causal information, which not only facilitates forecast and interpretation of flash droughts but may also be extended to broader extreme weather events.

Plain Language Summary Flash droughts are sudden and intense periods of soil drying that can disrupt agriculture, harm ecosystems, and cause economic losses. Predicting these events early is crucial but remains a big challenge. To address this, we developed a new deep learning model called ResAttCauRec, which incorporates additional causal information into advanced machine learning techniques to better understand how weather patterns over time and space affect soil moisture. Our tests showed that the new model significantly outperforms basic approaches, with a notable increase in prediction accuracy. By analyzing how the model works, we found that flash droughts are often triggered by rapid changes in weather conditions, such as daily maximum temperature, evaporation, and shifts in air pressure, while more gradual droughts are influenced by longer-term trends. We also observed that flash droughts are becoming more complex and harder to predict, raising concerns for the future. This research provides new insights into the triggers of flash droughts and offers improved tools for forecasting them. Beyond droughts, our approach of incorporating causal information into data-driven models could enhance predictions for other extreme weather events, offering better support for climate adaptation and mitigation efforts.

Key Points:

- A novel model integrates attention and causality into a CNN-LSTM backbone to capture the spatial-temporal dependence of soil moisture
- Causal information enhances model performance and generalization, facilitating effective forecasts of flash droughts
- The onset of flash drought is driven by distinct factors and is becoming increasingly complex, posing challenges for future predictions

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

S. Wang,
shuo.s.wang@polyu.edu.hk

Citation:

Tang, S., Wang, S., Jiang, J., & Zheng, Y. (2025). Incorporating causality into deep learning architectures to improve flash drought forecasts. *Water Resources Research*, *61*, e2024WR039470. https://doi.org/10.1029/2024WR039470

Received 18 NOV 2024
Accepted 24 SEP 2025

Author Contributions:

Conceptualization: Shuo Wang
Data curation: Sijie Tang
Formal analysis: Sijie Tang
Funding acquisition: Shuo Wang
Investigation: Shuo Wang
Methodology: Sijie Tang, Shuo Wang
Project administration: Shuo Wang
Resources: Shuo Wang
Software: Sijie Tang

1. Introduction

Drought is traditionally viewed as a gradual climatic phenomenon that may last for months or even years (Allen et al., 2010; A. K. Mishra & Singh, 2010). However, recent research indicates that under certain extreme atmospheric conditions—such as elevated temperatures, strong winds, or low humidity—drought can develop more intensely and rapidly due to insufficient precipitation or increased evaporative demand, disrupting both ecosystems and agricultural systems (Crausbay et al., 2017; Jing et al., 2025; Otkin et al., 2015; Qing & Wang, 2025). This form, often termed flash drought, is marked by its sudden onset and severe intensity (Senay et al., 2008). Unlike conventional droughts, flash droughts can have particularly acute impacts on agriculture, natural ecosystems, and society, owing to their abrupt nature (Otkin et al., 2018; Svoboda et al., 2002). This rapid

Supervision: Shuo Wang, Jiping Jiang, Yi Zheng
Validation: Sijie Tang
Visualization: Sijie Tang
Writing – original draft: Sijie Tang
Writing – review & editing: Shuo Wang, Jiping Jiang, Yi Zheng

development complicates timely warnings and effective response measures for resource managers and stakeholders (Anderson et al., 2013). The summer drought experienced in the American Midwest in 2012 serves as a notable instance of a historical flash drought event, resulting in serious harm to local crops and a staggering $35.7 billion economic loss (Smith, 2018). Consequently, there is an imperative to heighten awareness of flash droughts, foster adaptive strategies, deliver early warning, and provide emergency response (Otkin et al., 2022).

Despite its importance, relatively few studies have focused specifically on flash drought prediction. Traditional methods for predicting flash drought often rely on models based on drought indices, physical processes, or coupled systems, utilizing meteorological and hydrological parameters (Tyagi et al., 2022). Most index-based studies analyze interactions between flash drought indicators, such as soil moisture (SM) and related climatic drivers (L. G. Chen et al., 2019; Liang & Yuan, 2021). Additionally, indices like the Standardized Precipitation Index (SPI) and Standardized Precipitation Evapotranspiration Index (SPEI) are often applied within probabilistic frameworks to forecast drought propagation and identify flash drought events (AghaKouchak, 2014; Ho et al., 2021; Yevjevich, 1967). Although these methods are widely used, they may struggle to capture the nonlinear and dynamic relationships among different indicators and drivers due to uncertainties introduced by process simplifications and limitations in input data (Prodhan et al., 2021; Q. Yuan et al., 2020). Alternatively, physical models, including coupled models like Global Climate Models (GCMs) with the Variable Infiltration Capacity (VIC) model, have been used to simulate SM under future climate conditions (V. Mishra et al., 2021). However, the complex physical mechanisms underlying flash droughts make it difficult to predict their rapid onset using data with low spatial-temporal resolution (Huntingford et al., 2019; Q. Yuan et al., 2020).

Machine learning and deep learning methods have shown promise in long-term traditional drought prediction, although their potential for flash drought prediction has only been explored in a limited number of studies. Feng et al. (2024) used Long Short-Term Memory (LSTM) models to predict soil moisture in Eastern China and further to quantify the temporal contribution of drivers for identifying different mechanisms during drought onsets. J.-L. Zhang et al. (2024) introduced Swin Transformer model to execute spatiotemporal prediction of meteorological drought across multiple scales in Eastern Asia, and achieved acceptable results for flash drought, sustained drought and severe drought. Foroumandi et al. (2024) developed a Generative Adversarial Network (GAN) to predict the Standardized Soil Moisture Index (SSI) for monitoring flash drought in the United States. Du et al. (2024) adopted regression tree model to simulate the soil moisture in Montana based on satellite and geospatial data sets, and successfully captured various phases of the 2017 Montana flash drought. In conclusion, classical machine learning approaches, including Artificial Neural Networks (ANNs), Support Vector Machines (SVM), and Random Forest (RF), have demonstrated effectiveness in managing multicollinearity and capturing non-linear relationships among various drought indicators and drivers (Kumar & Tian, 2024; Park et al., 2016; Tufaner & Özbeyaz, 2020; Zhu et al., 2021). Deep learning methods like LSTM networks and Convolutional Neural Networks (CNN) offer advantages over traditional machine learning approaches by reducing the issue of overfitting spatial-temporal lag components, thereby enhancing performance in handling multi-scale and multi-dimensional data for flash drought prediction (Dai et al., 2025; Dikshit et al., 2021; L. Li, Dai, et al., 2022; C. Xiao et al., 2019; X. Xiao et al., 2024). However, current studies mainly focus on regional and local areas (e.g., China and the United States), and use grid-by-grid training strategy rather than a global incorporation of data samples. Besides, current DL/ML models usually utilize several relevant predictors such as precipitation and temperature and may ignore other underlying predictors. An advanced global-scale model considering spatial-temporal effects from more potential drivers may help improve the soil moisture and flash drought prediction.

However, data-driven machine learning and deep learning methods are primarily designed for prediction and classification, rather than for discovering and quantifying interdependencies within the underlying system (Runge et al., 2019). To address this, causal inference methods offer a valuable complement to predictive machine learning, enhancing theoretical understanding by modeling the system as a potential deterministic mathematical framework (Reichstein et al., 2019). Some studies have succeed in combining causality with ML model for flash drought prediction via a straightforward decoupling approach (Dai et al., 2025; Kumar & Tian, 2024). Causality analysis can identify spatial-temporally related drivers for subsequent prediction, but do not directly interfere with ML models. More studies have shown that incorporating additional insights from dynamical systems into data-driven models can substantially improve model performance and generalization (H. Cai et al., 2022; Hoedt et al., 2021; Jiang et al., 2020; Zhao et al., 2019). Therefore, a deep coupling of causality analysis and DL models is expected to achieve better performance in soil moisture and flash drought prediction. Furthermore, a more accurate model enhances the reliability of its interpretation through eXplainable Artificial Intelligence (XAI)

techniques (Molnar, 2020). Recently, XAI has attracted considerable research interest, demonstrating the capacity to reveal new insights in geoscience fields such as flooding mechanisms (Jiang et al., 2024), ecosystem response (W. Li, Dai, et al., 2022), and flash drought dynamics (Feng et al., 2024).

This study introduces a causality-integrated module into a traditional CNN-LSTM (convolutional neural network with long short-term memory) backbone to enable accurate soil moisture prediction, as soil moisture is one of the most widely used indicators for flash droughts. Using this model, an XAI technique was applied to uncover the mechanisms behind flash drought onset in comparison with normal droughts. Specifically, we developed a Residual Attention Causal Recurrent (ResAttCauRec) model that predicts soil moisture based on daily spatial-temporal meteorological driver maps. We then examined how the causality module contributes to model improvements over a baseline, particularly for forecast of flash drought events. Additionally, we employed an XAI model interpretation algorithm, expected gradients, to evaluate the contributions of key drivers to the onset of both flash and normal drought events. Our analysis revealed simultaneous trends of flash drought acceleration and increasing complexity, underscoring their potential impacts on future flash drought predictions. This causality-based framework introduces interpretable causal signals into the deep learning model, enhancing its performance and offering potential for extension to other extreme events, such as floods and heatwaves. Furthermore, our study advances the understanding of onset mechanisms for both flash and normal droughts.

2. Model Development

To enhance insights and forecast for flash droughts, we developed the Residual Attention Causal Recurrent (ResAttCauRec) model (Figure 1a). This model leverages 15 consecutive days of time-series data on various climatic and meteorological drivers across a 10×10 grid to predict soil moisture at the central point seven days into the future.

2.1. Model Architecture

Recent studies emphasize that flash drought events are influenced by multiscale land-atmosphere-ocean interactions involving local indicators and climatic drivers (Liang & Yuan, 2021; Nguyen et al., 2021). A CNN-LSTM backbone was thus chosen for its capability to effectively extract information from spatial-temporal data (X. Li et al., 2017; R. Yang, Li, et al., 2020), essential for accurate soil moisture prediction. Traditional deep learning models are generally considered to perform poorly beyond their calibration range, showing limited generalization, especially when forecasting extreme events (Plésiat et al., 2024; Trok et al., 2024). These limitations stem, in part, from the lack of embedded prior knowledge—such as physical laws, processes, and causal relationships—in purely data-driven models. In this study, we enhanced the CNN-LSTM backbone by integrating advanced components, including attention mechanisms and a causality module, to improve model performance and generate actionable insights for flash drought early warning and risk management (H. Cai et al., 2022; Hoedt et al., 2021; Jiang et al., 2020; Zhao et al., 2019).

The input data is structured as a four-dimensional array (sequence $\times$ channel $\times$ height $\times$ width), where "sequence" represents the length of the time series, "channel" indicates the number of climatic and meteorological forcings, and "height" and "width" define the focused area, set to a 10×10 grid in this study. The input is first fed into an Attention CNN module to extract the spatial information (Plésiat et al., 2024). Each timeframe in the time series is treated individually to facilitate parallel computing. An averaging pooling layer follows the attention CNN module, embedding spatial information into channels by reducing the area to a single point (A. Zhang et al., 2021). The reshaped data is then transmitted to a causality-informed LSTM module, which is responsible for the temporal information extraction using treated time series data for multiple drivers (Mousavi et al., 2020). The last hidden state of the causality-informed LSTM module is used to provide the soil moisture prediction through a fully connected layer.

2.2. Residual Attention Module

The ResAtt (Residual Attention) module functions as the attention CNN within the ResAttCauRec model, responsible for extracting spatial information. In this study, we employed a classic ResNet (Residual Network) architecture, known for stabilizing training and convergence in deep learning models (K. He et al., 2016), to capture the spatial attributes of surrounding climatic and meteorological conditions relevant to soil moisture at the target site (Venkatesan & Li, 2017). Inspired by human visual attention, attention mechanisms (Vaswani

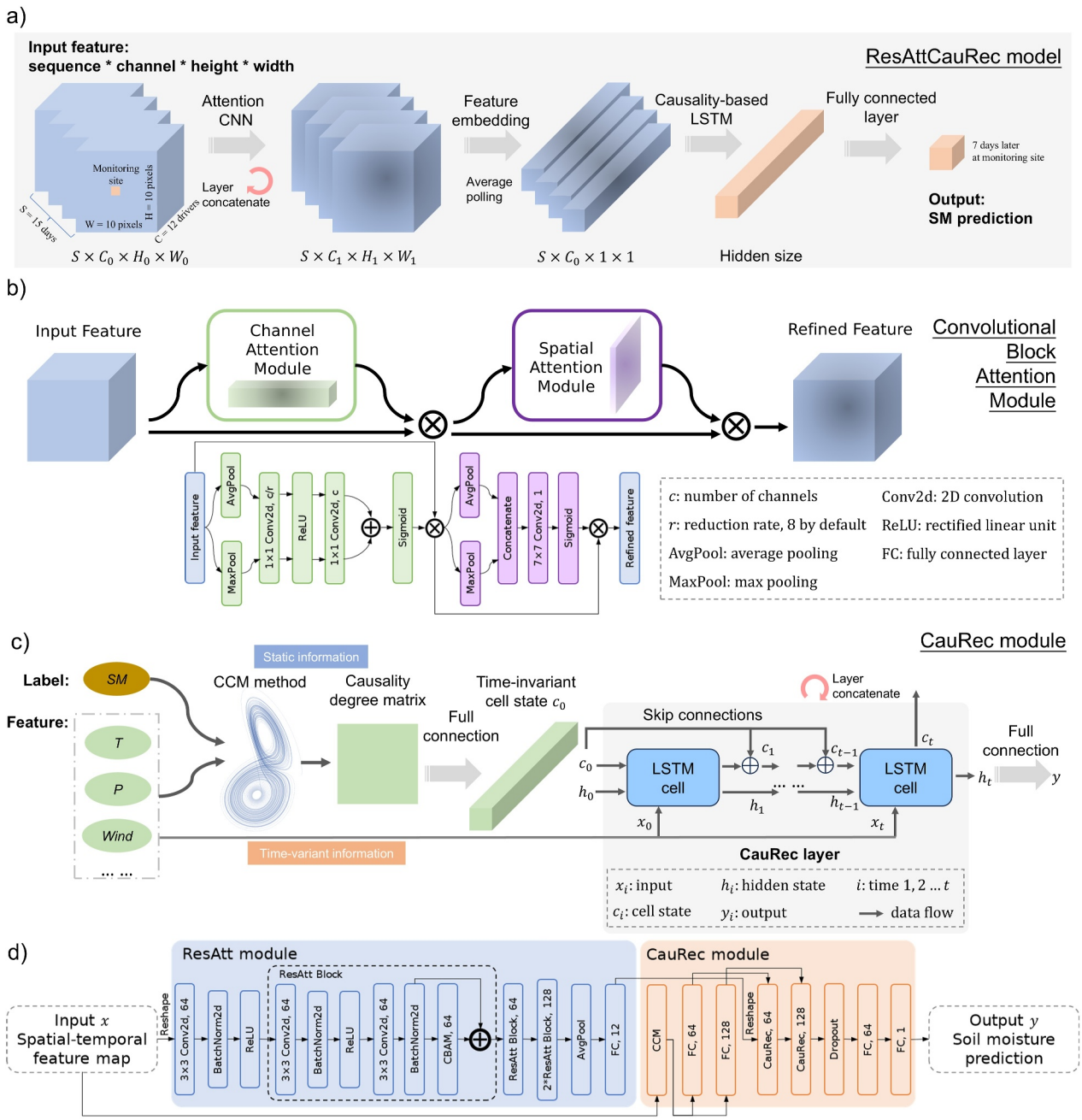

Figure 1. (a) Diagram of the proposed ResAttCauRec (Residual Attention Causal Recurrent) model, where soil moisture prediction is achieved through sequential attention CNN and causality-based LSTM modules, utilizing spatial-temporal meteorological driver maps. (b) Implementation of the Convolutional Block Attention Module (CBAM), the core component of the ResAtt (Residual Attention) module. (c) Schematic and implementation of the CauRec (Causal Recurrent) module, a causality-informed LSTM module. (d) Detailed architecture and data flow within the ResAttCauRec model.

et al., 2017), have been widely adopted in deep learning models to selectively focus on informative input features while reducing the impact of irrelevant or noisy data. By assigning varying weights to input elements based on their relevance to the prediction task, attention mechanisms improve model interpretability and robustness, particularly in complex and heterogeneous data scenarios. Although not traditionally classified as such, the attention mechanism can be viewed as a form of regularization. Vaswani et al. (2017) demonstrated that attention introduces a structured inductive bias by directing the model to focus on specific parts of the input, which aids in generalization by preventing overfitting to noisy or irrelevant data. Recent research on advanced natural language models, such as BERT (Devlin, 2018) and GPT (Achiam et al., 2023), indicates that the self-attention mechanism

often enhances generalization in subsequent tasks, suggesting it serves as an implicit regularizer by enabling the model to capture long-range dependencies more effectively. The ResAtt module integrates the ResNet architecture and attention mechanisms to effectively capture spatial information crucial for flash drought prediction.

The attention implementation (Figure 1b) follows the principles of the Convolutional Block Attention Module (CBAM) (Woo et al., 2018). Given a feature map $\mathbf{F} \in \mathbb{R}^{C\times H\times W}$ as input, CBAM sequentially generates attention maps across two distinct dimensions—channel and spatial—producing the final output as follows:

$$\mathbf{F}' = \mathbf{M_C}(\mathbf{F}) \odot \mathbf{F},$$
$$\mathbf{F}'' = \mathbf{M_S}(\mathbf{F}') \odot \mathbf{F}', \tag{1}$$

where $\mathbf{M_C}$ represents the channel attention map, $\mathbf{M_S}$ is the spatial attention map, and $\odot$ denotes element-wise multiplication. Channel attention focuses on identifying "what" is meaningful across the input channels. As illustrated in Figure S1b in Supporting Information S1, it can be computed as follows:

$$\mathbf{M_C}(\mathbf{F}) = \sigma(\mathbf{MLP}(\mathbf{AvgPool}(\mathbf{F})) + \mathbf{MLP}(\mathbf{MaxPool}(\mathbf{F})) \tag{2}$$

where σ represents the sigmoid function, and MLP denoted the shared multi-layer perceptron with one hidden layer.

In contrast to channel attention, spatial attention identifies "where" informative parts are located, complementing the channel attention (Figure S1c in Supporting Information S1). It is computed as follows:

$$\mathbf{M_S}(\mathbf{F}) = \sigma\left(f^{7\times7}([\mathbf{AvgPool}(\mathbf{F});\ \mathbf{MaxPool}(\mathbf{F})])\right) \tag{3}$$

where σ denotes the sigmoid function, and $f^{7\times7}$ represents a convolution operation with the filter size of 7×7.

Previous studies have indicates that this additional attention architecture enhances both classification and regression performance across various models (L. Chen et al., 2021; Mohammadi Foumani et al., 2024). As shown in Figure 1d, the basic ResAtt block in this study consists of two convolutional layers followed by a CBAM, and four ResAtt blocks form the backbone of the ResAtt module. Notably, the convolutional layers do not process temporal information in this context. Therefore, the temporal dimension S of the original input $\boldsymbol{x} \in \mathbb{R}^{n\times S\times C\times H\times W}$ is flatten into batch size n, to generate the input $\boldsymbol{x'} \in \mathbb{R}^{nS\times C\times H\times W}$ for the convolutional layers. After average pooling, which reduces the spatial information to a single point, another reshaping operation restores the time dimension. The ResAtt module's output is thus formatted to match the input size ($\boldsymbol{x''} \in \mathbb{R}^{n\times S\times C}$) for the subsequent recurrent module.

2.3. Causality Recurrent Module

The Causal Recurrent (CauRec) module is a causality-informed LSTM architecture designed to extract temporal information by embedding causal relationships within the cell state of a conventional LSTM cell, as shown in Figure 1c. The CauRec module categorizes LSTM inputs into two types: traditional time-varying information (forcing variables) and additional static information (causality degrees). Forcing variables, such as temperature, precipitation, and pressure, function as time-varying inputs of LSTM unit. The additional causal information, treated as static input, is derived from the CCM (Convergent Cross Mapping) causality degree between meteorological variables and soil moisture in neighboring pixels. This causality degree matrix is first processed through fully connected layers to create an initial cell state c_0, which is then skip-connected with all subsequent cell states to enhance the influence of static information throughout the sequence.

Typically, hidden and cell states in LSTM models are initialized to zero (Sak et al., 2014). Previous studies have shown that learning these initial states as parameters can improve model performance (Hwang, 2020; Pitis, 2016; Wenke & Fleming, 2019). In our approach, we train the initial states to encode interpretable causal information. The skip connections in our model helps retain static information, inspired by ResNet's gradient-preserving approach in deep learning (K. He et al., 2016). The skip connections allow all subsequent cell states to carry essential long-term information, specifically global causality. The two-step approach embeds causal signals

information directly into the model and is expected to provide weighting strategy of the feature dimension, encouraging the subsequent LSTM layers to prioritize crucial features. Such refinement is essential for modeling the sequential nature of soil moisture fluctuations, thereby improving the model's understanding of temporal dependencies on different drivers and the interpretability of its predictions (Lehmann et al., 2020).

LSTM networks, originally introduced by Hochreiter and Schmidhuber (1997), excel at learning long-term dependencies between input and output features in sequence prediction tasks. This capability is particularly beneficial for simulating complex drought dynamics, such as the progression of meteorological drought conditions to soil moisture drought, and for capturing the memory effects of meteorological inputs over extended periods. A typical LSTM unit consists of a memory cell and three key components: an input gate, a forget gate, and an output gate (Figure S2 in Supporting Information S1). The cell functions as a memory unit, while the gates control the flow of information in and out of the cell (Konapala et al., 2020). The gates are computed as follows:

$$\begin{aligned}\mathbf{I}_t &= \sigma(\boldsymbol{x}_t\boldsymbol{w}_{\mathrm{xi}} + \boldsymbol{h}_{t-1}\boldsymbol{w}_{\mathrm{hi}} + \boldsymbol{b}_{\mathrm{i}}),\\ \mathbf{F}_t &= \sigma\big(\boldsymbol{x}_t\boldsymbol{w}_{\mathrm{xf}} + \boldsymbol{h}_{t-1}\boldsymbol{w}_{\mathrm{hf}} + \boldsymbol{b}_{\mathrm{f}}\big),\\ \mathbf{O}_t &= \sigma(\boldsymbol{x}_t\boldsymbol{w}_{\mathrm{xo}} + \boldsymbol{h}_{t-1}\boldsymbol{w}_{\mathrm{ho}} + \boldsymbol{b}_{\mathrm{o}}),\end{aligned} \tag{4}$$

where $\mathbf{I}_t$ is the input gate, $\mathbf{F}_t$ is the forget gate, $\mathbf{O}_t$ is the output gate; $\boldsymbol{x}_t$ is the input, $\boldsymbol{h}_t$ represents the current hidden state; $\boldsymbol{w}_{\mathrm{xi}}, \boldsymbol{w}_{\mathrm{hi}}, \boldsymbol{w}_{\mathrm{xf}}, \boldsymbol{w}_{\mathrm{hf}}, \boldsymbol{w}_{\mathrm{xo}}, \boldsymbol{w}_{\mathrm{ho}}$ are weight parameters and $\boldsymbol{b}_{\mathrm{i}}, \boldsymbol{b}_{\mathrm{f}}, \boldsymbol{b}_{\mathrm{o}}$ are bias parameters. As depicted in Figure 1c, the hidden state $\boldsymbol{h}_t$ signifies short-term memory, while the cell state $\boldsymbol{c}_t$ represents long-term memory. Both hidden state and cell state are passed to the cell of the next time step, and are controlled by an input node $\tilde{\boldsymbol{c}}_t$:

$$\tilde{\boldsymbol{c}}_t = tanh(\boldsymbol{x}_t\boldsymbol{w}_{\mathrm{xc}} + \boldsymbol{h}_{t-1}\boldsymbol{w}_{\mathrm{hc}} + \boldsymbol{b}_{\mathrm{c}}), \tag{5}$$

where $\boldsymbol{w}_{\mathrm{xc}}$ and $\boldsymbol{w}_{\mathrm{hc}}$ are weight parameters and $\boldsymbol{b}_{\mathrm{c}}$ is bias parameter. The input gate $\mathbf{I}_t$ governs how much we take new data into account via $\tilde{\boldsymbol{c}}_t$ and the forget gate $\mathbf{F}_t$ addresses how much of the old cell state $\boldsymbol{c}_{t-1}$ we retain. In our modification, we aim for the cell state $\boldsymbol{c}_t$ capture and convey essential long-term information, specifically global causality. To achieve this, we use causality information to initialize the cell state $\boldsymbol{c}_0$, reinforcing this information in subsequent cell states through a skip connection. In mathematical expressions, these computations are performed as follows:

$$\begin{aligned}\boldsymbol{c}_0 &= \mathbf{MLP}(\mathbf{CCM}(\boldsymbol{X}, \boldsymbol{Y})),\\ \boldsymbol{c}_t &= \mathbf{F}_t \odot (\boldsymbol{c}_{t-1} + \boldsymbol{c}_0) + \mathbf{I}_t \odot \tilde{\boldsymbol{c}}_t,\\ \boldsymbol{h}_t &= \mathbf{O}_t \odot \boldsymbol{tanh}(\boldsymbol{c}_t)\end{aligned} \tag{6}$$

where $\mathbf{CCM}(\boldsymbol{X}, \boldsymbol{Y})$ represents the causality degree between input and output of training data set, calculated via the CCM approach. And $\odot$ represents elementwise product operator. As a comparison, basic LSTM networks generally initiate hidden state $\boldsymbol{h}_0$ and cell state $\boldsymbol{c}_0$ with zero or random matrices, and cell state $\boldsymbol{c}_t$ is directly derived from $\boldsymbol{c}_{t-1}$ instead of $\boldsymbol{c}_0$.

3. Data and Methods

3.1. FLUXNET2015 Data Set

In recent decades, satellite systems and land surface models have been employed to estimate surface soil conditions on a global scale. However, these satellite and reanalysis products have shown limited performance in soil moisture evaluation, with accuracy highly influenced by factors such as regional, climatic, land cover, and topographic variations (Kim et al., 2020; S. Yang, Li, et al., 2020; Zheng et al., 2024). Therefore, an in-situ measurement product, FLUXNET2015 data set, was selected as the ground truth of soil moisture labels in our model.

The FLUXNET2015 data set (Pastorello et al., 2020) provides comprehensive meteorological and biological measurements at the ecosystem level, with data collected from 212 sites worldwide. Covering over 1,500 site-

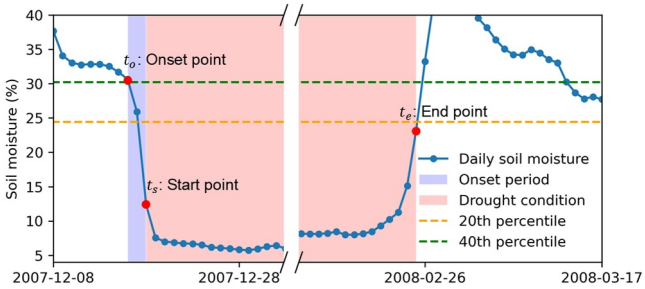

Figure 2. Definition of flash droughts, illustrated with a typical event observed at site AT-Neu (Neustift im Stubaital, Austria). In this event, soil moisture dropped from above the 40th percentile (onset point t_o) to below the twentieth percentile (start point t_s) within 2 days, and remained below the twentieth percentile for over 2 months until the final day (end point t_e). The orange and green dashed lines represent the wet condition (40th percentile) and the dry condition (twentieth percentile) condition of soil moisture during the entire monitoring period, respectively. The purple shaded area indicates the onset period of flash drought, while the pink shaded area represents the drought condition.

years up to 2014, the data set enabled us to focus on daily soil moisture observations (SWC_F_MDS_1) as the target labels. To ensure data quality, we filtered out sites with over 10% invalid values or less than one year of observations, resulting in 156 valid sites with an average coverage of 6.5 years, spanning from 1996 to 2014. We applied linear interpolation for two sites, CA-Gro and CA-Oas. At CA-Gro, 7.8% observations were invalid, with the longest consecutive gap spanning 8 days. At CA-Oas, 6.8% were missing, with the longest consecutive gap lasting 10 days.

Flash drought events were identified from soil moisture observations based on a rapid onset intensification rate (Qing et al., 2022). As illustrated in Figure 2, a typical flash drought event is characterized by a drop in soil moisture from the 40th to the 20th percentile, with an average decline rate of no less than one percentile per day; otherwise, the event is classified as a normal drought (X. Yuan et al., 2023). Given the limited duration of historical observations, we define percentiles over the entire observation period to allow for a comparison of absolute soil moisture changes. The last day when soil moisture exceeds the 40th percentile marks the onset point. The drought condition begins at the first point where soil moisture falls below the twentieth percentile and ends at the last point where it rises above the twentieth percentile. Additionally, the intensification rate during the onset period is mathematically defined as follows (Qing et al., 2022):

$$\text{intensification rate} = \frac{p_{t_o} - p_{t_s}}{t_s - t_o} \tag{7}$$

where p_t indicates the percentile corresponding to time t. An event with an intensification rate not less than one percentile per day is identified as a flash drought, otherwise it is a normal drought. Short-term fluctuations in soil moisture can artificially segment a single drought event into multiple occurrences (as shown in Figure S3 of Supporting Information S1). To address this, we now require that soil moisture must exceed the twentieth percentile threshold and remain above this level for at least one pentad (5 consecutive days) before a drought event is considered terminated (Mahto & Mishra, 2024). Using this definition, a total of 892 flash drought events were identified across 131 sites (Table S1 in Supporting Information S1). All identified drought events last more than 14 days, aligning with commonly used definitions that require drought conditions to persist for at least three pentads (X. Yuan et al., 2019) or 2 weeks (Osman et al., 2020).

3.2. ERA5 Reanalysis Data

To meet the needs of drought risk management, the ERA5 reanalysis data set (Hersbach et al., 2020) was adopted to provide near real-time, gap-free meteorological maps as model inputs. Produced by the European Centre for Medium-Range Weather Forecasts (ECMWF), ERA5 represents a state-of-the-art resource for studying global atmospheric and surface conditions at high spatial and temporal resolutions (0.25-degree grid and hourly intervals).

In this study, a comprehensive set of 12 variables was selected to assess and forecast soil moisture dynamics. These variables include 2 m dewpoint temperature (DewT), 2 m mean and maximum temperatures (MeanT, MaxT), convective available potential energy (CAPE), precipitation (Prec), evaporation (Evap), surface pressure (PSFC), relative humidity (Humid), 500 hPa geopotential height (Z500), 10 m wind speed components (Wind), and soil temperature and water content (SoilT, SoilW). Notably, the ERA5 data set provides 10 m u- and v-components of wind, representing the eastward and northward components, respectively. Here, we derived the absolute wind speed from these components using vector addition $\left(\sqrt{u^2 + v^2}\right)$, disregarding directional information.

The selection of meteorological drivers was based on their established influence on soil moisture dynamics and flash drought onset. Air temperature affects evapotranspiration rates and soil drying, while dewpoint temperature serves as an indicator of atmospheric moisture availability, influencing evaporative demand. Precipitation is the

primary source of soil moisture recharge, while evaporation and relative humidity regulate moisture loss to the atmosphere. These water- and temperature-related variables have been recurrently linked to flash drought occurrences (P. Li et al., 2024; Mahto & Mishra, 2023; Mo & Lettenmaier, 2015; L. Wang & Yuan, 2018; M. Zhang et al., 2022). Surface pressure reflects large-scale atmospheric conditions influencing precipitation and moisture transport (Loehrer & Johnson, 1995; G. Yuan et al., 2021). Z500 represents mid-tropospheric circulation patterns associated with drought formation (Bakke et al., 2023; Faranda et al., 2023). CAPE indicates atmospheric instability, which affects convective rainfall and subsequent soil moisture changes (Barkiđija & Fuchs, 2013; Myoung & Nielsen-Gammon, 2010). In addition, pressure-driven wind directly influences surface drying by enhancing evaporation and altering moisture transport (Davarzani et al., 2014; McVicar et al., 2012). Finally, soil temperature and soil water content directly govern soil moisture retention and energy balance, making them critical for forecasting soil moisture and drought evolution.

The prediction horizon is set to 7 days, following previous studies (Feng et al., 2024; Kumar & Tian, 2024; L. Li, Dai, et al., 2022). This timeframe allows for timely management interventions in response to emerging soil moisture anomalies, particularly during the critical early stages of flash drought development. In addition, we conducted lead-time sensitivity analysis to illustrate the model performance with different prediction horizons (from one day to 2 weeks). To ensure the model captures sufficient spatial information, daily data from a 10×10 grid (0.25-degree resolution) surrounding the target site were extracted from the original data set and used as model inputs. We chose a 15-day input window based on references to previous studies that adopted input lengths ranging from 10 days (L. Li, Dai, et al., 2022) to 17 pentads (Feng et al., 2024), particularly for short-to medium-term forecasting tasks. We conducted a sensitivity analysis on this hyperparameter to discuss the impact of the length of input features on flash drought prediction.

3.3. Convergent Cross Mapping (CCM) for Causal Discovery

In the field of hydrology, various approaches have emerged in recent decades to infer causal relationships from observational data. These methodologies can be broadly categorized into four groups: those based on linear and nonlinear autoregressive modeling (e.g., Granger causality and Transfer Entropy), graph-based techniques (such as the Peter-Clark algorithm), and methods grounded in the theory of time-delay embedding (e.g., Convergent Cross Mapping, CCM). The first three approaches assume that the system is primarily stochastic, whereas CCM is grounded in dynamical systems theory, which considers causality as arising from underlying deterministic interactions.

CCM was specifically designed to identify nonlinear, state-dependent causal relationships in complex systems (Sugihara et al., 2012). Unlike stochasticity-based approaches, CCM reconstructs the system's attractor using Takens' theorem (Takens, 2006), allowing it to infer causality even in cases where the relationship between variables is weak or indirect. Additionally, CCM can distinguish between direct causal interactions and apparent correlations caused by shared external forcing (Sugihara et al., 2012).

Land–atmosphere interactions, which drive Earth's surface water and energy budgets (Indu et al., 2022), are governed by multiple coupled differential equations (Brubaker & Entekhabi, 1996). In these interactions, soil moisture plays a key role in partitioning energy and water fluxes. However, the underlying dynamics exhibit nonlinear and weakly coupled behavior, often displaying chaotic characteristics that make them difficult to analyze using purely observational data (Lorenz & Haman, 1996; Shen et al., 2021). Since phase-space reconstruction can effectively capture the properties of such a deterministic dynamical system (Takens, 2006), CCM provides a more suitable framework for quantifying causal relationships in this context compared to methods that assume stochasticity (Y. Wang et al., 2018).

The basic principle of CCM involves reconstructing system states $Y = f(X, Y)$ from two time series variables and then quantifying the causality between them using nearest neighbor forecasting (Sugihara & May 1990). Generally, we use the shadow manifolds generated by the method of time delay embedding because we do not know the true manifold of the system (Packard et al., 1980). Given Takens' theorem, time delay embedding uses successive lags of a single time series to compute shadow manifolds which can cross map with 1:1 correspondence to the true manifold of the system (Takens, 2006). The shadow manifolds M_x and M_y can be regarded as summaries of X and Y (Luo et al., 2014). In the case of a system where X causes Y, information from X get embedded in Y. CCM quantifies this relationship using simplex projection to estimate the X from M_y, that is

$\hat{X}(t)|M_y$. Finally, the accuracy of the estimation can be the metric for causality. Besides, the convergence in CCM means that for the variables with causalities, the longer the time series, the better performance of prediction can be expected (Sugihara et al., 2012).

Specifically, given two time series $X = \{X(1), X(2), ..., X(L)\}$ and $Y = \{Y(1), Y(2), ..., Y(L)\}$ where L is the length of time series, to check whether X forces Y or not, the phase space reconstruction of Y should be firstly conducted.

1. An E-dimensional reconstruction uses E successive lags of Y, each separated by a time step τ: $\underline{y}(t) = < Y(t), Y(t-\tau), Y(t-2\tau), ..., Y(t-(E-1)\tau) >$ for $t \in [1 + (E-1)\tau, L]$. The value of embedding dimension E depends on several factors including system complexity, time series length, and noise (Kennel et al., 1992), can be optimally identified using search algorithms (e.g., $E = 2$ in this study via grid search). Since most time series were not overly sampled in time, we fixed the time lag $\tau = 1$. Then the shadow manifold of Y can be defined as $M_y = \left\{\underline{y}(t) \text{ for each } t \in [1 + (E-1)\tau, L]\right\}$.
2. From the selected vector $\underline{y}(t)$, find $E + 1$ nearest neighbors. The prediction model of X can be given as $\hat{X}(t)|M_y = \sum w_i X(t_i)$ where $i = 1, 2 \ldots E + 1$. The weight w_i can be calculated as follows:

$$w_i = \frac{u_i}{\sum u_i},$$

$$u_i = \mathbf{exp}\left[-\frac{d\left(\underline{x}(t), \underline{x}(t_i)\right)}{d\left(\underline{x}(t), \underline{x}(t_1)\right)}\right] \tag{8}$$

   where $d\left(\underline{x}(t), \underline{x}(t_i)\right)$ is a Euclidean distance.
3. If X causes Y, the estimate $\hat{X}(t)\big|M_y$ is expected to converge to $Y(t)$ as L increases because a denser cluster of $E + 1$ points will be used in prediction (Sugihara et al., 2012). Therefore, we can plot the correlation coefficients between $\hat{X}(t)\big|M_y$ and $Y(t)$ check the significance and convergency. Figure S4 in Supporting Information S1 shows a unidirectional causation from X to Y.

In this study, we evaluated the bidirectional causality between meteorological variables and soil moisture within the training data set. Because FLUXNET sites do not perfectly align with ERA5 grid cells, we paired each site with the four nearest ERA5 pixels to minimize spatial mismatch and potential bias. It is important to note that this procedure was used exclusively for calculating the local causality degree, not for model input, which was still based on the surrounding 10×10 grid cells. The correlation coefficients and their significance levels from these four pixels were combined into a causality degree matrix, which was then fed into fully connected layers to generate static information. This approach enhances the representation of persistent causal signals within the LSTM cell states via skip connections (Figure 1c), reinforcing the influence of physically plausible drivers throughout the sequence.

3.4. Model Implementation and Training

The ResAttCauRec model proposed in this study was trained on the global ERA5 and FLUXNET data set. For each site, the time series was chronologically divided into training, validation, and test data sets in a 70:15:15 ratio (Bishop & Nasrabadi, 2006; Y. Xu & Goodacre, 2018). The validation set was used to monitor the training process, and the model exhibiting the best performance on the validation set was selected to effectively prevent overfitting (Prechelt, 2002). The test set was reserved for the final evaluation of the model's generalization ability. To facilitate model convergence (Grus, 2019), we applied z-score normalization to all input features, calculated as:

$$z = \frac{x - \mu}{\sigma} \tag{9}$$

where z represents the standard score after conversion, x denotes the raw score; μ is the mean of the population, and σ is the standard deviation. The normalization parameters (μ and σ) were computed exclusively from the training data set and subsequently applied to the validation and test data sets to avoid data leakage. To ensure independence among samples despite their sequential nature, input features (with dimensions $15 \times 12 \times 10 \times 10$) and corresponding labels (of length 1) were paired and then shuffled independently within each of the three data sets (Han et al., 2019).

Model architecture hyperparameters (e.g., the number of layers in the convolutional and recurrent modules) were selected based on a trade-off between performance and model complexity (Bengio, 2012; Goodfellow et al., 2016). Due to limited computational resources, it was impractical to conduct an exhaustive hyperparameter search. Instead, we manually tuned the architecture through empirical testing, guided by general heuristics such as the ten-times rule and scaling laws (Bahri et al., 2024). The detailed parameter sizes of the ResAttCauRec model are provided in Table S2 of Supporting Information S1.

The training hyperparameters define how the model learns during training. Among them, the learning rate and batch size jointly have a strong impact on convergence behavior and training stability (F. He et al., 2019). We experimented with various combinations of batch sizes (ranging from 32 to 512) and learning rates (from 1e−4 to 1e−1) to identify an optimal setting (Balles et al., 2016). Based on convergence performance and computational efficiency, we selected a batch size of 512 and a learning rate of 1e−4. The Adam optimizer, a variant of stochastic gradient descent, was used for model training because its adaptive learning rates make it generally less sensitive to the choice of initial learning rate compared to other optimization algorithms (Kingma & Ba, 2014).

The number of training epochs was treated as a hyperparameter to be optimized in order to prevent overfitting. We trained the model for 50 epochs, with each epoch processing approximately 260,000 samples. The epoch that achieved the best performance on the validation data set was selected for subsequent evaluation and analysis. This training process took around 5.5 hr on a single NVIDIA GeForce RTX 3060 12G. In addition, we conducted 10 independent training runs to ensure robustness and stability of the results (Mienye & Sun, 2022).

The aforementioned data set partitioning scheme allows for the evaluation of models' temporal generalization. However, its applicability to other regions with varying climatic conditions and soil properties remains insufficiently assessed. To address this limitation and provide a more comprehensive evaluation of the model's spatial generalization, a spatial Monte Carlo cross-validation approach (Q.-S. Xu & Liang, 2001) was used, involving five independent random splits (70% training, 15% validation, 15% test). The average performance across the five test sets serves as a robust estimate of the model's spatial generalizability.

3.5. Ablation Study and Metrics

We conducted an ablation study to evaluate the effectiveness of the ResAttCauRec model. By removing the causality module, we obtained the ResAttRec model. Further removing the attention module resulted in the ResRec model, which serves as the basic CNN-LSTM backbone. The baseline Rec model is a fundamental LSTM model that focuses solely on the observation series near the monitoring site, disregarding spatial information from more distant grids.

The ablation models are not fully independent. They serve as components or simplified versions of ResAttCauRec. Therefore, it is essential to include an independent benchmark model with a distinct architecture to more clearly demonstrate the advantages of the proposed method. U-Net, a classic CNN-based architecture originally designed for image segmentation (Ronneberger et al., 2015), and has been widely adopted in geoscience applications such as atmospheric river detection (Tian et al., 2024), precipitation prediction (Tong et al., 2024) as well as flash drought identification (Foroumandi et al., 2024). In this study, we implement a modified U-Net model as the benchmark. It retains the standard encoder–decoder structure, with a convolutional encoder for downsampling and a symmetric decoder for upsampling (Ronneberger et al., 2015). To adapt U-Net for our soil moisture regression task, we replace the Conv2d layers with Conv3d layers to accommodate temporal sequences. Additionally, we append a global average pooling layer followed by a fully connected layer at the end, allowing the model to output a scalar value. The feature sizes of the convolutional layers are set to [24, 32, 40, 64], ensuring the total number of trainable parameters (848,473) is comparable to that of the ResAttCauRec model (839,961).

Mean Squared Error (MSE) is the most widely used loss function for regression tasks. In this study, we utilized MSE as the loss function to train our models and as a criterion to assess their performance. MSE assigns greater weight to larger errors compared to smaller ones, meaning it penalizes outliers more severely. It is also referred to as the L2 norm or the Euclidean distance. MSE measures the average of the squared differences between the simulated values S_i and actual observations O_i, and can be calculated as follow:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(S_i - O_i)^2 \tag{10}$$

where N is the total sample number. The Root Mean Squared Error (RMSE), which is the square root of the MSE, was also used in this study.

Nash-Sutcliffe coefficient of Efficiency (NSE) is a popular criterion in geoscience, which quantifies how well the model simulation can predict the outcome (Nash & Sutcliffe, 1970). The definition of NSE is as follow:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N}(S_i - O_i)^2}{\sum_{i=1}^{N}(O_i - \overline{O})^2} \tag{11}$$

where $\overline{O}$ represents the average of observations. NSE ranges from $-\infty$ to 1. The closer the NSE is to 1, the better the simulation result is.

The Kling–Gupta efficiency (KGE) is based on a decomposition of NSE into its constitutive components (correlation, variability bias and mean bias) (Gupta et al., 2009), and is increasingly used for model calibration and evaluation:

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{\mathrm{sim}}}{\sigma_{\mathrm{obs}}} - 1\right)^2 + \left(\frac{\mu_{\mathrm{sim}}}{\mu_{\mathrm{obs}}} - 1\right)^2} \tag{12}$$

where σ_{sim} is the standard deviation of simulations, σ_{obs} is the standard deviation of observations, μ_{sim} is the simulation mean, and μ_{obs} is the observation mean. Similar to NSE, KGE = 1 indicates a perfect simulation.

Generalizability is a crucial issue for data-driven models, indicating that the models can maintain strong performance when transferred to new data sets. A relevant criterion for assessing generalization ability (GA) evaluates this aspect by comparing the model's performance on the test set to its performance on the training set (Gorgij et al., 2023):

$$\mathrm{GA} = \frac{\mathrm{RMSE}_{\mathrm{train}}}{\mathrm{RMSE}_{\mathrm{test}}} = \frac{\sqrt{\mathrm{MSE}_{\mathrm{train}}}}{\sqrt{\mathrm{MSE}_{\mathrm{test}}}} \tag{13}$$

where RMSE refers to Root Mean Square Error. If the GA value exceeds one, the model is considered undertrained or underfitting; conversely, if the GA value is less than one, the model is regarded as overtrained or overfitting. While overfitting should generally be minimized, a moderate degree of it can be tolerated in deep learning models, especially when balanced by techniques like regularization and cross-validation. However, underfitting is typically undesirable, as it indicates that the model has failed to capture essential patterns in the data (Bartlett et al., 2020). Therefore, the closer the GA value is to one when it is less than one, the better the model's generalization.

We expect the models to effectively identify flash drought events based on soil moisture simulations. The prediction of drought events involves classification rather than regression. Four metrics are used to evaluate classification quality: accuracy, precision, recall, and F1-score (Powers, 2020). These metrics are calculated based on true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), using binary classification as an example. Accuracy is defined as the proportion of all correct classifications, regardless of whether they are positive or negative (Menditto et al., 2007):

$$\mathrm{Accuracy} = \frac{\text{correct classifications}}{\text{total classifications}} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{14}$$

In a balanced data set with a similar number of samples across all classes, accuracy can serve as a coarse measure of model quality. However, when dealing with extreme events that constitute a small proportion of the total days, a high accuracy score may be misleading. Recall, also known as the true positive rate, is mathematically defined as the proportion of actual positives that were correctly identified (Powers, 2020):

$$\mathrm{Recall} = \frac{\text{correctly classified actual positives}}{\text{all actual positives}} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{15}$$

Recall represents the probability of detection, indicating the fraction of flash drought events detected by the models. Precision, on the other hand, is the proportion of all positive predictions that are actually positive. It is mathematically defined as follows (Menditto et al., 2007):

$$\text{Precision} = \frac{\text{correctly classified actual positives}}{\text{all classified positives}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{16}$$

Precision can measure the fraction of actual flash drought events that are classified as flash droughts.

Precision improves as the number of false positives decreases, while recall improves as the number of false negatives decreases. However, false positives and false negatives often exhibit an inverse relationship, leading to a trade-off between precision and recall. To balance this trade-off, the F1 score is derived from the harmonic mean of precision and recall (Powers, 2020):

$$\frac{2}{\text{F1}} = \frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}$$

$$\text{F1} = 2 * \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{17}$$

This metric balances the importance of precision and recall, making it preferable to accuracy for class-imbalanced data sets. When precision and recall are similar in value, the F1 score will also be close to that value. Conversely, when precision and recall differ clearly, the F1 score will resemble the lower of the two metrics.

3.6. XAI-Based Model Interpretation

It is essential to employ Explainable Artificial Intelligence (XAI) techniques to gain insight into the implicit linear or nonlinear interactions between potential drivers and soil moisture, thereby enhancing our understanding of the mechanisms behind flash droughts. In this study, we utilized the Expected Gradients (EG) method, developed by Erion et al. (2021) as an extension of the Integrated Gradients (IG) method (Sundararajan et al., 2017). The details of EG can be found in Text S1 in Supporting Information S1.

Given the spatial-temporal component of the EG values for features in this study, we aggregated the EG values across spatial-temporal grids to represent the overall contribution of feature i (Lundberg, 2017). A positive EG value for a specific feature indicates its role in increasing the predicted value relative to the baseline prediction, while a negative EG value suggests the opposite.

Building on the feature importance scores derived from EG values, we quantify the event-specific variability in feature importance across different drought onset intensification rates, a concept we refer to as drought complexity (Jiang et al., 2024). This metric captures the heterogeneity of physical processes underlying drought onset with varying intensification rates. Specifically, drought complexity is defined as the number of key driving features that substantially influence soil moisture dynamics during the onset period of a drought event. We identify key drivers as those features whose absolute EG values exceed the 80th percentile threshold. The daily count of such key features is used to characterize the complexity on each day. For each drought event, we average this daily count over the onset period to obtain an event-level measure of drought complexity. A higher complexity value indicates a more multifaceted combination of meteorological drivers contributing to drought onset.

4. Results and Discussion

4.1. Soil Moisture Prediction

Figures 3a and 3b show the loss curves of the ResAttCauRec model. Notably, the validation loss decreases in parallel with the training loss during the first 35 epochs. However, in the final 15 epochs, the validation loss stabilizes while the training loss continues to decline, indicating the onset of overfitting. We retained the models that achieved the best validation performance within the 50 training epochs for subsequent analysis. To evaluate the contributions of the attention and causality modules integrated into the CNN-LSTM backbone, we conducted an ablation study. By successively removing these modules, we obtained simplified variants: ResAttRec, ResRec,

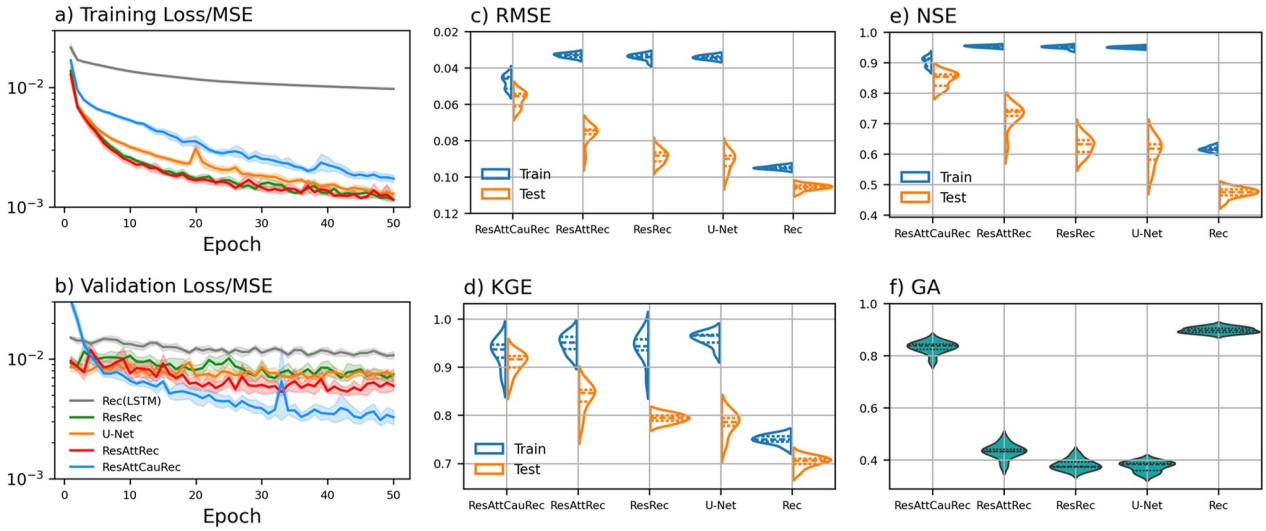

Figure 3. Comparison of the ResAttCauRec model with four other models: ResAttRec, ResRec, Rec (LSTM), and the benchmark U-Net model, in terms of their comprehensive performance across 10 complete training periods. Panels (a, b) display the training and validation loss curves from the 10 training sessions. The light solid line represents the average results of these sessions, with a 95% confidence interval. Panels (c–e) present the performance metrics—RMSE, KGE, and NSE—for the models on both the training and validation sets. (f) Illustrates the generalization ability of the five models. The dashed lines inside the kernel density estimation of violin plots indicate the quartiles.

and the baseline Rec model (a standard LSTM). Additionally, we implemented a U-Net model as an independent benchmark. Although the ResAttCauRec model exhibited lower training set performance, it consistently outperformed the simplified models and benchmark on the test set, underscoring the effectiveness of the added modules (Figures 3c–3e). Figure 3f further highlights that the ResAttCauRec model achieved a mean generalizability score of 0.83, which is noticeably higher than ResAttRec (0.43), ResRec (0.38), and U-Net (0.38), and only slightly lower than the Rec model (0.90).

The results indicate that the baseline Rec model partially captures the temporal signals of soil moisture dynamics, providing relatively reliable predictions, a finding supported by previous studies (Datta & Faroughi, 2023; P. Gao et al., 2021). The incorporation of the spatial information extractor (i.e., residual blocks) substantially enhances model performance; however, this additional spatial information accounts for only a 0.15 increase in NSE on the test data set. In contrast, the NSE on the training set increased by 0.33, primarily due to the extensive model parameters associated with the spatial module.

To bridge the gap in model performance between the training and test sets, regularization techniques are crucial for preventing overfitting and enhancing generalization ability. The introduction of attention improved model generalization, with the average GA increasing from 0.38 (ResRec) to 0.43 (ResAttRec). This improvement can be attributed to enhanced performance on the test set rather than merely mitigating overfitting on the training data. Additionally, we observed that the Conv3d-based U-Net benchmark achieved performance comparable to the CNN-LSTM-based ResRec model. This aligns with previous findings that Conv3d architectures effectively capture spatiotemporal dependencies by simultaneously convolving over spatial and temporal dimensions (Tran et al., 2015). Bai et al. (2018) also reported that well-designed convolutional models can perform comparably to recurrent models in sequence modeling tasks.

The results indicate that the causality module enhances model generalization as a form of regularization technique. Unlike the attention module, which enables the model to assign varying weights to different spatial locations, the causality module is designed to provide explicit strategic weighting of the feature dimension, encouraging the subsequent LSTM layers to prioritize crucial features (Hwang, 2020; Pitis, 2016; Wenke & Fleming, 2019). This is evidenced by the decrease in the NSE metric on the training set (from 0.95 with ResAttRec to 0.90 with ResAttCauRec, on average) and the increase on the test set (from 0.72 to 0.84, on average). The ResAttCauRec model exhibits similar generalization to the Rec model, with an average GA of 0.83, while achieving substantial improvements in model metrics on unseen data. This suggests that the proposed ResAttCauRec model effectively balances model complexity and generalization. Additionally, the results

highlight the potential benefits of guiding data-driven models with causality-based signals (H. Cai et al., 2022; Hoedt et al., 2021; Jiang et al., 2020; Zhao et al., 2019), particularly for dynamical systems like soil moisture. The detailed regression performance of the proposed model at each site is presented in Table S3 of Supporting Information S1.

To provide a more comprehensive evaluation of the model's spatial generalization capability, we conducted a spatial Monte Carlo cross-validation. Unlike temporal splits where the test data come from the same spatial distribution as the training data, spatial cross-validation forces the model to make predictions at entirely unseen locations. During training, we observed greater fluctuations in validation loss (Figure S5a in Supporting Information S1) compared to the temporal split scenario (Figure 3b), suggesting higher spatial uncertainty and heterogeneity in the data distribution. This is expected, as geophysical processes (e.g., soil properties, land cover, or microclimate conditions) influencing flash droughts may vary significantly across regions, making generalization more difficult. Overall, model performance under spatial cross-validation was slightly lower across all evaluation metrics (Figures S5b–S5d in Supporting Information S1). This aligns with findings from prior studies (Wadoux et al., 2021), which suggest that spatial cross-validation represents a lower-bound estimate of model performance due to the need for spatial extrapolation beyond the training domain. Nevertheless, the proposed ResAttCauRec model achieved an average GA score of 0.84 (Figure S5e in Supporting Information S1), which is comparable to the temporal split scenario, indicating that the model maintains a strong capacity to generalize across space.

4.2. Flash Drought Forecasts

Although previous studies have focused on soil moisture prediction and achieved acceptable results using various models, such as Y. Cai et al. (2019) and Kornelsen and Coulibaly (2014), limited studies have reported on the application of these models for drought event alarms. Data-driven models, particularly those trained on large data sets, tend to prioritize the most frequent patterns. This bias enables them to perform well under normal conditions but may lead to underperformance during rare events, such as flash droughts.

Figure 4 illustrates the continuous soil moisture predictions of the proposed model for drought event identification. Taking the 4-year soil moisture monitoring data from site BE-Vie (Vielsalm, Belgium) as an example, an annual drought occurrence pattern emerges, with drought events typically starting in July and ending in October. The baseline Rec model's predictions exhibit an overall upward bias, failing to track the periodic changes effectively. In contrast, the ResAttCauRec model successfully captured almost all drought events, except for one flash drought event that occurred in September 2020, during which the model slightly overestimated soil moisture levels. Notably, the flash drought event in September 2020 did not display extremely low soil water content but rather values close to the threshold (20th percentile) of drought conditions. This slight bias around the threshold led to a substantially different classification result.

In Figure 4b, a similar phenomenon is observed at site US-Me2 (Metolius, USA), where the daily minimum soil water content is approximately 10%. While most normal drought events were correctly identified, a mismatch occurred in July 2010: a predicted drought event closely followed an actual one but did not temporally overlap. Despite this discrepancy in classification, the predicted soil moisture closely matched the observed values during the period. This sensitivity of classification to specific thresholds underscores the challenges associated with drought early warning systems.

Although the ResAttCauRec model achieves strong performance in soil moisture regression, as indicated by consistently lower test set errors and higher GA values compared to other models, it exhibits an even greater relative advantage in accurately capturing flash drought events. As shown in Table 1, the accuracy of all five models correlates closely with their fitting ability illustrated in Figure 3. The ResAttCauRec model achieved the highest overall accuracy of 0.87 in classifying all days of the test data set into various event categories. However, it does not have an overwhelming advantage; the category of no drought, contributing 79.11% of all 55,700 calendar days, noticeably influences model accuracy. Even the baseline Rec model performs well in such scenarios, achieving an F1 score of 0.84 without extreme soil moisture values. The most evident disparity among these models appears in the classification of flash drought events.

Notably, the baseline model struggles to identify the onset period of flash droughts. The three metrics for the Rec model hover around 0.06, indicating that nearly all (93%) flash drought onset days were incorrectly identified and that almost all (96%) predicted flash drought onset days were inaccurate. This inefficient and high-risk prediction

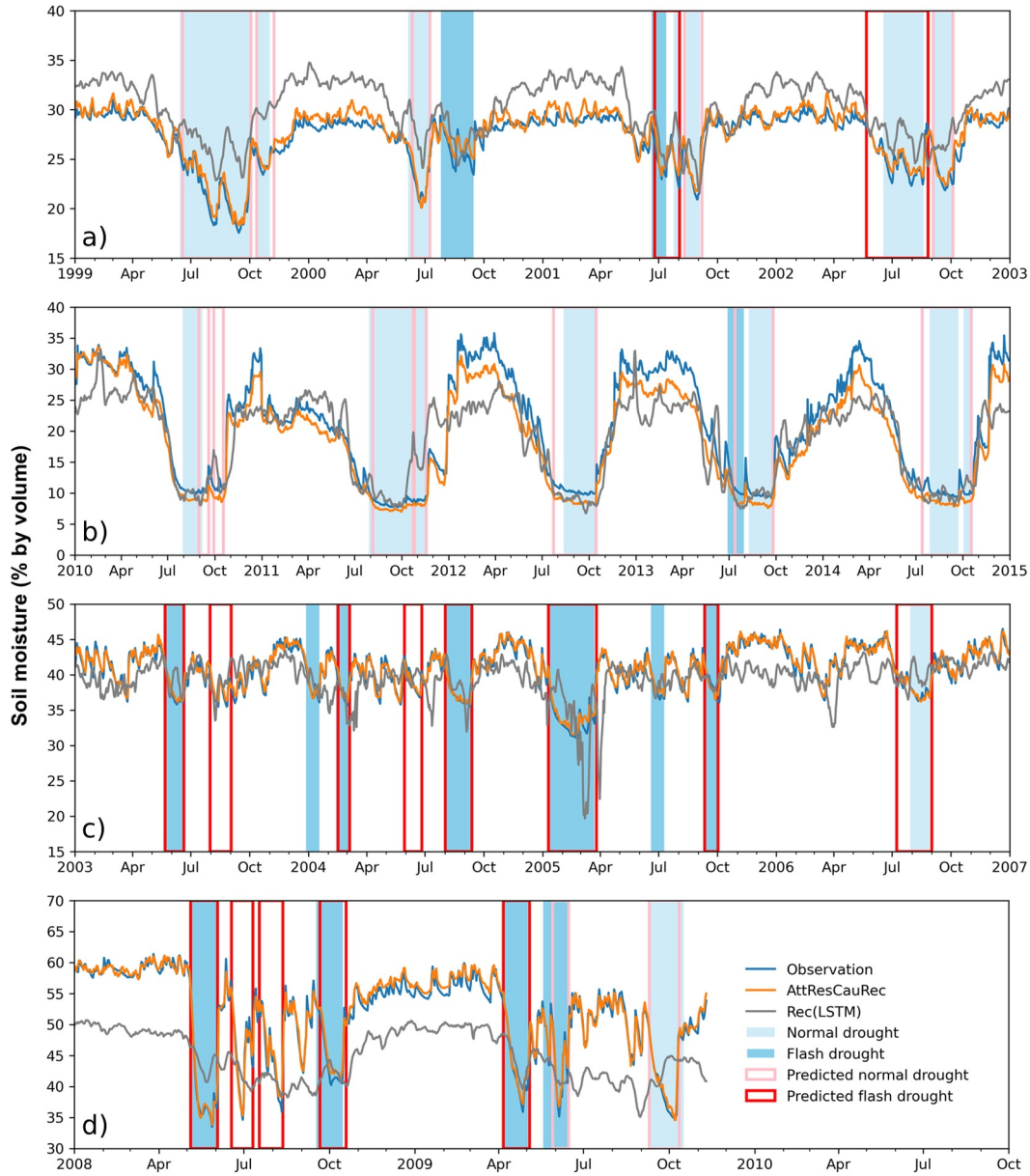

Figure 4. Comparison of flash drought predictions between the proposed model and the baseline model. (a, b) Typical soil moisture trends with an annual periodicity observed at monitoring sites BE-Vie (Vielsalm, Belgium) and US-Me2 (Metolius, USA). (c, d) Long-term soil moisture trends without a periodic pattern at monitoring sites MY-PSO (Pasoh, Malaysia) and CH-Cha (Chamau, Switzerland).

is not useful for flash drought event alarms. In contrast, the ResAttCauRec model exhibits relatively good classification performance for flash drought events, successfully identifying approximately 39% of flash drought onset days in a timely manner, with a positive prediction accuracy of 44% for this category. The detailed classification performance of the proposed model at each site is presented in Table S3 of Supporting Information S1.

ResAttRec, ResRec, and the benchmark U-Net model show similar classification performance. Although these models differ in their ability to predict continuous soil moisture values, their regression performance is not linearly related to classification accuracy. This discrepancy arises because regression and classification capture different aspects of model behavior and are evaluated using distinct metrics (Chicco & Jurman, 2020). To further assess the robustness of classification performance, we increased and decreased the soil moisture thresholds for drought starting and ending points. The results, summarized in Table S4 of Supporting Information S1, reveal

Table 1
Model Performance Based on the Daily Event Classification Result for the Test Data Set

		Categories				
Model	Metrics	Normal drought onset	Normal drought condition	Flash drought onset	Flash drought condition	No drought
ResAttCauRec model	Precision	0.60	0.47	**0.44**	**0.76**	0.96
	Recall	0.72	0.64	**0.39**	**0.60**	0.96
	F1-score	0.66	0.54	**0.41**	**0.67**	0.96
	Accuracy			0.87		
ResAttRec model	Precision	0.28	0.27	**0.05**	**0.32**	0.82
	Recall	0.19	0.21	**0.05**	**0.18**	0.92
	F1-score	0.23	0.24	**0.05**	**0.23**	0.86
	Accuracy			0.72		
ResRec model	Precision	0.31	0.30	**0.07**	**0.29**	0.80
	Recall	0.17	0.20	**0.08**	**0.19**	0.89
	F1-score	0.22	0.24	**0.08**	**0.23**	0.84
	Accuracy			0.71		
U-Net model	Precision	0.36	0.31	**0.07**	**0.27**	0.81
	Recall	0.11	0.12	**0.08**	**0.27**	0.89
	F1-score	0.16	0.17	**0.07**	**0.27**	0.85
	Accuracy			0.71		
Rec model (LSTM)	Precision	0.24	0.18	**0.06**	**0.31**	0.83
	Recall	0.20	0.20	**0.07**	**0.25**	0.86
	F1-score	0.22	0.19	**0.06**	**0.28**	0.84
	Accuracy			0.69		
Proportion		6.46%	5.24%	**1.85%**	**12.66%**	73.80%

Note. The days are categorized into five groups based on whether they are identified as part of a drought event and the phase of drought event they belong to. Bold values indicate the categories of interest. Cell background colors follow a green–yellow–red scale, where green represents higher values, yellow represents intermediate values, and red represents lower values.

performance gaps between the proposed ResAttCauRec model and the others that are consistent with those observed in Table 1.

Additionally, the lead time of prediction greatly influences model performance. Table S5 in Supporting Information S1 presents the classification performance of the proposed model across varying lead times, ranging from 1 day to 2 weeks. The model achieved a substantial improvement (0.16–0.17) in F1-score for both drought onset and condition classification when the lead time was as short as one day. However, performance declined sharply by up to 0.42 when the lead time was extended to 2 weeks. These findings highlight the model's sensitivity to prediction horizon, which is consistent with previous studies reporting similar trends (Bommer et al., 2025). While a one-day lead time yields the highest performance, it holds limited value in operational contexts due to insufficient reaction time. Conversely, although a 2-week lead time offers more time for decision-making, its low predictive accuracy makes it impractical for reliable use. Therefore, a 1-week lead time was chosen as the default setting in this study, as it strikes a reasonable balance between forecast accuracy and actionable foresight.

The choice of input sequence length represents a fundamental trade-off between capturing sufficient temporal dependencies and managing computational cost and model complexity, particularly given the large number of sites and variables considered in this study. As shown in Table S6 of Supporting Information S1, our sensitivity analysis reveals that doubling the input sequence length yields modest improvements in classification performance across all categories, consistent with previous findings that longer temporal windows can provide beneficial contextual information (Levy et al., 2024). Conversely, reducing the input length by half leads to a more pronounced decline in model performance, although it does offer proportional reductions in computational resource requirements for the convolutional modules. It is important to note, however, that excessively long input sequences may introduce irrelevant information and increase the risk of overfitting, especially when distant

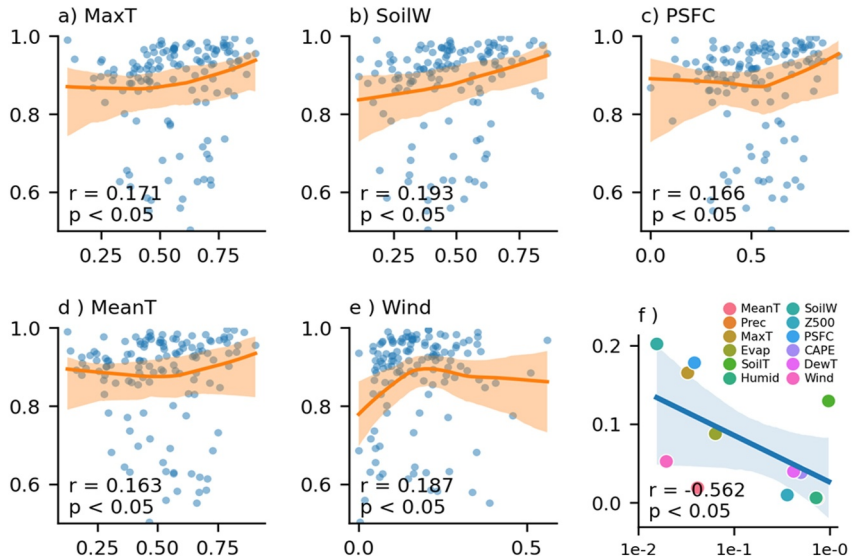

Figure 5. Causality enhances model performance. Panels (a–e) depict the relationship between causality degree (*x*-axis) and model performance (NSE, *y*-axis) for five main drivers, using the locally weighted scatterplot smoothing (LOWESS) of the points and 95% confidence interval from 200 bootstraps; metrics report the Sperman rank-order correlation and corresponding *p*-values. Panels (f) shows the correlation strength (*x*-axis for p-values) related to the feature importance (*y*-axis) of drivers. See Section 3.2 for details on abbreviations of drivers.

historical signals have limited relevance to short-term drought onset (Chakraborty et al., 2022). Although our primary objective is not to achieve state-of-the-art results but to demonstrate the effectiveness of the proposed approach in enhancing backbone models. Future work could address these challenges for better model performance by investigating adaptive input window strategies or attention-based architectures that dynamically assess the importance of different time steps (Iqbal et al., 2020).

The distribution of model predictions, as illustrated in Figure S6 of Supporting Information S1 provides insights into classification performance. It is evident that the prediction distribution for the first percentile group (drought condition days with soil moisture lower than the 20th percentile) varies greatly among the four models. Notably, only the ResAttCauRec model avoids producing unreasonable outlier predictions within this group; the other four models exhibit numerous such outliers, particularly in the test data set, which was not overfitted. Additionally, the highest soil water content group demonstrates the largest variance and interquartile range, highlighting a common disadvantage among the models when addressing contrasting types of extreme events.

4.3. Improvement of Predictive Performance Through Causality

To investigate how causality aids in pattern identification along the feature dimension, we calculated and compared the causality degree based on CCM (Figure S7 in Supporting Information S1) and the importance metric derived from expected gradients (Figure S8 in Supporting Information S1) for all input features. No apparent overlap between the two distributions was observed. However, the relationship between the feature-mean causality degree and model performance is noteworthy, despite its statistical insignificance, as depicted in Figure S9 in Supporting Information S1. Specifically, we found that the causality degree significantly enhances model performance for the top three important drivers identified by feature importance (Figure 5). In contrast, no such relationship exists for less important drivers (Figure S10 in Supporting Information S1). We propose that the coupling of causality and model performance can be amplified by feature importance. In other words, the causality degree represents the bonding strength between features and soil moisture, determining whether soil moisture changes in response to driving features. The extent of these changes is related to feature importance, which can be conceptually regarded as the model's sensitivity to input. We utilized the p-values, as shown in Figures 5a–5e and Figure S8 in Supporting Information S1, to quantify this coupling strength, thereby substantiating our hypothesis in Figure 5f.

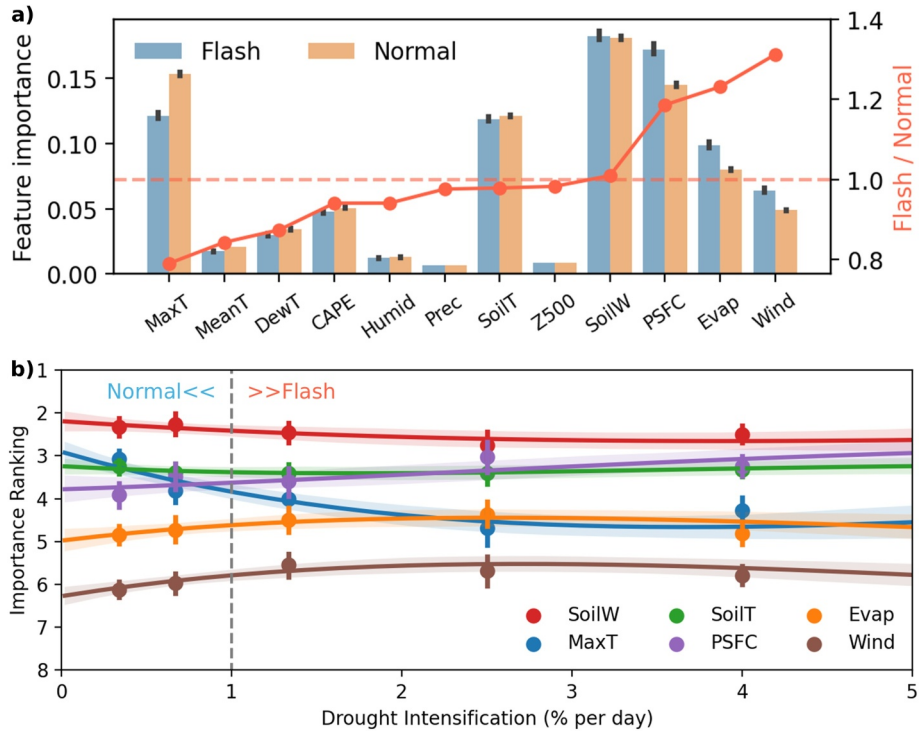

Figure 6. (a) Feature importance of different drivers for normal and flash drought onset. Dots above the orange dashed line indicate drivers that have a greater impact on flash drought onset, while those below indicate drivers more influential for normal drought onset. The error bars represent a 95% confidence interval based on all events. (b) Variation in feature importance ranking with drought intensification, including the top six important drivers. All events are grouped into five evenly sized bins, with a polynomial regression line (order = 4) plotted along with a 95% confidence interval. See Section 3.2 for details on abbreviations of drivers.

The amplification effect of the coupling strength not only supports our rationale for introducing the causality module but also elucidates why additional causal information enhances model performance. Deep learning models are generally deterministic when used as predictors. Incorporating causal information introduces an extra dimension, akin to confidence intervals, which enhances the models' reliability and robustness in the face of input uncertainty.

4.4. Mechanisms Behind Flash Drought Onset

Given the model's capabilities in fitting soil moisture and identifying drought events, we further leveraged the feature importance derived from the EG-based XAI method to explore the mechanisms underlying flash and normal drought onset. Specifically, we extracted the feature importance scores for each variable during the onset period of each drought event. Figure 6a presents the average importance scores for both drought types, aggregated across all historical events and sites. Figure 6b illustrates the distribution of feature importance in relation to the drought intensification rate, offering a more detailed view of variable influence under different drought dynamics.

Despite the considerable uncertainty and bias in ERA5 soil moisture predictions (Kim et al., 2020; S. Yang, Li, et al., 2020; Zheng et al., 2024), soil moisture continues to play a crucial role in forecasting future soil water content. Additionally, soil temperature, another key indicator of soil status, has long been recognized as having a strong correlation with soil moisture. This relationship has been consistently validated in the literature (Lakshmi et al., 2003; Riveros-Iregui et al., 2007) and effectively applied in soil moisture modeling (Dong et al., 2016).

Surface pressure and wind speed were identified as the second and sixth most important drivers for both flash and normal droughts. Variations in atmospheric pressure and surface wind are recognized as key mechanisms responsible for air movement into and out of soils (Buckingham, 1904), a phenomenon often referred to as “pressure pumping.” This mechanism greatly influences the exchange fluxes of water between the soil and

atmosphere (Fukuda, 1955; Louge et al., 2022; Yu et al., 2020) and affects the rate of evaporation (the fifth most important driver in this study) from soil water (Ishihara et al., 1992; Zeng et al., 2011).

Additionally, daily maximum temperature plays an important role in drought onset. Numerous studies on compound drought-heatwave events have demonstrated that heatwaves characterized by high daily maximum temperature contribute to the development of drought conditions, and conversely, droughts can exacerbate heatwave occurrences (Bevacqua et al., 2022; C. Wang et al., 2023). Interestingly, our findings indicate that precipitation has a minimal contribution to soil moisture predictions. This is largely due to the rapid infiltration process in the top layer of soil, which means the soil's response to precipitation occurs at a sub-daily scale (Eltahir, 1998). As a result, the impact of individual precipitation events on long-term soil water content is limited (Rahmati et al., 2024).

The orange ratio line in Figure 6a illustrates the differences in the importance of various drivers for normal versus flash drought onset. In the case of normal droughts, which typically develop over extended periods, temperature plays a critical role in increasing evapotranspiration rates (Alfieri et al., 2020), leading to a gradual decline in soil moisture levels (Masson-Delmotte et al., 2021). Elevated daily maximum temperatures can accelerate moisture loss from both soil and vegetation over time, thereby contributing to the onset and persistence of drought conditions (Rahmati et al., 2024).

The contributions of evaporation and surface pressure to flash drought onset are approximately 20% higher than their contributions to normal droughts, while the contribution of wind is 30% higher. This increased influence can be attributed to the "pressure pumping" mechanism, where the interplay between pressure and wind facilitates enhanced air movement between the soil, vegetation, and atmosphere. This dynamic process substantially accelerates evapotranspiration rates, resulting in a faster moisture loss. The rapid evaporation dries the top layer of soil, leading to the swift onset of flash droughts, which often occur within hours to days, and persist as long as the wind conditions remain favorable (Rahmati et al., 2024). Additionally, variations in surface pressure can impact weather patterns and large-scale atmospheric circulation, such as the formation of strong quasi-stationary ridges characterized by positive geopotential height anomalies and associated high surface pressure (Ford & Labosier, 2017). These persistent blocking conditions frequently lead to extended periods of clear skies, which increase incoming solar radiation, elevate temperatures, and reduce precipitation (Ford & Labosier, 2017; Hoerling et al., 2014; Mo & Lettenmaier, 2016). All of these factors can rapidly deplete soil moisture.

Generally, normal droughts are primarily influenced by long-term temperature trends, whereas flash droughts are more responsive to rapid atmospheric changes. In the latter case, surface pressure, wind speed, and evaporation are crucial factors that contribute to the swift depletion of soil moisture. This distinction is also evident in the absolute contribution rankings of features. Despite a stable annual ranking (Figure S11 in Supporting Information S1), the rank lines for maximum temperature and surface pressure intersect at the boundary between the two types of droughts (Figure 6b and Figure S12 in Supporting Information S1). Notably, the importance of surface pressure continues to rise with the intensification rate, while maximum temperature becomes less important in the context of more rapidly occurring flash droughts.

4.5. Acceleration of Drought Onset Due To Increasing Complexity

Existing studies have highlighted the accelerating trend of onset for both flash droughts (Qing et al., 2022) and normal drought (X. Yuan et al., 2023) from global reanalysis and assimilation grids. This is consistent with the temporal trend shown in Figure 7a, which is derived from a monitoring site network. Over the 17 valid year cycles, the average intensification rate during the onset period has increased by 0.95 percentile/day. Concurrently, the global average complexity of drought—defined as the number of primary drivers during the drought onset period (refer to Section 3.5 for details)—has also shown a significant upward trend. Especially after 2004, changes in complexity have synchronized with those in the intensification rate. Previous mechanistic analyses in Section 4.4 suggest that complex interactions among the identified drivers may compound the impacts of drought, thereby accelerating the onset process. Given the widespread occurrence of this synchronization phenomenon across global sites (Figures 7c and 7d), it underscores the importance of understanding flash drought risks from a multivariate perspective (Brunner et al., 2021).

To evaluate the influence of high drought complexity on the reliability of flash drought predictions, we assessed the average estimation errors that may occur when predicting soil moisture during the drought onset period. The

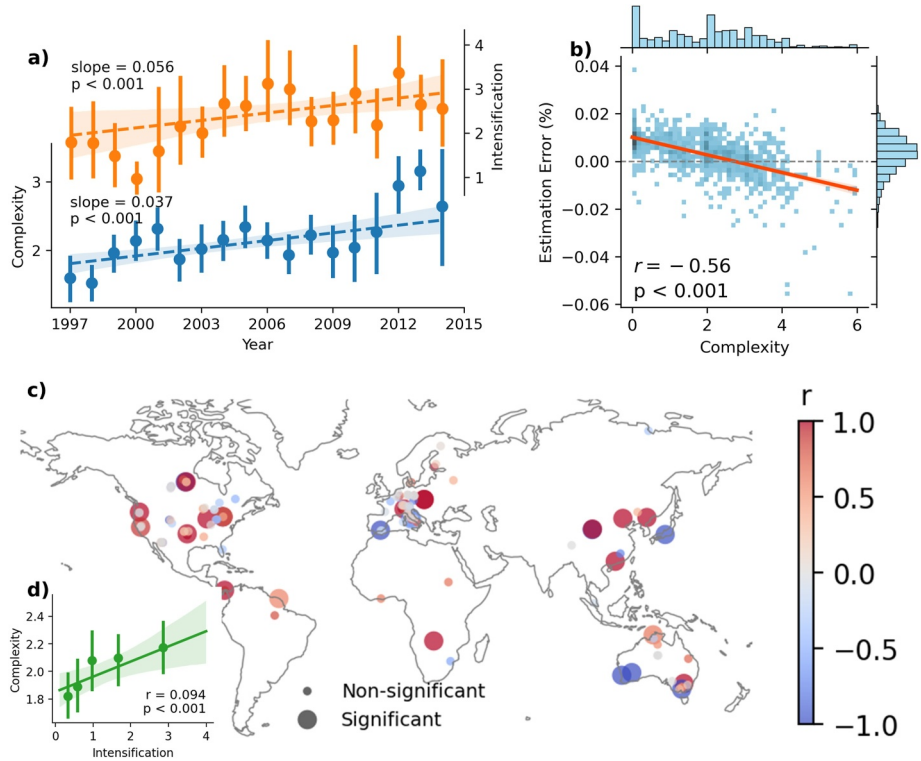

Figure 7. (a) The intensification rate and complexity of drought onset have increased during the period from 1997 to 2014. (b) Distribution of event-scale estimation errors in relation to complexity, along with corresponding linear regression analysis. Pixels above the gray line ($x = 0$) indicate overestimation of soil moisture, while those below indicate underestimation. (c) Synchronization between the intensification rate and complexity is observed in most (73%) of the global monitoring network. (d) Relationship based on event-scale analysis.

analysis reveals a tendency to overestimate soil moisture in drought events characterized by higher complexity, while an underestimation is observed in events with lower complexity (Figure 7b). The strong negative correlation between drought complexity and estimation error ($r = -0.56$, $P < 0.001$) indicates potential risks for future drought early warning and management, particularly in the context of increasing drought complexity and intensification rates. Importantly, this correlation is unlikely to stem from a decrease in the sample size of highly complex droughts, as similar results were obtained using percentile groups of complexity (Figure S13 in Supporting Information S1). The high variance and underestimation in the high complexity groups further underscore the challenges in predicting and issuing warnings for accelerating multi-driver flash droughts.

4.6. Discussion

Historically, large-scale flash droughts have resulted in billions of dollars in agricultural losses and damaged ecosystems. Notable flash drought events include those across much of the USA in 2012, the northern USA in 2017 (Hoell et al., 2020; Otkin et al., 2018), and in southern China in 2011 and 2013 (X. Yuan et al., 2019). As flash droughts are fast-developing and hard to forecast, predictive models are expected to overreact to potential rapid soil moisture depletion. In practice, false alarms may be preferable to missed events (Hembach-Stunden et al., 2024), given the disproportionate socioeconomic cost of underestimating such extremes. Therefore, future flash drought models may be designed with a higher tolerance for false positives (i.e., higher recall), especially in operational settings. Nevertheless, to better learn the underlying dynamics, models should initially be developed for continuous regression tasks, rather than classification, as the former retains richer information and improves representation learning.

Despite the promising performance of our ResAttCauRec model, several limitations remain. First, from a data perspective, the quality and quantity of observational soil moisture data constrain model generalization. The current data set may not fully capture the diversity of land cover, soil types, vegetation, and hydroclimatic

regimes, which are critical to reflecting the full range of soil moisture dynamics (Cosby et al., 1984; Fang et al., 2016; Joshi et al., 2011). In particular, some ecosystems prone to flash droughts remain underrepresented (D'Odorico et al., 2007). While spatial Monte Carlo cross-validation provides a more realistic assessment of the model's spatial generalizability, the test set in each split may still fail to capture the full diversity of environmental conditions, such as extreme climate zones, unique soil textures, or land-use types. As a result, the estimated generalization ability might be optimistic for entirely unrepresented regions. Moreover, some locations in the test set might still share similar climatological or ecological characteristics with training sites, which may unintentionally favor the model's performance. This data sparsity also limits the model's capacity to accurately characterize and learn the complex land–atmosphere interactions underlying the observed soil moisture dynamics which is critical for subsequent attribution analysis. To address this, future work should prioritize the inclusion of more geographically and ecologically diverse observations. Additionally, while more satellite/reanalysis products (e.g., SMAP, MERRA) and land surface models (e.g., VIC, CLM) can be used to generate synthetic data for training (Plésiat et al., 2024; Trok et al., 2024), careful bias correction and downscaling are required to bridge the gap between synthetic and real-world data. Increased computational resources and potentially more complex model structures are also needed for such large-scale data sets.

In this study, the model architecture was designed based on empirical tuning, balancing performance with computational feasibility. Due to limited computational resources and domain-specific constraints, exhaustive hyperparameter optimization (e.g., via Bayesian optimization or Neural Architecture Search) was not feasible (Frazier, 2018; White et al., 2023). The current model may not be the optimal design for all drought types. As deep learning tools continue to mature, more sophisticated models—potentially larger and better tailored—could further improve flash drought predictability, especially when paired with fine-tuning and transfer learning.

Another frontier lies in weather forecasting. Flash drought development is highly sensitive to short-term meteorological variability, such as high evapotranspiration and precipitation deficits. The integration of state-of-the-art weather forecasting models, such as Pangu-Weather (Bi et al., 2023) or GenCast (Price et al., 2025), could provide more accurate environmental drivers and extend the lead time of drought early warning systems. Designing a seamless coupling between high-resolution weather forecasts and hydrological deep learning models remains a valuable future direction.

While our study introduces causal information to guide feature learning, we acknowledge that this does not directly reveal whether the model's internal representations align with known physical processes. Lees et al. (2021) provide a promising direction for interpreting LSTM states by probing their internal memory vectors to uncover hydrologically meaningful representations, such as variable-capacity soil moisture stores and snow cover. Future research could extend the framework by applying similar probing techniques to assess whether the causality-informed LSTM states encode interpretable physical signals related to drought onset mechanisms.

The results emphasize the importance of integrating causality into deep learning models. By embedding causal information directly into the architecture, the model strategically weights different features, guiding the subsequent LSTM layers to prioritize the most important ones. This enhancement is crucial for accurately modeling the sequential dynamics of soil moisture fluctuations, improving the model's reliability and robustness in the face of input uncertainty. It also enhances the model's understanding of temporal dependencies and driver interactions, leading to better interpretability of its predictions (Lehmann et al., 2020). As CCM method focuses solely on temporal causality, a hybrid causal discovery method that accounts for both spatial and temporal causality may further strengthen model performance (B. Gao et al., 2023).

5. Conclusions

To implement effective flash drought risk management and adaptation strategies, this study proposes a novel deep learning framework, the ResAttCauRec model, for improving soul moisture flash drought predictions. This framework employs a CNN-LSTM backbone to capture the dependence of soil moisture on spatial-temporal meteorological variables. An attention mechanism is integrated into the CNN module to enable the model to assign varying weights to different spatial locations. Additionally, a novel CCM-based causality module is incorporated into the conventional LSTM, providing strategic weighting of the feature dimension by initiating and controlling cell states.

The results of the ablation study reveal that (a) the proposed causality module, along with the attention module, serves as a regularization technique that improves generalization (GA = 0.83) and performance (NSE = 0.84 on the test set) compared to the baseline model. (b) The ResAttCauRec model excels in capturing soil moisture extremes, enabling effective forecasts of flash drought events, achieving an F1 score of 0.41 for flash drought onset, in contrast to 0.06 for the baseline model and 0.07 for the benchmark model. (c) Further XAI-based model interpretation shows that the causality degree significantly enhances performance for the top important drivers, supporting the introduction of the causality module. (d) In addition to soil temperature and soil water content, daily maximum temperature, evaporation, and surface pressure are identified as the most important drivers of soil moisture dynamics. (e) Normal droughts are primarily influenced by long-term temperature trends, while flash droughts are more responsive to rapid atmospheric changes, with surface pressure, wind speed, and evaporation playing crucial roles in the swift depletion of soil moisture. (f) The analysis also reveals a concerning trend of increasing drought complexity and intensification, which poses challenges for reliable prediction and early warning systems.

This work highlights the necessity for adaptive management strategies that take into account the complex nature of flash drought events. As the landscape of flash drought risk evolves, our study provides valuable insights into the mechanisms driving flash drought onset, advocating for improved predictive models that can better inform agricultural and ecological practices while mitigating the impacts of these events. Additionally, this study introduces an effective approach to enhance data-driven deep learning models by incorporating additional causal information, which not only facilitates forecast and interpretation of flash droughts but may also be extended to broader extreme weather events.

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The soil moisture from the FLUXNET2015 data set (Pastorello et al., 2020) are publicly available at https://fluxnet.org/; the ERA5 reanalysis data (Hersbach et al., 2020) are obtained from European Centre for Medium-Range Weather Forecasts (ECMWF) at https://cds.climate.copernicus.eu/.

All code in this study is in Python. The core code for model building and training can be found in a GitHub repository at https://github.com/nantekoto/ResAttCauRec/. The calculations for the CCM causality were performed using the causal-ccm package (Javier, 2021). Model realizations are based on PyTorch package (Paszke et al., 2019); the XAI-based model interpretation is performed using the SHAP package (Lundberg, 2017); all visualizations are completed by Matplotlib (Hunter, 2007) and Seaborn (Waskom, 2021) packages. All packages can be found via Python Package Index, the official repository of software for the Python programming language, at https://pypi.org/.

Acknowledgments
We sincerely thank the Editor and anonymous reviewers for their constructive comments, which significantly improved the quality of this work. This research was supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU/RGC 15232023) and the Hong Kong Polytechnic University (Project No. P0043040, P0045957). The computational resources in this study were supported by the Center for Computational Science and Engineering at Southern University of Science and Technology. Additional support was provided by the High-level University Special Fund (Grant G030290001).

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AghaKouchak, A. (2014). A baseline probabilistic drought forecasting framework using standardized soil moisture index: Application to the 2012 United States drought. *Hydrology and Earth System Sciences*, *18*(7), 2485–2492. https://doi.org/10.5194/hess-18-2485-2014

Alfieri, J., Kustas, W., & Anderson, M. (2020). A brief overview of approaches for measuring evapotranspiration. *Agroclimatology: Linking Agriculture to Climate*, *60*, 109–127. https://doi.org/10.2134/agronmonogr60.2016.0034

Allen, C. D., Macalady, A. K., Chenchouni, H., Bachelet, D., McDowell, N., Vennetier, M., et al. (2010). A global overview of drought and heat-induced tree mortality reveals emerging climate change risks for forests. *Forest Ecology and Management*, *259*(4), 660–684. https://doi.org/10.1016/j.foreco.2009.09.001

Anderson, M. C., Hain, C., Otkin, J., Zhan, X., Mo, K., Svoboda, M., et al. (2013). An intercomparison of drought indicators based on thermal remote sensing and NLDAS-2 simulations with US drought monitor classifications. *Journal of Hydrometeorology*, *14*(4), 1035–1056. https://doi.org/10.1175/jhm-d-12-0140.1

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., & Sharma, U. (2024). Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, *121*(27), e2311878121. https://doi.org/10.1073/pnas.2311878121

Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

Bakke, S. J., Ionita, M., & Tallaksen, L. M. (2023). Recent European drying and its link to prevailing large-scale atmospheric patterns. *Scientific Reports*, *13*(1), 21921. https://doi.org/10.1038/s41598-023-48861-4

Balles, L., Romero, J., & Hennig, P. (2016). Coupling adaptive batch sizes with learning rates. *arXiv preprint arXiv:1612.05086*.

Barkiđija, S., & Fuchs, Ž. (2013). Precipitation correlation between convective available potential energy, convective inhibition and saturation fraction in middle latitudes. *Atmospheric Research*, *124*, 170–180. https://doi.org/10.1016/j.atmosres.2012.12.010
Bartlett, P. L., Long, P. M., Lugosi, G., & Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, *117*(48), 30063–30070. https://doi.org/10.1073/pnas.1907378117
Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade* (2nd ed., pp. 437–478). Springer.
Bevacqua, E., Zappa, G., Lehner, F., & Zscheischler, J. (2022). Precipitation trends determine future occurrences of compound hot–dry events. *Nature Climate Change*, *12*(4), 350–355. https://doi.org/10.1038/s41558-022-01309-5
Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, *619*(7970), 533–538. https://doi.org/10.1038/s41586-023-06185-3
Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning* (Vol. 4). Springer.
Bommer, P. L., Kretschmer, M., Spuler, F. R., Bykov, K., & Höhne, M. M.-C. (2025). Deep learning meets teleconnections: Improving S2S predictions for European winter weather. *arXiv preprint arXiv:2504.07625*, *1*(1), 015002. https://doi.org/10.1088/3049-4753/ade9c2
Brubaker, K. L., & Entekhabi, D. (1996). Analysis of feedback mechanisms in land-atmosphere interaction. *Water Resources Research*, *32*(5), 1343–1357. https://doi.org/10.1029/96wr00005
Brunner, M. I., Slater, L., Tallaksen, L. M., & Clark, M. (2021). Challenges in modeling and predicting floods and droughts: A review. *Wiley Interdisciplinary Reviews: Water*, *8*(3), e1520. https://doi.org/10.1002/wat2.1520
Buckingham, E. (1904). *Contributions to our knowledge of the aeration of soils*. Department of Agriculture, Bureau of Soils.
Cai, H., Liu, S., Shi, H., Zhou, Z., Jiang, S., & Babovic, V. (2022). Toward improved lumped groundwater level predictions at catchment scale: Mutual integration of water balance mechanism and deep learning method. *Journal of Hydrology*, *613*, 128495. https://doi.org/10.1016/j.jhydrol.2022.128495
Cai, Y., Zheng, W., Zhang, X., Zhangzhong, L., & Xue, X. (2019). Research on soil moisture prediction model based on deep learning. *PLoS One*, *14*(4), e0214508. https://doi.org/10.1371/journal.pone.0214508
Chakraborty, M., Li, W., Faber, J., Rümpker, G., Stoecker, H., & Srivastava, N. (2022). A study on the effect of input data length on a deep-learning-based magnitude classifier. *Solid Earth*, *13*(11), 1721–1729. https://doi.org/10.5194/se-13-1721-2022
Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., & Miao, Y. (2021). Review of image classification algorithms based on convolutional neural networks. *Remote Sensing*, *13*(22), 4712. https://doi.org/10.3390/rs13224712
Chen, L. G., Gottschalck, J., Hartman, A., Miskus, D., Tinker, R., & Artusa, A. (2019). Flash drought characteristics based on US drought monitor. *Atmosphere*, *10*(9), 498. https://doi.org/10.3390/atmos10090498
Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*, 1–13. https://doi.org/10.1186/s12864-019-6413-7
Cosby, B., Hornberger, G., Clapp, R., & Ginn, T. (1984). A statistical exploration of the relationships of soil moisture characteristics to the physical properties of soils. *Water Resources Research*, *20*(6), 682–690. https://doi.org/10.1029/wr020i006p00682
Crausbay, S. D., Ramirez, A. R., Carter, S. L., Cross, M. S., Hall, K. R., Bathke, D. J., et al. (2017). Defining ecological drought for the twenty-first century. *Bulletin of the American Meteorological Society*, *98*(12), 2543–2550. https://doi.org/10.1175/bams-d-16-0292.1
Dai, H., Xiong, L., Ma, Q., & Duan, Z. (2025). Deep learning model for drought prediction based on large-scale spatial causal network in the Yangtze River Basin. *Journal of Hydrology*, *654*, 132808. https://doi.org/10.1016/j.jhydrol.2025.132808
Datta, P., & Faroughi, S. A. (2023). A multihead LSTM technique for prognostic prediction of soil moisture. *Geoderma*, *433*, 116452. https://doi.org/10.1016/j.geoderma.2023.116452
Davarzani, H., Smits, K., Tolene, R. M., & Illangasekare, T. (2014). Study of the effect of wind speed on evaporation from soil through integrated modeling of the atmospheric boundary layer and shallow subsurface. *Water Resources Research*, *50*(1), 661–680. https://doi.org/10.1002/2013wr013952
Devlin, J. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
Dikshit, A., Pradhan, B., & Alamri, A. M. (2021). Long lead time drought forecasting using lagged climate variables and a stacked long short-term memory model. *Science of the Total Environment*, *755*, 142638. https://doi.org/10.1016/j.scitotenv.2020.142638
D'Odorico, P., Caylor, K., Okin, G. S., & Scanlon, T. M. (2007). On soil moisture–vegetation feedbacks and their possible effects on the dynamics of dryland ecosystems. *Journal of Geophysical Research*, *112*(G4). https://doi.org/10.1029/2006jg000379
Dong, J., Steele-Dunne, S. C., Ochsner, T. E., & Van De Giesen, N. (2016). Determining soil moisture and soil properties in vegetated areas by assimilating soil temperatures. *Water Resources Research*, *52*(6), 4280–4300. https://doi.org/10.1002/2015wr018425
Du, J., Kimball, J., Jencso, K., Hoylman, Z., Brust, C., Ketchum, D., et al. (2024). Machine-learning based multi-layer soil moisture forecasts—An application case study of the Montana 2017 flash drought. *Water Resources Research*, *60*(10), e2023WR036973. https://doi.org/10.1029/2023wr036973
Eltahir, E. A. (1998). A soil moisture–rainfall feedback mechanism: 1. Theory and observations. *Water Resources Research*, *34*(4), 765–776. https://doi.org/10.1029/97wr03499
Erion, G., Janizek, J. D., Sturmfels, P., Lundberg, S. M., & Lee, S.-I. (2021). Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, *3*(7), 620–631. https://doi.org/10.1038/s42256-021-00343-w
Fang, X., Zhao, W., Wang, L., Feng, Q., Ding, J., Liu, Y., & Zhang, X. (2016). Variations of deep soil moisture under different vegetation types and influencing factors in a watershed of the Loess Plateau, China. *Hydrology and Earth System Sciences*, *20*(8), 3309–3323. https://doi.org/10.5194/hess-20-3309-2016
Faranda, D., Pascale, S., & Bulut, B. (2023). Persistent anticyclonic conditions and climate change exacerbated the exceptional 2022 European-mediterranean drought. *Environmental Research Letters*. https://doi.org/10.1088/1748-9326/acbc37
Feng, J., Li, J., Xu, C. Y., Wang, Z., Zhang, Z., Wu, X., et al. (2024). Viewing soil moisture flash drought onset mechanism and their changes through XAI lens: A case study in eastern China. *Water Resources Research*, *60*(6), e2023WR036297. https://doi.org/10.1029/2023wr036297
Ford, T. W., & Labosier, C. F. (2017). Meteorological conditions associated with the onset of flash drought in the eastern United States. *Agricultural and Forest Meteorology*, *247*, 414–423. https://doi.org/10.1016/j.agrformet.2017.08.031
Foroumandi, E., Gavahi, K., & Moradkhani, H. (2024). Generative adversarial network for real-time flash drought monitoring: A deep learning study. *Water Resources Research*, *60*(5), e2023WR035600. https://doi.org/10.1029/2023wr035600
Frazier, P. I. (2018). A tutorial on Bayesian optimization. *arXiv preprint arXiv:1807.02811*.
Fukuda, H. (1955). Air and vapor movement in soil due to wind gustiness. *Soil Science*, *79*(4), 249–256. https://doi.org/10.1097/00010694-195504000-00002
Gao, B., Yang, J., Chen, Z., Sugihara, G., Li, M., Stein, A., et al. (2023). Causal inference from cross-sectional Earth system data with geographical convergent cross mapping. *Nature Communications*, *14*(1), 5875. https://doi.org/10.1038/s41467-023-41619-6

Gao, P., Qiu, H., Lan, Y., Wang, W., Chen, W., Han, X., & Lu, J. (2021). Modeling for the prediction of soil moisture in litchi orchard with deep long short-term memory. *Agriculture*, *12*(1), 25. https://doi.org/10.3390/agriculture12010025

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1). MIT Press Cambridge.

Gorgij, A. D., Askari, G., Taghipour, A., Jami, M., & Mirfardi, M. (2023). Spatiotemporal forecasting of the groundwater quality for irrigation purposes, using deep learning method: Long short-term memory (LSTM). *Agricultural Water Management*, *277*, 108088. https://doi.org/10.1016/j.agwat.2022.108088

Grus, J. (2019). *Data science from scratch: First principles with python*. O'Reilly Media.

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Han, Z., Zhao, J., Leung, H., Ma, K. F., & Wang, W. (2019). A review of deep learning models for time series prediction. *IEEE Sensors Journal*, *21*(6), 7833–7848. https://doi.org/10.1109/jsen.2019.2923982

He, F., Liu, T., & Tao, D. (2019). Control batch size and learning rate to generalize well: Theoretical and empirical evidence. *Advances in Neural Information Processing Systems*, *32*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Hembach-Stunden, K., Vorlaufer, T., & Engel, S. (2024). False and missed alarms in seasonal forecasts affect individual adaptation choices. *Q Open*, *4*(1), qoad031. https://doi.org/10.1093/qopen/qoad031

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., et al. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Ho, S., Tian, L., Disse, M., & Tuo, Y. (2021). A new approach to quantify propagation time from meteorological to hydrological drought. *Journal of Hydrology*, *603*, 127056. https://doi.org/10.1016/j.jhydrol.2021.127056

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G. S., et al. (2021). Mc-lstm: Mass-conserving lstm. Paper presented at the International conference on machine learning.

Hoell, A., Parker, B.-A., Downey, M., Umphlett, N., Jencso, K., Akyuz, F. A., et al. (2020). Lessons learned from the 2017 flash drought across the US Northern Great Plains and Canadian Prairies. *Bulletin of the American Meteorological Society*, *101*(12), E2171–E2185. https://doi.org/10.1175/bams-d-19-0272.1

Hoerling, M., Eischeid, J., Kumar, A., Leung, R., Mariotti, A., Mo, K., et al. (2014). Causes and predictability of the 2012 Great Plains drought. *Bulletin of the American Meteorological Society*, *95*(2), 269–282. https://doi.org/10.1175/bams-d-13-00055.1

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/mcse.2007.55

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, *14*(12), 124007. https://doi.org/10.1088/1748-9326/ab4e55

Hwang, J. (2020). Modeling financial time series using LSTM with Trainable Initial Hidden States. *arXiv preprint arXiv:2007.06848*.

Indu, J., Nair, A. S., Pradhan, A., Mangla, R., Krishnan, S., Verma, K., & Huggannavar, V. (2022). Terrestrial water budget through radar remote sensing. In *Radar remote sensing* (pp. 123–148). Elsevier.

Iqbal, W., Berral, J. L., & Carrera, D. (2020). Adaptive sliding windows for improved estimation of data center resource utilization. *Future Generation Computer Systems*, *104*, 212–224. https://doi.org/10.1016/j.future.2019.10.026

Ishihara, Y., Shimojima, E., & Harada, H. (1992). Water vapor transfer beneath bare soil where evaporation is influenced by a turbulent surface wind. *Journal of Hydrology*, *131*(1–4), 63–104. https://doi.org/10.1016/0022-1694(92)90213-f

Javier, P. (2021). Causal-CCM a python implementation of convergent cross mapping. *Causal-CCM a Python Implementation of Convergent Cross Mapping*.

Jiang, S., Tarasova, L., Yu, G., & Zscheischler, J. (2024). Compounding effects in flood drivers challenge estimates of extreme river floods. *Science Advances*, *10*(13), eadl4005. https://doi.org/10.1126/sciadv.adl4005

Jiang, S., Zheng, Y., & Solomatine, D. (2020). Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning. *Geophysical Research Letters*, *47*(13), e2020GL088229. https://doi.org/10.1029/2020gl088229

Jing, Y., Wang, S., Chan, P.-W., & Yang, Z.-L. (2025). Gross primary productivity is more sensitive to accelerated flash droughts. *Communications Earth & Environment*, *6*(1), 34. https://doi.org/10.1038/s43247-025-02013-w

Joshi, C., Mohanty, B. P., Jacobs, J. M., & Ines, A. V. (2011). Spatiotemporal analyses of soil moisture from point to footprint scale in two different hydroclimatic regions. *Water Resources Research*, *47*(1). https://doi.org/10.1029/2009wr009002

Kennel, M. B., Brown, R., & Abarbanel, H. D. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A*, *45*(6), 3403–3411. https://doi.org/10.1103/physreva.45.3403

Kim, H., Wigneron, J.-P., Kumar, S., Dong, J., Wagner, W., Cosh, M. H., et al. (2020). Global scale error assessments of soil moisture estimates from microwave-based active and passive satellites and land surface models over forest and mixed irrigated/dryland agriculture regions. *Remote Sensing of Environment*, *251*, 112052. https://doi.org/10.1016/j.rse.2020.112052

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Konapala, G., Kao, S.-C., Painter, S. L., & Lu, D. (2020). Machine learning assisted hybrid models can improve streamflow simulation in diverse catchments across the conterminous US. *Environmental Research Letters*, *15*(10), 104022. https://doi.org/10.1088/1748-9326/aba927

Kornelsen, K. C., & Coulibaly, P. (2014). Root-zone soil moisture estimation using data-driven methods. *Water Resources Research*, *50*(4), 2946–2962. https://doi.org/10.1002/2013wr014127

Kumar, S., & Tian, D. (2024). Causal discovery analysis reveals global sources of predictability for regional flash droughts. *Water Resources Research*, *60*(11), e2024WR038391. https://doi.org/10.1029/2024wr038391

Lakshmi, V., Jackson, T. J., & Zehrfuhs, D. (2003). Soil moisture–temperature relationships: Results from two field experiments. *Hydrological Processes*, *17*(15), 3041–3057. https://doi.org/10.1002/hyp.1275

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., et al. (2021). Hydrological concept formation inside long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences Discussions*, *2021*, 1–37.

Lehmann, J., Kretschmer, M., Schauberger, B., & Wechsung, F. (2020). Potential for early forecast of Moroccan wheat yields based on climatic drivers. *Geophysical Research Letters*, *47*(12), e2020GL087516. https://doi.org/10.1029/2020gl087516

Levy, M., Jacoby, A., & Goldberg, Y. (2024). Same task, more tokens: The impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*, 15339–15353. https://doi.org/10.18653/v1/2024.acl-long.818

Li, L., Dai, Y., Shangguan, W., Wei, Z., Wei, N., & Li, Q. (2022). Causality-structured deep learning for soil moisture predictions. *Journal of Hydrometeorology*, *23*(8), 1315–1331. https://doi.org/10.1175/jhm-d-21-0206.1

Li, P., Jia, L., Lu, J., Jiang, M., & Zheng, C. (2024). A new evapotranspiration-based drought index for flash drought identification and monitoring. *Remote Sensing*, *16*(5), 780. https://doi.org/10.3390/rs16050780

Li, W., Migliavacca, M., Forkel, M., Denissen, J. M., Reichstein, M., Yang, H., et al. (2022). Widespread increasing vegetation sensitivity to soil moisture. *Nature Communications*, *13*(1), 3959. https://doi.org/10.1038/s41467-022-31667-9

Li, X., Zhang, Y., Zhang, J., Chen, S., Marsic, I., Farneth, R. A., & Burd, R. S. (2017). Concurrent activity recognition with multimodal CNN-LSTM structure. *arXiv preprint arXiv:1702.01638*.

Liang, M., & Yuan, X. (2021). Critical role of soil moisture memory in predicting the 2012 Central United States flash drought. *Frontiers in Earth Science*, *9*, 615969. https://doi.org/10.3389/feart.2021.615969

Loehrer, S. M., & Johnson, R. H. (1995). Surface pressure and precipitation life cycle characteristics of PRE-STORM mesoscale convective systems. *Monthly Weather Review*, *123*(3), 600–621. https://doi.org/10.1175/1520-0493(1995)123<0600:spaplc>2.0.co;2

Lorenz, E. N., & Haman, K. (1996). The essence of chaos. *Pure and Applied Geophysics*, *147*(3), 598–599.

Louge, M., Valance, A., Xu, J., Ould el-Moctar, A., & Chasle, P. (2022). Water vapor transport across an arid sand Surface—Non-Linear thermal coupling, wind-driven pore advection, subsurface waves, and exchange with the atmospheric boundary layer. *Journal of Geophysical Research: Earth Surface*, *127*(4), e2021JF006490. https://doi.org/10.1029/2021jf006490

Lundberg, S. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Luo, C., Zheng, X., & Zeng, D. (2014). Causal inference in social media using convergent cross mapping. Paper presented at the 2014 IEEE Joint Intelligence and Security Informatics Conference.

Mahto, S. S., & Mishra, V. (2023). Increasing risk of simultaneous occurrence of flash drought in major global croplands. *Environmental Research Letters*, *18*(4), 044044. https://doi.org/10.1088/1748-9326/acc8ed

Mahto, S. S., & Mishra, V. (2024). Global evidence of rapid flash drought recovery by extreme precipitation. *Environmental Research Letters*, *19*(4), 044031. https://doi.org/10.1088/1748-9326/ad300c

Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S. L., Péan, C., Berger, S., et al. (2021). Climate change 2021: The physical science basis. *Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*, *2*(1), 2391.

McVicar, T. R., Roderick, M. L., Donohue, R. J., Li, L. T., Van Niel, T. G., Thomas, A., et al. (2012). Global review and synthesis of trends in observed terrestrial near-surface wind speeds: Implications for evaporation. *Journal of Hydrology*, *416*, 182–205. https://doi.org/10.1016/j.jhydrol.2011.10.024

Menditto, A., Patriarca, M., & Magnusson, B. (2007). Understanding the meaning of accuracy, trueness and precision. *Accreditation and Quality Assurance*, *12*(1), 45–47. https://doi.org/10.1007/s00769-006-0191-z

Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, *10*, 99129–99149. https://doi.org/10.1109/access.2022.3207287

Mishra, A. K., & Singh, V. P. (2010). A review of drought concepts. *Journal of Hydrology*, *391*(1–2), 202–216. https://doi.org/10.1016/j.jhydrol.2010.07.012

Mishra, V., Aadhar, S., & Mahto, S. S. (2021). Anthropogenic warming and intraseasonal summer monsoon variability amplify the risk of future flash droughts in India. *Npj Climate and Atmospheric Science*, *4*(1), 1. https://doi.org/10.1038/s41612-020-00158-3

Mo, K. C., & Lettenmaier, D. P. (2015). Heat wave flash droughts in decline. *Geophysical Research Letters*, *42*(8), 2823–2829. https://doi.org/10.1002/2015gl064018

Mo, K. C., & Lettenmaier, D. P. (2016). Precipitation deficit flash droughts over the United States. *Journal of Hydrometeorology*, *17*(4), 1169–1184. https://doi.org/10.1175/jhm-d-15-0158.1

Mohammadi Foumani, N., Miller, L., Tan, C. W., Webb, G. I., Forestier, G., & Salehi, M. (2024). Deep learning for time series classification and extrinsic regression: A current survey. *ACM Computing Surveys*, *56*(9), 1–45. https://doi.org/10.1145/3649448

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Mousavi, S. M., Ellsworth, W. L., Zhu, W., Chuang, L. Y., & Beroza, G. C. (2020). Earthquake transformer—An attentive deep-learning model for simultaneous earthquake detection and phase picking. *Nature Communications*, *11*(1), 3952. https://doi.org/10.1038/s41467-020-17591-w

Myoung, B., & Nielsen-Gammon, J. W. (2010). The convective instability pathway to warm season drought in Texas. Part I: The role of convective inhibition and its modulation by soil moisture. *Journal of Climate*, *23*(17), 4461–4473. https://doi.org/10.1175/2010jcli2946.1

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Nguyen, H., Wheeler, M. C., Hendon, H. H., Lim, E.-P., & Otkin, J. A. (2021). The 2019 flash droughts in subtropical eastern Australia and their association with large-scale climate drivers. *Weather and Climate Extremes*, *32*, 100321. https://doi.org/10.1016/j.wace.2021.100321

Osman, M., Zaitchik, B. F., Badr, H. S., Christian, J. I., Tadesse, T., Otkin, J. A., & Anderson, M. C. (2020). Flash drought onset over the contiguous United States: Sensitivity of inventories and trends to quantitative definitions. *Hydrology and Earth System Sciences Discussions*, *2020*, 1–21.

Otkin, J. A., Anderson, M. C., Hain, C., & Svoboda, M. (2015). Using temporal changes in drought indices to generate probabilistic drought intensification forecasts. *Journal of Hydrometeorology*, *16*(1), 88–105. https://doi.org/10.1175/jhm-d-14-0064.1

Otkin, J. A., Svoboda, M., Hunt, E. D., Ford, T. W., Anderson, M. C., Hain, C., & Basara, J. B. (2018). Flash droughts: A review and assessment of the challenges imposed by rapid-onset droughts in the United States. *Bulletin of the American Meteorological Society*, *99*(5), 911–919. https://doi.org/10.1175/bams-d-17-0149.1

Otkin, J. A., Woloszyn, M., Wang, H., Svoboda, M., Skumanich, M., Pulwarty, R., et al. (2022). Getting ahead of flash drought: From early warning to early action. *Bulletin of the American Meteorological Society*, *103*(10), E2188–E2202. https://doi.org/10.1175/bams-d-21-0288.1

Packard, N. H., Crutchfield, J. P., Farmer, J. D., & Shaw, R. S. (1980). Geometry from a time series. *Physical Review Letters*, *45*(9), 712–716. https://doi.org/10.1103/physrevlett.45.712

Park, S., Im, J., Jang, E., & Rhee, J. (2016). Drought assessment and monitoring through blending of multi-sensor indices using machine learning approaches for different climate regions. *Agricultural and Forest Meteorology*, *216*, 157–169. https://doi.org/10.1016/j.agrformet.2015.10.011

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., et al. (2020). The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Scientific Data*, *7*(1), 225. https://doi.org/10.1038/s41597-020-0534-3

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, *32*.

Pitis, S. (2016). Non-zero initial states for recurrent neural networks. Retrieved from https://r2rt.com/non-zero-initial-states-for-recurrent-neural-networks.html

Plésiat, É., Dunn, R. J., Donat, M. G., & Kadow, C. (2024). Artificial intelligence reveals past climate extremes by reconstructing historical records. *Nature Communications*, *15*(1), 9191. https://doi.org/10.1038/s41467-024-53464-2

Powers, D. M. (2020). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Prechelt, L. (2002). Early stopping-but when? In *Neural networks: Tricks of the trade* (pp. 55–69). Springer.

Price, I., Sanchez-Gonzalez, A., Alet, F., Andersson, T. R., El-Kadi, A., Masters, D., et al. (2025). Probabilistic weather forecasting with machine learning. *Nature*, *637*(8044), 84–90. https://doi.org/10.1038/s41586-024-08252-9

Prodhan, F. A., Zhang, J., Yao, F., Shi, L., Pangali Sharma, T. P., Zhang, D., et al. (2021). Deep learning for monitoring agricultural drought in South Asia using remote sensing data. *Remote Sensing*, *13*(9), 1715. https://doi.org/10.3390/rs13091715

Qing, Y., & Wang, S. (2025). Soil drying intensification increases the connection between dry and hot extremes in a changing climate. *Earth's Future*, *13*(5), e2024EF005151. https://doi.org/10.1029/2024ef005151

Qing, Y., Wang, S., Ancell, B. C., & Yang, Z.-L. (2022). Accelerating flash droughts induced by the joint influence of soil moisture depletion and atmospheric aridity. *Nature Communications*, *13*(1), 1139. https://doi.org/10.1038/s41467-022-28752-4

Rahmati, M., Amelung, W., Brogi, C., Dari, J., Flammini, A., Bogena, H., et al. (2024). Soil moisture memory: State-of-the-art and the way forward. *Reviews of Geophysics*, *62*(2), e2023RG000828. https://doi.org/10.1029/2023rg000828

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat, F. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Riveros-Iregui, D. A., Emanuel, R. E., Muth, D. J., McGlynn, B. L., Epstein, H. E., Welsch, D. L., et al. (2007). Diurnal hysteresis between soil CO2 and soil temperature is controlled by soil water content. *Geophysical Research Letters*, *34*(17). https://doi.org/10.1029/2007gl030938

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Paper presented at the Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18* (pp. 234–241). https://doi.org/10.1007/978-3-319-24574-4_28

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., et al. (2019). Inferring causation from time series in Earth system sciences. *Nature Communications*, *10*(1), 2553. https://doi.org/10.1038/s41467-019-10105-3

Sak, H., Senior, A., & Beaufays, F. (2014). Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *arXiv preprint arXiv:1402.1128*.

Senay, G. B., Budde, M., Brown, J., & Verdin, J. (2008). *Mapping flash drought in the US: Southern Great Plains*. Paper presented at the 22nd conference on hydrology. AMS.

Shen, B.-W., Pielke Sr, R. A., Zeng, X., Baik, J.-J., Faghih-Naini, S., Cui, J., & Atlas, R. (2021). Is weather chaotic? Coexistence of chaos and order within a generalized Lorenz model. *Bulletin of the American Meteorological Society*, *102*(1), E148–E158. https://doi.org/10.1175/bams-d-19-0165.1

Smith, A. (2018). NOAA national centers for environmental information (NCEI). US billion-dollar weather and. *Climate Disasters*.

Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., & Munch, S. (2012). Detecting causality in complex ecosystems. *Science*, *338*(6106), 496–500. https://doi.org/10.1126/science.1227079

Sugihara, G., & May, R. M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, *344*(6268), 734–741. https://doi.org/10.1038/344734a0

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. Paper presented at the International conference on machine learning.

Svoboda, M., LeComte, D., Hayes, M., Heim, R., Gleason, K., Angel, J., et al. (2002). The drought monitor. *Bulletin of the American Meteorological Society*, *83*(8), 1181–1190. https://doi.org/10.1175/1520-0477-83.8.1181

Takens, F. (2006). Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980: Proceedings of a symposium held at the University of Warwick 1979/80* (pp. 366–381). Springer.

Tian, Y., Zhao, Y., Li, J., Chen, B., Deng, L., & Wen, D. (2024). East Asia atmospheric river forecast with a deep learning method: GAN-UNet. *Journal of Geophysical Research: Atmospheres*, *129*(5), e2023JD039311. https://doi.org/10.1029/2023jd039311

Tong, X., Zhou, W., & Xia, J. (2024). Improving Boreal summer precipitation predictions from the global NMME through Res34-Unet. *Geophysical Research Letters*, *51*(2), e2023GL106391. https://doi.org/10.1029/2023gl106391

Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)* (pp. 4489–4497). https://doi.org/10.1109/iccv.2015.510

Trok, J. T., Barnes, E. A., Davenport, F. V., & Diffenbaugh, N. S. (2024). Machine learning–based extreme event attribution. *Science Advances*, *10*(34), eadl3242. https://doi.org/10.1126/sciadv.adl3242

Tufaner, F., & Özbeyaz, A. (2020). Estimation and easy calculation of the Palmer Drought Severity Index from the meteorological data by using the advanced machine learning algorithms. *Environmental Monitoring and Assessment*, *192*(9), 1–14. https://doi.org/10.1007/s10661-020-08539-0

Tyagi, S., Zhang, X., Saraswat, D., Sahany, S., Mishra, S. K., & Niyogi, D. (2022). Flash drought: Review of concept, prediction and the potential for machine learning, deep learning methods. *Earth's Future*, *10*(11), e2022EF002723. https://doi.org/10.1029/2022ef002723

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Venkatesan, R., & Li, B. (2017). *Convolutional neural networks in visual computing: A concise guide*. CRC Press.

Wadoux, A. M.-C., Heuvelink, G. B., De Bruin, S., & Brus, D. J. (2021). Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, *457*, 109692. https://doi.org/10.1016/j.ecolmodel.2021.109692

Wang, C., Li, Z., Chen, Y., Ouyang, L., Li, Y., Sun, F., et al. (2023). Drought-heatwave compound events are stronger in drylands. *Weather and Climate Extremes*, *42*, 100632. https://doi.org/10.1016/j.wace.2023.100632

Wang, L., & Yuan, X. (2018). Two types of flash drought and their connections with seasonal drought. *Advances in Atmospheric Sciences*, *35*(12), 1478–1490. https://doi.org/10.1007/s00376-018-8047-0

Wang, Y., Yang, J., Chen, Y., De Maeyer, P., Li, Z., & Duan, W. (2018). Detecting the causal effect of soil moisture on precipitation using convergent cross mapping. *Scientific Reports*, *8*(1), 12171. https://doi.org/10.1038/s41598-018-30669-2

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Wenke, S., & Fleming, J. (2019). Contextual recurrent neural networks. *arXiv preprint arXiv:1902.03455*.

White, C., Safari, M., Sukthanker, R., Ru, B., Elsken, T., Zela, A., et al. (2023). Neural architecture search: Insights from 1000 papers. *arXiv preprint arXiv:2301.08727*.

Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. Paper presented at the Proceedings of the European conference on computer vision (ECCV). *Lecture Notes in Computer Science*, 3–19. https://doi.org/10.1007/978-3-030-01234-2_1

Xiao, C., Chen, N., Hu, C., Wang, K., Gong, J., & Chen, Z. (2019). Short and mid-term sea surface temperature prediction using time-series satellite data and LSTM-AdaBoost combination approach. *Remote Sensing of Environment*, *233*, 111358. https://doi.org/10.1016/j.rse.2019.111358

Xiao, X., Ming, W., Luo, X., Yang, L., Li, M., Yang, P., et al. (2024). Leveraging multisource data for accurate agricultural drought monitoring: A hybrid deep learning model. *Agricultural Water Management*, *293*, 108692. https://doi.org/10.1016/j.agwat.2024.108692

Xu, Q.-S., & Liang, Y.-Z. (2001). Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, *56*(1), 1–11. https://doi.org/10.1016/s0169-7439(00)00122-2

Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, *2*(3), 249–262. https://doi.org/10.1007/s41664-018-0068-2

Yang, R., Singh, S. K., Tavakkoli, M., Amiri, N., Yang, Y., Karami, M. A., & Rai, R. (2020). CNN-LSTM deep learning architecture for computer vision-based modal frequency detection. *Mechanical Systems and Signal Processing*, *144*, 106885. https://doi.org/10.1016/j.ymssp.2020.106885

Yang, S., Li, R., Wu, T., Hu, G., Xiao, Y., Du, Y., et al. (2020). Evaluation of reanalysis soil temperature and soil moisture products in permafrost regions on the Qinghai-Tibetan Plateau. *Geoderma*, *377*, 114583. https://doi.org/10.1016/j.geoderma.2020.114583

Yevjevich, V. M. (1967). *An objective approach to definitions and investigations of continental hydrologic droughts* (Vol. 23). Colorado State University.

Yu, L., Zeng, Y., & Su, Z. (2020). Understanding the mass, momentum, and energy transfer in the frozen soil with three levels of model complexities. *Hydrology and Earth System Sciences*, *24*(10), 4813–4830. https://doi.org/10.5194/hess-24-4813-2020

Yuan, G., Zhang, L., & Liu, Y. (2021). Impacts of soil moisture and atmospheric moisture transport on the precipitation in two typical regions of China. *Atmospheric Research*, *247*, 105151. https://doi.org/10.1016/j.atmosres.2020.105151

Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., et al. (2020). Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sensing of Environment*, *241*, 111716. https://doi.org/10.1016/j.rse.2020.111716

Yuan, X., Wang, L., Wu, P., Ji, P., Sheffield, J., & Zhang, M. (2019). Anthropogenic shift towards higher risk of flash drought over China. *Nature Communications*, *10*(1), 4661. https://doi.org/10.1038/s41467-019-12692-7

Yuan, X., Wang, Y., Ji, P., Wu, P., Sheffield, J., & Otkin, J. A. (2023). A global transition to flash droughts under climate change. *Science*, *380*(6641), 187–191. https://doi.org/10.1126/science.abn6301

Zeng, Y., Su, Z., Wan, L., & Wen, J. (2011). A simulation analysis of the advective effect on evaporation using a two-phase heat and mass flow model. *Water Resources Research*, *47*(10). https://doi.org/10.1029/2011wr010701

Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2021). Dive into deep learning. *arXiv preprint arXiv:2106.11342*.

Zhang, J.-L., Huang, X.-M., & Sun, Y.-Z. (2024). Multiscale spatiotemporal meteorological drought prediction: A deep learning approach. *Advances in Climate Change Research*, *15*(2), 211–221. https://doi.org/10.1016/j.accre.2024.04.003

Zhang, M., Yuan, X., Otkin, J. A., & Ji, P. (2022). Climate warming outweighs vegetation greening in intensifying flash droughts over China. *Environmental Research Letters*, *17*(5), 054041. https://doi.org/10.1088/1748-9326/ac69fb

Zhao, W. L., Gentine, P., Reichstein, M., Zhang, Y., Zhou, S., Wen, Y., et al. (2019). Physics-constrained machine learning of evapotranspiration. *Geophysical Research Letters*, *46*(24), 14496–14507. https://doi.org/10.1029/2019gl085291

Zheng, Y., Coxon, G., Woods, R., Power, D., Rico-Ramirez, M. A., McJannet, D., et al. (2024). Evaluation of reanalysis soil moisture products using cosmic ray neutron sensor observations across the globe. *Hydrology and Earth System Sciences*, *28*(9), 1999–2022. https://doi.org/10.5194/hess-28-1999-2024

Zhu, Q., Luo, Y., Zhou, D., Xu, Y.-P., Wang, G., & Tian, Y. (2021). Drought prediction using in situ and remote sensing products with SVM over the Xiang River Basin, China. *Natural Hazards*, *105*(2), 2161–2185. https://doi.org/10.1007/s11069-020-04394-x