International Journal of Applied Earth Observation and Geoinformation 140 (2025) 104563

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

journal homepage: www.elsevier.com/locate/jag

Calibration and uncertainty quantification for deep learning-based drought detection

Mengxue Zhang [a,*], Miguel-Ángel Fernández-Torres [b], Kai-Hendrik Cohrs [a], Gustau Camps-Valls [a]

[a] Image Processing Laboratory (IPL), Universitat de València, Paterna, València, 46980, Spain
[b] Department of Signal Theory and Communications, Universidad Carlos III de Madrid (UC3M), Madrid, 28911, Spain

ARTICLE INFO

Keywords:
Extreme climate events
Drought detection
Deep learning
Model calibration
Uncertainty quantification

ABSTRACT

Droughts are hydrometeorological extreme events influenced by highly intricate land–atmosphere feedback mechanisms and climate variability. Deep learning models have recently succeeded in detecting extreme climate events and promise to uncover and understand droughts further. There are two main challenges of reliability and trustworthiness limiting their applications: miscalibration and inherent uncertainty. However, they remain rarely explored because deep learning models are overparameterized and seldom tractable. To address this shortcoming, we introduce methodologies for model calibration and entropy-based uncertainty quantification for deep learning-based drought detection. The calibration algorithm can deal with calibration errors by reducing distributional shifts and alleviating overconfident predictions. The uncertainty framework, in turn, decomposes and quantifies the total uncertainty according to several components: data uncertainty, procedural variability, parametric variability, and latent variability. Thus, our method identifies uncertain predictions and supports robust evaluations, benefiting the credibility of the decision-making process. Empirical evidence of performance in a wide range of European drought events is given, justifying the effectiveness of our approach. The calibration methodology yields the lowest expected calibration error (0.31%) and the precision of the uncertainty-based decision-making is improved from 72.27% to 74.06% and 76.59%, based on ensemble predictions and rejecting the predictions for the top 20% uncertain negative samples, respectively. In summary, our approach significantly enhances drought detection's reliability and classification accuracy, constituting a key step toward more trustworthy and actionable climate decision-making.

1. Introduction

Droughts are hydrometeorological phenomena driven by complex climate and weather cycles and processes (Van Loon, 2015). Recent studies reveal droughts are relevant to large-scale atmospheric circulation (Trenberth et al., 2014), land–atmosphere feedbacks (Schumacher et al., 2022), and climate variability (Jiang and Zhou, 2023). However, the dynamics involved in droughts are highly non-linear, and a thorough understanding of the key drivers is still lacking (Miralles et al., 2019; Koppa et al., 2024). Meanwhile, droughts are among the most harmful natural disasters, causing severe damage to the ecological environment, agricultural production, and socioeconomic impacts (Mishra and Singh, 2010). The effect of droughts under an aggravated global warming scenario is likely to be intensified in the future (Trenberth et al., 2014), thereby effective detection and understanding of drought dynamics and the development of management interventions are becoming more necessary (Haile et al., 2020).

Index-based approaches are mechanistic tools for drought characterization. Among the most frequently used approaches, we can highlight the Palmer Drought Severity Index (PDSI) (Palmer, 1965), the Standardized Precipitation Index (SPI) (McKee et al., 1993), and the Standardized Precipitation–Evapotranspiration Index (SPEI) (Vicente-Serrano et al., 2010). PDSI measures the wetness and dryness based on the supply and demand concept of water balance. SPI incorporates multi-scalar characteristics of drought, enabling the monitoring and managing of different water resources. SPEI combines the water balance and multi-scalar character to present drought quantification. Besides, indices based on combined indicators (Sepulcre-Canto et al., 2012), empirical soil moisture (Carrão et al., 2016), and multi-source remote sensing data are also used for drought monitoring (Shen et al., 2019; Zhang et al., 2022; Zhao et al., 2023). They enrich the characterization of multivariate droughts. However, these indices may carry over some errors and uncertainties from the independent variables and the

* Corresponding author.
E-mail address: mengxue@alumni.uv.es (M. Zhang).

https://doi.org/10.1016/j.jag.2025.104563
Received 8 October 2024; Received in revised form 24 March 2025; Accepted 23 April 2025
Available online 9 May 2025

computational process (Hoffmann et al., 2020). They can be sensitive to the length and quality of historical data and, by relying on the Penman–Monteith equation, might overestimate future PET rates (Ault, 2020). Therefore, calibration and uncertainty assessment are required for their application to assess drought conditions. Notable calibration efforts include the self-calibrating PDSI (Wells et al., 2004), a framework for integrating water storage changes into the Global Hydrology Model (Schumacher et al., 2018), and recent work on aerodynamic resistance calibration for surface energy balance (de Andrade et al., 2021).

Regarding uncertainty quantification, the PDSI assessment revealed that the data source variability and the potential evaporation modeling may cause divergences in analyzing global warming (Sheffield et al., 2012). SPI is undermined by the effect of climate change, as assessed in Zargar et al. (2014). The same issue happens with the epistemic uncertainty on SPI and SPEI (Laimighofer and Laaha, 2022). The formulation of Global Climate Models becomes the dominant source of spread by the end of the 21st century, especially for soil moisture drought (Orlowsky and Seneviratne, 2013). Indeed, aleatoric uncertainties are challenging for the reliable drought assessment under climate variability (Mukherjee et al., 2018).

More recently, deep learning-based (DL) models have exhibited huge success in Earth science (Camps-Valls et al., 2021) and are increasingly applied to drought-relevant variables estimation and forecasting. Validated models include, but are not limited to, Long short-term memory (LSTM) networks (Xiao et al., 2019), convolutional-LSTMs (Foroumandi et al., 2023), and transformers (Gonzalez-Calabuig et al., 2024). Further examples of probabilistic DL models applied for Earth science applications are deep auto-regressive networks (Chattopadhyay et al., 2020), generative adversarial networks (Ravuri et al., 2021), variational auto-encoders (Rúvźivćka et al., 2022), and normalizing flows (McDonald et al., 2022).

There are two main challenges of reliability and trustworthiness limiting DL in drought applications: miscalibration and inherent uncertainty. Existing studies have shown DL models may have poor calibration (Guo et al., 2017), i.e., their outputs are overconfident or underconfident and have low operational value, as they cannot be trusted for further decision-making (Faghani et al., 2023). Moreover, DL models have inherent uncertainty, which can cause errors in drought detection, forecasting, characterization, and attribution. Conversely, accounting for uncertainty in water management has led to lower costs and better decisions (McMillan et al., 2017), as well as improved hydrological process understanding by enabling robust reasoning and the identification of spatial and temporal patterns that relate to the true process rather than just uncertainty (McMillan et al., 2018).

However, few studies discuss the calibration and the uncertainty of DL models for extreme climate events because they are complex, overparameterized, and seldom tractable models, limiting their reliability and trustworthiness. A data–model uncertainty coupling framework has been recently proposed to estimate the predictive uncertainty of precipitation forecasting (Xu et al., 2021). Parameterized DL models have also been coupled with Bayesian model averaging for rapid flood risk quantification (Yosri et al., 2024). Currently, the DL drought literature lacks investigations on calibration and uncertainty. As the number of DL models applied for drought detection increases, it is essential to study the calibration and the uncertainty to make models reliable and trustworthy (Materia et al., 2024; Camps-Valls et al., 2025).

To address the above research limitations, we propose methodologies for the calibration and uncertainty quantification of DL-based drought detection models, particularly those trained on historical events. Specifically, the following essential points are concerned: *(1) how to reduce calibration errors; (2) how to measure and quantify the different sources of uncertainty; (3) how to benefit decision-making through the estimated uncertainties.* To this end, a post-processing parametric model calibration algorithm is first introduced to deal with the calibration errors by reducing distributional shifts of both labels prior and likelihood, alleviating the overconfident predictions. Then, an entropy-based uncertainty framework is proposed to estimate four types of inherent uncertainty in DL models, identifying uncertain posteriors and supporting robust evaluations or comparisons. Last but not least, two strategies based on the uncertainty framework are explored to improve the credibility of decision-making.

2. Background

2.1. Data preparation

We exploit ERA-5[1] as the fifth generation European Centre for Medium-Range Weather Forecasts reanalysis for the global climate and weather for the past eight decades. Data is available from 1940 to the present. It provides physically consistent essential climate variables (ECVs) (Mahecha et al., 2020). We collected total precipitation and air temperature covering the period from Jan 2003 to Dec 2018.

The Köppen-Geiger system[2] is then considered as the reference for our study of the spatio-temporal heterogeneity of drought events. This global climate classification includes five main classes and 30 sub-types. The categorization is based on threshold values and seasonality of monthly air temperature and precipitation (Beck et al., 2018). We followed the Köppen-Geiger system to split Europe into different climate-consistent zones (15 sub-types) for calibration and uncertainty analysis.

To train the drought detection models and evaluate their performance, we leverage the International Disaster Database (EM-DAT.[3]) EM-DAT is a global database recording the occurrence and effects of more than 21,000 natural disasters worldwide from 1900 to the present. EM-DAT records human and economic losses at the country level for disasters, fulfilling at least one of the following criteria: 10 fatalities, 100 affected people, a declaration of a state of emergency, and a call for international assistance. Therefore, all recorded disaster events are objective and influential (Guha-Sapir et al., 2015). We utilized the Geocoded Disasters (GDIS) database (Rosvold and Buhaug, 2021), which provides a raster data structure for the historical drought events in EM-DAT.

Harmonizing data and preprocessing is necessary to conduct the different experiments. In this sense, all ECVs, climate-consistent zones, and GDIS were re-gridded to the same spatial resolution ($0.5° \times 0.5°$). Time-series data are further resampled to a fixed temporal resolution of 8-days. We split the data into distinct periods for the training, validation, and test sets to ensure no overlap. In the end, the training data spans from Jan 2003 to Dec 2008, the validation data spans from Jan 2009 to Dec 2010, and the testing data spans from Jan 2011 to Dec 2018. Climate-consistent zones and the drought events on the training and validation subsets are visualized in Fig. 1(a), and the spatial descriptions of all drought events considered in this study are provided in Zhang et al. (2024).

2.2. Deep learning for drought detection

Drought detection involves identifying droughts happening geographically over time to differentiate them from regular conditions (Camps-Valls et al., 2025). DL-based models for drought detection incorporate time series of ECVs as input and yield corresponding estimated drought probabilities per each location and time step. Our previous study[4] introduced two probabilistic generative models for

[1] Copernicus ERA5 overview document.
[2] www.gloh2o.org/koppen.
[3] https://public.emdat.be/.
[4] https://github.com/mengxue-rs/DK-VRN.

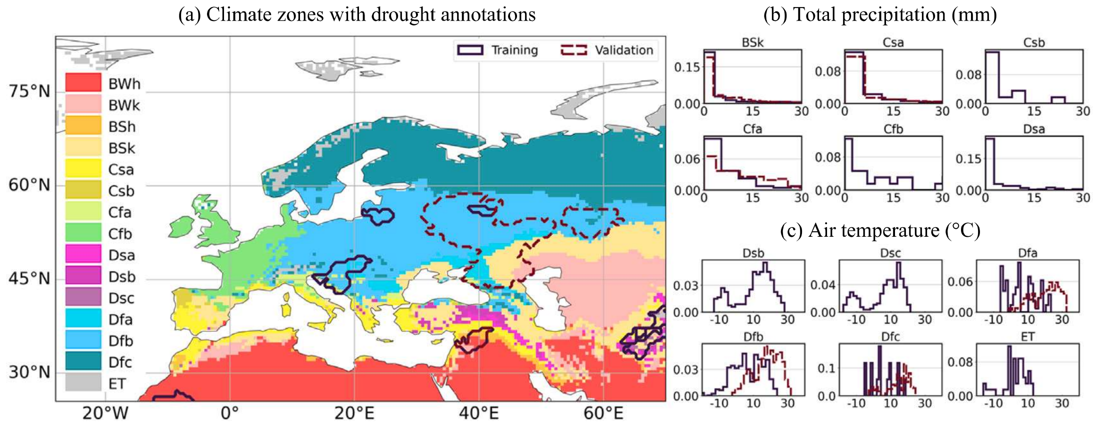

Fig. 1. Visualization of the extended European study region and data. Climate zones and the locations for the drought events covered by the training and validation subsets are highlighted in (a), while the distribution of ECVs (total precipitation, air temperature) for six selected climate zones, considering only those spatio-temporal locations where and when droughts happened, are shown in (b) and (c), respectively. The different climate zones are BWh (hot desert), BWk (cold desert), BSh (hot semi-arid), BSk (cold semi-arid), Csa (hot-summer Mediterranean), Csb (warm-summer Mediterranean), Cfa (humid subtropical), Cfb (subtropical highland), Dsa (hot, dry-summer continental), Dsb (warm, dry-summer continental), Dsc (dry-summer subarctic), Dfa (hot-summer humid continental), Dfb (warm-summer humid continental), Dfc (subarctic), and ET (tundra). The color map used follows the official settings. A more detailed description of the climate zones can be found in Beck et al. (2018). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

drought detection: a variational long short-term memory network (VLSTM) and a domain knowledge-driven variational recurrent network (DK-VRN). They model temporal dependencies between total precipitation and air temperature variables at their input and are trained on the historical drought events in EM-DAT, which are identified based on socio-economic criteria, independent from these ECVs (Zhang et al., 2024). These networks constitute the basis of this study. Their hyperparameters, as well as the data preprocessing techniques used (i.e., the computation of monthly anomalies and the undersampling strategy for the training subset) are kept the same. The architectures of VLSTM and DK-VRN are detailed in Table S.1. As perceived in the following experiments, the outputs of these architectures are uncalibrated probabilities prepared for model calibration and uncertainty quantification.

2.3. Motivation

Here we motivate the necessity of model calibration and uncertainty quantification for DL-based drought detection. Let us formalize the problem of drought detection. Notationally, X is a selection of input variables, i.e., ECVs, $Y \in \mathbb{C} = \{0, 1\}$ is the corresponding drought event label (non-drought and drought, respectively) per time step, and p denotes the uncalibrated drought probabilities (posterior) provided at the output of the DL models above.

First, a Bayesian perspective can help us understand the distributional shifts and justify the need for model calibration:

$$p \approx P_t(Y|X = x) = \frac{P_t(X = x|Y)P_t(Y)}{P_t(X = x)}, \tag{1}$$

where $P_t(Y|X = x)$ is the true posterior, $P_t(X = x|Y)$ is the true likelihood, $P_t(Y)$ is the true label prior distribution, and $P_t(X = x)$ is the true data prior distribution. However, the estimated posterior from DL models is not necessarily consistent with the true posterior, as distributional shifts are common in the environment. For instance, the non-stationary variability of ECVs under droughts in different scenarios ($P_t(X = x|Y)$), and the definitive ambiguities or spatial–temporal heterogeneity of drought events ($P_t(Y)$), to name a few, are possible reasons for the distributional shifts. Accordingly, the reliability of the estimated posterior is limited by the distributional shifts without effective model calibration.

The distributions of drought events are imbalanced and spatiotemporally heterogeneous, as is exemplified in Fig. 1(a) by the existing distributional shift of the label prior between the training and the validation subsets. Besides, the distributional shift of the likelihood between six different climate zones can be observed in Fig. 1(b)–(c). These distributional shifts can result in overconfident posterior estimates, as is shown in the results derived from our experiments.

Second, several sources of inherent uncertainty associated with different components of the DL drought detection model can be identified: (1) the variability of the ECVs, (2) the training procedures, (3) the neural network parameters, and (4) the latent variables. These sources influence the estimated posterior of DL models and consequently result in uncertainty, which can cause errors in decisions. Analogous conclusions are discussed for different DL models in Gawlikowski et al. (2023). Accordingly, it is necessary to quantify uncertainty to ensure trustworthy decision-making.

3. Method

3.1. Model calibration

Calibration of DL models intends to maximize the agreement between the estimated probabilities and true frequencies. Chen and Su (2023) discovered it is difficult for calibration algorithms to achieve reasonable results for imbalanced training data, while (Dong et al., 2024) provided a comprehensive survey about the calibration under class imbalance data.

As motivated in Section 2.3, uncalibrated probabilities can be affected by two distributional shifts: label prior shift and likelihood shift. Dal Pozzolo et al. (2015) discovered the label prior shift caused by undersampling affecting the probabilistic calibration of machine learning models. Tian et al. (2020) investigated the integration of prior re-balancing and likelihood flattening. Guilbert et al. (2024) explored exponential and polynomial families of calibration functions. However, existing parametric post-processing calibration methods theoretically rely on distributional assumptions or require a high-quality calibration set. For drought detection, collecting a consistent calibration subset may be difficult because historical records of drought events are imbalanced and scarce, as the event occurrences are spatiotemporally heterogeneous.

To tackle these issues, we propose a post-processing parametric calibration algorithm, which is depicted in Fig. 2. Specifically, the algorithm includes two steps designed to tackle the prior shift and likelihood shift, respectively. First, we divide the region we analyze into N climatic zones following Beck et al. (2018). The uncalibrated probabilities of the DL models trained on the extended European region are then projected by new climatic calibration functions learned from their corresponding climate zone locations in the training subset. The

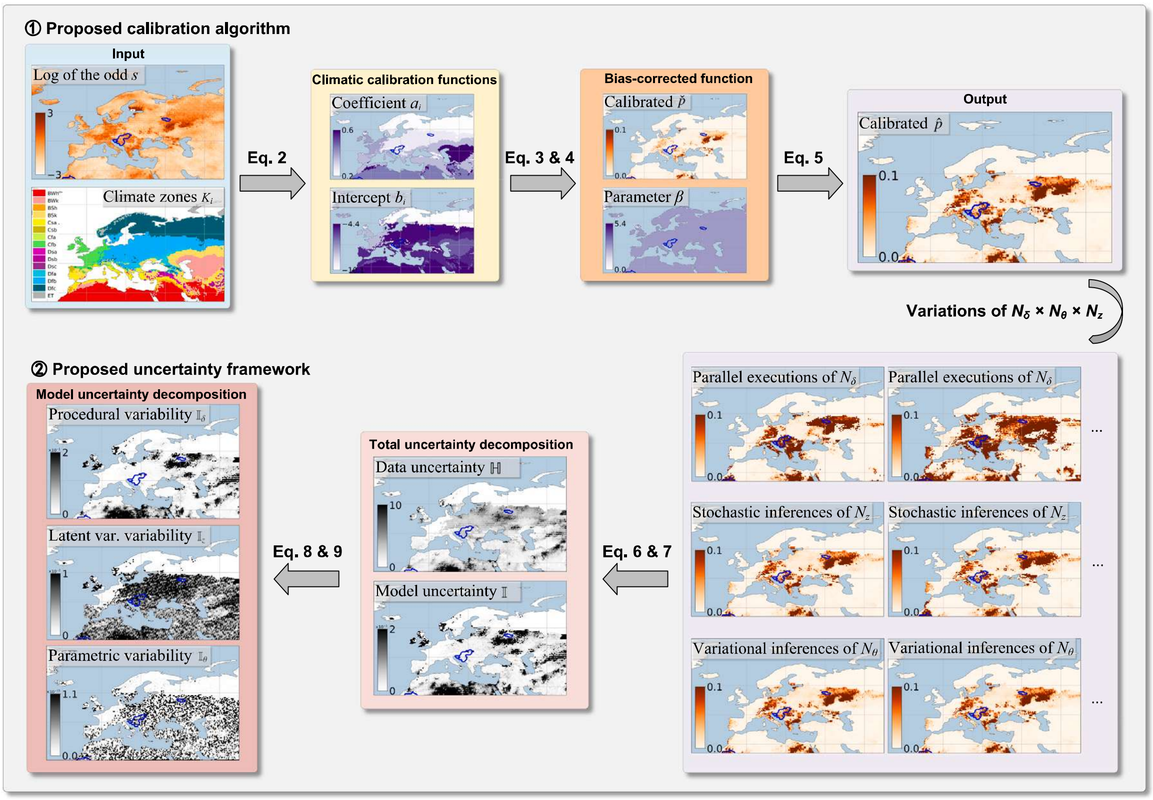

Fig. 2. Graphical illustration of the proposed calibration and uncertainty frameworks. All maps are depicted with the extended European data on 2003-04-03.

motivation is twofold: (1) climate zones with varying geographical patterns lead to distinct label priors; (2) under-sampling in the training subset has little impact on the likelihood. For each climate zone K_i, with $i \in \{1, \ldots, N\}$, we aim to alleviate the label prior shift caused by imbalanced classes of spatial–temporal occurrences:

$$(a_i{}^*, b_i{}^*) = \arg\min_{(a_i, b_i)\in\mathbb{R}^2} - \sum_{D_i} \sum_{\mathbb{C}} y \log(\check{p}), \tag{2}$$

$$\check{p} = \frac{1}{1 + \exp(-s \cdot a_i{}^* - b_i{}^*)}, \tag{3}$$

where a_i and b_i are the learned coefficient and intercept parameters, respectively; $s = \log \frac{p}{1-p}$ denotes the logarithm of the odds ratio; and D_i denotes the set of samples for each climate zone in the training subset (before under-sampling). The parameter values are derived by minimizing the negative log-likelihood (NLL).

Second, the generated probabilities from the first step are stretched with a non-linear distribution transformation. We assume that, after rectifying the label prior shift, the validation (or unseen test) subset has a similar label prior to the training subset, indicating that the likelihood shift is the primary driver of their distributional discrepancies. In this context, we extend the equation taken from Dal Pozzolo et al. (2015) to account for the only likelihood shift caused by the variability of ECVs conditioning on droughts. To reduce the intractable shift, a bias-corrected function is optimized on the validation subset as a regularization to avoid overfitting on the training subset:

$$\beta^* = \arg\min_{\beta\in(0,+\infty)} \sum_{m=1}^{M} \frac{|B_m|}{N_t} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|, \tag{4}$$

$$\hat{p} = \frac{\beta^* \check{p}}{\beta^* \check{p} - \check{p} + 1}, \tag{5}$$

where β is a learned parameter, M is the number of bins, N_t is the total number of samples, and B_m is the set of samples whose prediction confidence falls into the interval $I_m = (\frac{m-1}{M}, \frac{m}{M}]$. The expected calibration error (ECE) is minimized to achieve the lowest calibration error over the validation subset.

Ultimately, the proposed parametric algorithm corrects uncalibrated probabilities, alleviating distributional shifts in drought data to achieve less calibration error while preserving the classification accuracy.

3.2. Uncertainty quantification

Uncertainty is associated with the mismatch between models and reality, an inevitable consequence of abstracting perceptions of the phenomenal world into closed formal constructs (Hall, 2003). Drought modeling is naturally subject to a range of uncertainties (Beven et al., 2016), due to either 'Incomplete knowledge' or 'unknowable future scenarios' (Ghosh and Mujumdar, 2007). 'Incomplete knowledge' arises from inadequate information and understanding about the underlying geophysical process, leading to limitations in the modeling, known as model (epistemic) uncertainty (New and Hulme, 2000). Uncertainty due to an 'unknowable future scenario' is related to the data collection, the inherent noise, and the unpredictability in the forecast of future socioeconomic and human behavior resulting in future greenhouse gas emission scenarios, termed as data (aleatoric) uncertainty (Ghosh and Mujumdar, 2007).

Uncertainty decomposition of DL models differing in the metrics includes entropy-based and variance-based (Depeweg et al., 2018). Entropy is beneficial when the predictive posterior distribution is Bernoulli or categorical, or when the shape is complex. Depeweg et al. (2017) first studied the entropy-based decomposition of DL models. Liu et al. (2019) then separated the model uncertainty into parametric and structural components. Huang et al. (2024) dissected the model uncertainty into model approximation error, data variability, and procedural variability. Last but not least, Gawlikowski et al. (2023) provided a comprehensive review of uncertainty quantification in DL, including all kinds of uncertainty sources. However, these methods address only total model uncertainty (Depeweg et al., 2017) or rely on specialized Bayesian approaches (Liu et al., 2019; Huang et al., 2024). Bayesian approaches provide a rigorous probabilistic framework but are computationally demanding (Bishop and Nasrabadi, 2006). A tractable uncertainty framework that accounts for various uncertainty sources is needed for robust evaluations and comparisons, enabling actionable climate decision-making.

We introduce an entropy-based uncertainty framework to estimate four sources of uncertainty in DL models for droughts. First, the probability P_t of a drought conditioned on the state of some ECVs on the test set is given by:

$$P_t(Y|X=x) = \int_{\Delta,\Theta,Z} P_t(Y|X=x,(\delta,\theta,z))d\delta d\theta dz, \tag{6}$$

where $X = x$ denotes samples of ECVs where data uncertainty is reflected. The variable δ denotes DL model training algorithm randomness identified with the procedural variability, θ denotes model parameters that imply the parametric variability, and z denotes model stochastic variables that give rise to the latent variable variability. In what follows, we will omit $X = x$ for ease of notation, as each term will be conditioned on that.

First, we decompose the *entropy* in order to capture the different uncertainty terms:

$$\mathbb{H}(Y) = \mathbb{I}(Y;(\delta,\theta,z)) + \mathbb{H}(Y|(\delta,\theta,z)), \tag{7}$$

where the first term $\mathbb{H}(Y)$ is the total entropy, the second term $\mathbb{I}$ measures the mutual information representing the model uncertainty (Smith and Gal, 2018), and the third term $\mathbb{H}$ represents the conditional entropy of the data uncertainty. The mutual information is minimal when the knowledge about the model does not increase the information in the final prediction. Thus, it can be interpreted as a measure of model uncertainty (Gawlikowski et al., 2023). However, DL models are inherent to different uncertainties. The overall model uncertainty is not enough for comprehensive uncertainty assessment.

In this regard, we construct different sources of uncertainty by repeatedly introducing conditioning into the mutual information, being the model uncertainty $\mathbb{I}(Y;(\delta,\theta,z))$ further decomposed as follows:

$$\begin{aligned}\mathbb{I}(Y;(\delta,\theta,z)) = &\underbrace{\mathbb{I}(Y;(\delta,\theta,z)) - \mathbb{I}(Y;(\theta,\delta)|z^*)}_{\text{latent variable variability } \mathbb{I}_z} \\ &+ \underbrace{\mathbb{I}(Y;(\theta,\delta)|z^*) - \mathbb{I}(Y;\delta|z^*,\theta^*)}_{\text{parametric variability } \mathbb{I}_\theta} \\ &+ \underbrace{\mathbb{I}(Y;\delta|z^*,\theta^*)}_{\text{procedural variability } \mathbb{I}_\delta},\end{aligned} \tag{8}$$

where z^* and θ^* correspond to ideal cases when their uncertainties are fully eliminated. The analytical solutions for the conditional mutual information can be obtained using identities from information theory (Crooks, 2017) as follows:

$$\begin{aligned}\mathbb{I}_z &= \mathbb{H}(Y) - \mathbb{H}(Y|z), \\ \mathbb{I}_\theta &= \mathbb{H}(Y) - \mathbb{H}(Y|(\theta,z)) - \mathbb{I}_z, \\ \mathbb{I}_\delta &= \mathbb{H}(Y) - \mathbb{H}(Y|(\delta,\theta,z)) - \mathbb{I}_z - \mathbb{I}_\theta.\end{aligned} \tag{9}$$

Intuitively, these values are non-negative since conditioning always reduces entropy, and their sum always equals the overall model uncertainty by construction.

The implementation of the estimators for the uncertainty terms is described below. Specifically, the procedural variability of the training algorithm is realized with parallel executions of N_δ different randomly initial seeds. Each neural network seed triggers a stochastic optimization process converging to a sample of the posterior. The parametric variability is obtained via N_θ samples with Monte Carlo (MC) Dropout (Gal and Ghahramani, 2016), which approximates variational inference over the network weights. The latent variable variability is captured via N_z samples from the latent distribution of the VAE (Kingma and Welling, 2013). The total number of variations tested is equal to $N_\delta \times N_\theta \times N_z$. The proposed uncertainty framework is graphically illustrated in Fig. 2. To our knowledge, this framework is novel in the literature on uncertainty quantification for DL models.

Table 1
Overall calibration errors (%) for different parametric calibration algorithms in the test subset, given the DK-VRN model.

	Uncalibrated	Temp	Platt	Poly	Proposed
ECE	4.06	3.64	0.40	0.35	**0.31**
RMSCE	3.40	2.48	0.32	0.28	**0.23**
Training (s)	–	17.8	30.41	3.17	72.0
Inference (s)	–	0.07	0.07	0.03	0.07

4. Experiments and results

This section presents thorough empirical evidence of the performance of the different calibration strategies and the uncertainty evaluation framework proposed. The source code to reproduce the experiments is publicly available in GitHub.[5]

4.1. Experimental setup and evaluation metrics

First, we evaluate the proposed calibration algorithm performance by considering calibration metrics and analyzing mechanisms. The proposed algorithm is compared with uncalibrated, temperature scaling (Guo et al., 2017), Platt scaling (Platt et al., 1999), and the polynomial algorithm (Guilbert et al., 2024). Then, we provide overall metrics for the entire test subset and individual metrics for the fifteen climatic zones. Quantitative calibration metrics include expected calibration error (ECE) and root mean square calibration error (RMSCE) (Kumar et al., 2019). Next, the top-predicted calibration experiments are completed for imbalanced data. Finally, an ablation study and analyses based on reliability diagrams and distributional shifts reveal the mechanism. It should be noted that all calibration results are only reported based on DK-VRN, but the analysis is also compatible with the VLSTM model.

On the other hand, the overall uncertainty is decomposed as data uncertainty $\mathbb{H}$, latent variable variability $\mathbb{I}_z$, parametric variability $\mathbb{I}_\theta$, and procedural variability $\mathbb{I}_\delta$. At first, we analyze the impact of hyperparameters on the proposed uncertainty framework. Then, we evaluate the effectiveness of the proposed uncertainty framework by comparing different predictions and DL models. Lastly, the benefits of the proposed uncertainty framework for decision-making are examined based on ensemble predictions and rejecting uncertain predictions. The equations of all employed metrics are included in the supplementary material for clarity.

4.2. Results on model calibration

4.2.1. Analysis of calibration errors

Table 1 presents the overall calibration errors of different calibration algorithms. Notably, the proposed algorithm achieves the lowest ECE and RMSCE. About time complexity, the polynomial method is faster than others because only a linear regression is trained after feature polynomial transformation. Comparatively, the proposed algorithm comprises two-stage optimizations. Nevertheless, their time complexities are tolerable and the inferences are fast once their parameters have been determined. Thus, the proposed algorithm realizes the best overall calibration performance, with computational considerations balanced against its optimization process.

Fig. S.1 illustrates the calibration errors for the fifteen climate zones shown in Fig. 1. The proposed algorithm achieves the lowest calibration

[5] https://github.com/mengxue-rs/calibration-uncertainty-framework.

Table 2
Results (%) in the test subset for the ablation study conducted on the different components of the proposed calibration algorithm, given the DK-VRN model.

Climatic calibration	Bias-corrected	ECE	RMSCE
✗	✗	4.06	3.40
✗	✓	3.98	3.34
✓	✗	**0.31**	0.26
✓	✓	**0.31**	**0.23**

✗: not applied, ✓: applied.

Table 3
Quantified uncertainties of the DK-VRN model for the test subset (%).

	Corrected samples			Erroneous samples		
	TP+TN	TP	TN	FP+FN	FP	FN
$\mathbb{I}_z$	5.4e−3	2.7e−2	5.4e−3	3.7e−2	3.8e−2	7.4e−3
$\mathbb{I}_\theta$	2.5e−5	4.0e−16	2.5e−5	1.8e−3	1.9e−3	9.2e−17
$\mathbb{I}_\delta$	3.7e−2	3.3e+0	3.4e−2	3.1e+0	3.2e+0	6.2e−2
$\mathbb{H}$	1.7e+0	2.8e+1	1.7e+0	1.4e+1	1.5e+1	3.6e+0
Total	1.74	31.33	1.74	17.14	18.24	3.77

errors in the test subset over nine out of fifteen zones, including Csa, Cfa, Cfb, Dsa, Dsb, Dfa, Dfb, Dfc, and ET. A few exceptions are BWh, BWk, BSh, and Dsc, where other algorithms perform better. Intuitively, the reason may be the ECVs likelihood of the validation does not perfectly represent the likelihood of the test over particular zones because the validation set does not cover all climate zones (see Fig. 1(b)–(c), where validation likelihoods are missing over particular zones, such as Dsc). In contrast, all the climate zones in the train subset are covered in the test one, which allows for a comprehensive evaluation of our procedure.

Fig. S.2 depicts the calibration errors for the top predictions, from 50% to 5% (percentiles). The top predictions represent the exclusive subset of predictions for which relatively high probabilities (confidences) were estimated by DL models. It can be observed that the calibration errors of the uncalibrated and the temperature scaling approaches are sensitive to the top predictions. Conversely, the proposed calibration algorithm achieves the lowest ECE and RMSCE for the top predictions. Accordingly, the proposed algorithm can effectively reduce the calibration errors of imbalanced data and improve the reliability of crucial predictions.

4.2.2. Mechanism analysis of model calibration

In this experiment, we first perform an ablation study to analyze the contribution of the different components of the proposed calibration algorithm to its performance. Results for the variants reported in Table 2 are established by assessing different combinations of the steps followed. As can be seen, the climatic calibration functions constitute the main component contributing to the reduction of ECE and RMSCE. In contrast, the bias-corrected function decreases the error slightly. In the end, their combination achieves the best calibration performance.

The distributional shifts before and after calibration are then analyzed for the complete extended European region considered in this study and the different climate zones. Fig. S.3 depicts reliability diagrams and overall distributions for the training and the validation subset. It is noticeable that the climatic calibration functions rectify the over-confidence of predictive probabilities on the training subset. Therefore, the label prior shift before and after undersampling is effectively reduced. Then, the bias-corrected function fine-tunes the predictive probabilities to achieve the lowest validation error, reasonably decreasing the effects of likelihood shift and avoiding the model overfitting.

4.3. Results on uncertainty quantification

4.3.1. Impact of hyperparameters on uncertainty

The impact of the number of experiments conducted to estimate the different uncertainty terms is first studied. As shown in Fig. 3, all the values of uncertainties tend to converge as the sampling size (N) increases. For their intercomparison, the procedural variability $\mathbb{I}_\delta$ and the data uncertainty $\mathbb{H}$ are more stable than the other two. The reason is that the latent variable variability $\mathbb{I}_z$ and the parametric variability $\mathbb{I}_\theta$ constitute samplings from a higher-dimensional space. In this context, the proposed framework demonstrates stable calibration, where uncertainty quantification becomes more robust with larger sample sizes. This behavior aligns with fundamental statistical principles such as the

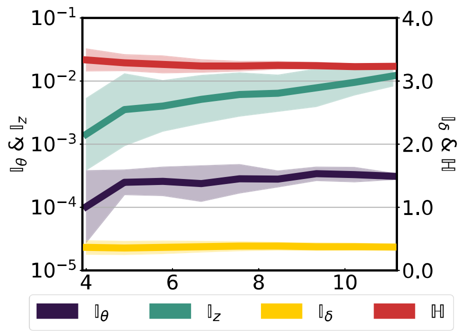

Fig. 3. Impact of the sampling size N (x-axis) on the proposed uncertainty decomposition, given the test subset (%) and the DK-VRN model.

Law of Large Numbers and the Central Limit Theorem, ensuring that empirical uncertainty estimates converge to their true values. In the following experiments, N_δ, N_θ, and N_z are empirically fixed as ten for consistency and efficiency.

Then, the effect of the MC dropout rate is reported in Fig. S.4(a). As can be seen, the MC dropout rate does not affect the uncertainty associated with the data. Comparably, the rate value affects all three variabilities related to models. Specifically, the parametric variability directly controlled by the rate intensely increases as the rate rises. Another two variabilities are indirectly affected by the rates and gradually become stable when the rate surpasses 0.5. In subsequent experiments, the MC dropout is set to 0.5, following Zhang et al. (2024).

Lastly, the effect of calibration on the uncertainty is shown in Fig. S.4(b), where uncertainty decomposition for different calibration algorithms is inspected. It can be found that the values of the estimated uncertainties are related to the calibration algorithm. Empirically, a calibration algorithm with higher calibration errors (see Table 1) results in high uncertainty estimations. For instance, temperature scaling yields the highest estimation for the data uncertainty and the three model variabilities. Conversely, the proposed calibration algorithm produces the lowest estimations for them.

4.3.2. Evaluation of uncertainty

We now separate the entire test subset into exclusive sets based on the confusion matrix: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), and the uncertainties are computed for each set. As reported in Table 3, total and data uncertainty $\mathbb{H}$ for FN are higher than those estimated for TN. It makes sense as the same probabilities but with relatively higher uncertainty usually imply mistakes. However, these estimated uncertainties are misleading when dealing with FP and TP. Conversely, $\mathbb{I}_z$ yields higher uncertainties for FP over TP. The reason may be that the data and model uncertainty influence predictions differently. The data uncertainty is helpful for FN and TN because FN samples may arise from unknowable drought samples. The model uncertainty is helpful because FP arises from the model's limitations for discriminating drought and non-drought samples. Similar results disaggregated by climatic zones can be observed in Fig. S.5. Therefore, the uncertainty framework permits further processing and analysis for uncertain samples.

Table 4
Classification performance of VLSTM and DK-VRN for different decision-making strategies, given the test subset (%).

	VLSTM		DK-VRN	
	ROC-AUC	PR-AUC	ROC-AUC	PR-AUC
Uncalibrated	66.39	0.91	72.27	2.29
Calibrated	66.39	0.91	72.27	2.29
Uncalibrated ensemble	69.96	1.32	73.61	2.45
Calibrated ensemble	69.52	1.99	**74.06**	**2.59**

ROC-AUC: Receiver Operating Characteristic Area Under the Curve.
PR-AUC: Precision-Recall Area Under the Curve.
These two metrics measure the accuracy of drought detection (classification).

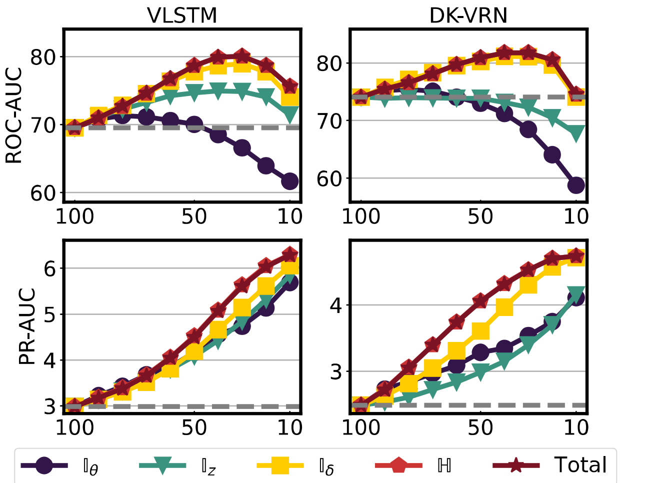

Fig. 4. Accuracy–rejection curves (%) for the VLSTM and DK-VRN models, given uncertain negative samples in the test subset. The *x*-axis denotes the keeping rate from 100% to 10%. For example, 80% keeping rate here represents rejecting the top 20% negative samples with higher uncertainty. The *y*-axis denotes the classification accuracy (ROC-AUC or PR-AUC). The calculation processes of metrics are included in the supplementary material. The dashed gray line denotes the raw results (without rejection).

Last but not least, the estimated uncertainties for VLSTM and DK-VRN are compared in Table S.2. For each variability considered to quantify model uncertainty (latent variable, parametric, procedural), accuracy and standard deviation values are computed given all the variations tested for the two remaining variabilities (see Section 3.2). DK-VRN yields lower total uncertainty, procedural variability, and latent variable variability. Our previous research claimed encoding domain knowledge as a prior/constraint distribution is beneficial to reliability and stability (Zhang et al., 2024). Besides, Table S.3 demonstrates that higher standard deviations of accuracies accompany higher uncertainties. Therefore, these uncertainties constitute an alternative to standard deviations for robustness measurements when the true labels are unknown.

4.3.3. *Decision-making based on uncertainty quantification*

This subsection discusses two practices to improve decision-making based on the proposed uncertainty framework.

Ensembling predictions. Table 4 summarizes the results for the uncalibrated models (the averaging accuracy of individuals), the calibrated models, the uncalibrated model ensemble (the accuracy of averaging probabilities provided by all the variations), and the calibrated model ensemble. First, the calibration algorithm reduces the calibration error while preserving the classification performance. The average performance for the uncalibrated models is the same as that of the calibrated ones, while the accuracy of the model ensemble is slightly altered. Second, the model ensemble yields higher performance than single models, which implies that uncertain posteriors from the proposed uncertainty framework can be synthesized to achieve more accurate estimations.

Rejecting uncertain predictions. As depicted in Fig. 4 and Fig. S.6, rejecting highly uncertain predictions can produce higher ROC-AUC and PR-AUC. For example, rejecting the top 20% uncertain negative samples based on data uncertainty $\mathbb{H}$ (Burke and Brown, 2008) yields ROC-AUCs equal to 72.72% and 76.59%, and PR-AUCs of 2.38% and 3.05%, for VLSTM and DK-VRN models, respectively, which are 3.2% and 2.53%, as well as 0.39% and 0.24% higher than the raw results. Similarly, rejecting the top 20% uncertain positive samples based on $\mathbb{I}_z$ achieves ROC-AUCs of 71.41% and 75.37%, and PR-AUCs of 3.64% and 4.65%, for VLSTM and DK-VRN models, respectively, which are 1.89% and 1.31%, as well as 1.65% and 2.66% higher than the raw results. Nevertheless, rejecting predictions means some samples do not have accessible probabilities for decision. To reject misspecified uncertain negative samples can promote warning awareness and alleviate risks. Meanwhile, unaccessible possible positive samples can yield an underestimation of losses in hazard-sensitive areas. Incorporating additional real-time hydrological information or cooperating with expert systems to make decisions for these samples in real applications is accordingly necessary.

5. Discussion

Before concluding this study, some important points related to visualization, limitations, and prospects of the approaches are worth mentioning.

5.1. *Visualizations of aggregated signals*

We first visually analyze the spatially aggregated signals of drought probabilities and uncertainties over time. The analysis overlaps five drought events across four different regions. For each drought event, all signals are determined by calculating the average output value of all pixels within each respective region per time step and then normalizing this value to lie in the range [0, 1] along the time axis.

The normalized drought probability signals of the uncalibrated, the calibrated, and the normalized uncertainty signals for the DK-VRN model are depicted in Fig. 5. First, a few false alarms occurred, such as false drought events in Afghanistan and Italy in 2014. However, all the probability signals for droughts are dominant compared to signals for non-droughts, suggesting the capability of detecting true drought events. Besides, their signals have little difference because the proposed calibration does not change the detection performance per climatic zone. For instance, the uncalibrated and calibrated signals are close in Afghanistan since the region has homogeneous climatic properties. In comparison, larger differences may reside when the area has heterogeneous climatic properties, such as in Russia.

Furthermore, false predictions are reasonably accompanied by excessive uncertainties. For instance, abnormal procedural variabilities ($\mathbb{I}_\delta$) can be inspected at the two false positives in Afghanistan, which indicate the usefulness of estimated uncertainty to ameliorate the detection performance. Moreover, the proposed uncertainty framework based on entropy can further benefit the rejecting predictions. For example, procedural variabilities present less uncertainty at the true positives while data variabilities ($\mathbb{H}$) proclaim high uncertainty at the false positives in Italy. The integration of different uncertainties may contribute to better decisions. A similar analysis for the signals resulting from the VLSTM model can be observed in Fig. S.7. Therefore, we can conclude that the proposed uncertainty framework provides more opportunities for the analysis of predictions and can be profitable for the trustworthiness of models in decision-making.

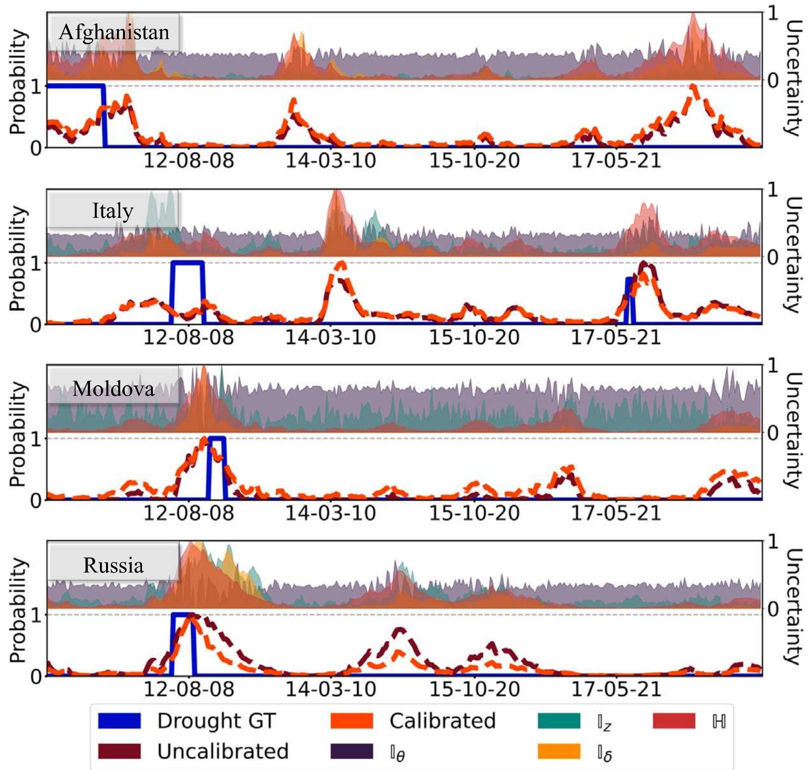

Fig. 5. Spatially aggregated signals of drought probabilities and uncertainties for the DK-VRN model, given the period 2011–2018 (test subset).

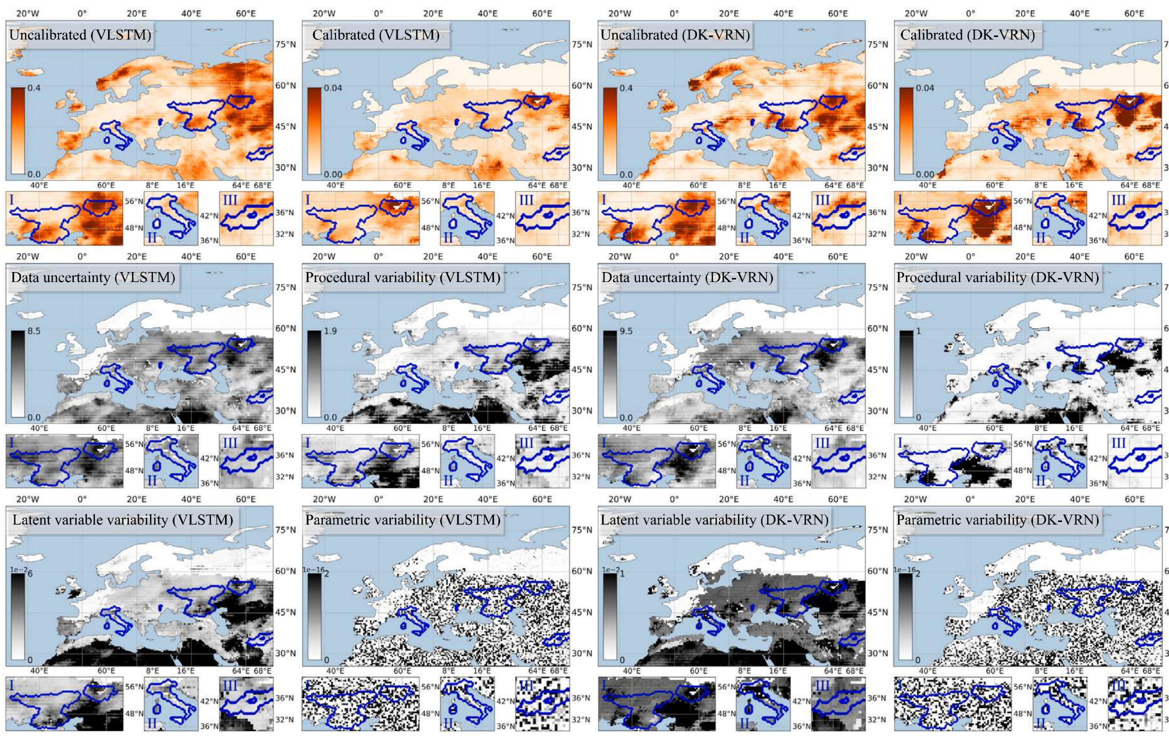

Fig. 6. Temporally aggregated (averaging all time steps) drought probability and uncertainty maps for VLSTM and DK-VRN models, given the period 2011–2018 (test subset). Three drought events occurred in this period: I (Russia), II (Italy), and III (Afghanistan). The ground truth is highlighted with blue contours. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.2. Visualizations of aggregated maps

We now visualize the temporally aggregated (averaging all time steps) maps of drought probabilities and uncertainties over Europe. Maps are presented for the uncalibrated and calibrated probabilities, as well as the different uncertainty estimates (data, latent variable, parametric, and procedural variabilities). As depicted in Fig. 6, maps for the uncalibrated and calibrated probabilities demonstrate dissimilar scales and spatial patterns. First, the scales vary from [0, 0.4] to [0, 0.04], which implies the proposed calibration algorithm generally decreases the overconfident probabilities. Specifically, the locations of zones with fewer drought samples (lower label prior distribution) are re-assigned to lower probabilities by the climatic calibration functions. For example, the drought probabilities in Northern Europe are depressed based on historical drought data, resulting in less calibration error of Dfc, as reported in Fig. S.1.

Moreover, distinct ranges and patterns are presented in uncertainty maps. These ranges correspond to the contribution of each uncertainty component over the total uncertainty. For instance, data uncertainty is the most significant, while parametric variability is the least significant, where a remarkable logarithmic difference can be observed based on the scales of the lower left colorbars in Fig. 6. Analogous conclusions were found for drought indices (Laimighofer and Laaha, 2022). The proposed uncertainty framework allows further understanding and comparison of DL models. For instance, DK-VRN provides less latent variable and procedural variability than VLSTM, indicating that the effective incorporation of domain knowledge can reduce the corresponding uncertainty. Therefore, the proposed uncertainty framework provides opportunities for uncovering reasons for instability and supporting robust evaluations or comparisons in unpredictable scenarios.

5.3. Limitations and prospects

While the calibration and the uncertainty on DL-based drought detection are investigated, the following limitations should be outlined. Firstly, different data sources are not taken into account. The proposed uncertainty framework estimates the data uncertainty only considering ERA-5 data. Future efforts may count multi-source remote sensing precipitation datasets (Huffman et al., 2015; Ashouri et al., 2015) with different temporal and spatial resolutions. Incorporating new input variables could help to understand or even decrease uncertainty, and this is an interesting direction to explore in future research, based on the proposed methodology in this paper. Secondly, the label noise or error is not considered. Admittedly, the labels of drought events are published by responsible specialists who undertake periodic quality control. However, the criteria for drought are controversial. Techniques like noisy labels or weak supervision may be helpful in this case (Cortés-Andrés et al., 2024). For instance, label correction can improve model regularization by adding similar structural labels, making them soft, and correcting noisy semantic classes. Domain knowledge can also be encoded, providing multiple sources of weak supervision as signals for inferring the latent labels. Last but not least, testing the proposed uncertainty framework on different climate projections can provide valuable insights into understanding how DL models perform in unpredictable long-term scenarios with differences in greenhouse gas emissions (Stocker et al., 2013).

6. Conclusion

We have studied the calibration and uncertainty quantification for DL-based drought detection, bridging the gap between existing DL models and trustworthy drought detection. A parametric calibration algorithm was proposed to mitigate the distributional shift of probabilities to achieve less calibration error. Then, an entropy-based uncertainty framework is proposed for uncertainty decomposition and quantification. Experimental results on ERA-5 data and considering European drought events have validated the effectiveness of the proposed calibration algorithm and uncertainty framework. We showcased that the proposed calibration algorithm improves the reliability of predictive probabilities, while the proposed uncertainty framework benefits the trustworthiness of the decision-making. Although our experimental results are encouraging, we plan to investigate further from two perspectives: (1) quantify data uncertainty for drought detection given multi-source data; (2) explore model calibration and uncertainty quantification under different climate projections.

CRediT authorship contribution statement

Mengxue Zhang: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Miguel-Ángel Fernández-Torres:** Writing – review & editing, Supervision, Formal analysis, Conceptualization. **Kai-Hendrik Cohrs:** Writing – review & editing, Formal analysis, Conceptualization. **Gustau Camps-Valls:** Writing – review & editing, Supervision, Software, Resources, Project administration, Funding acquisition, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

M Z appreciates the financial support from the China Scholarship Council (CSC) through the State Scholarship Fund for Overseas Study (No. 202106710031). All authors acknowledge the support from the European Research Council (ERC) under the ERC Synergy Grant USMILE (grant agreement 855187), the European Union‘s Horizon 2020 research and innovation program within the project ‘XAIDA: Extreme Events - Artificial Intelligence for Detection and Attribution,’ (GA 101003469), ‘AI4PEX: Artificial Intelligence and Machine Learning for Enhanced Representation of Processes and Extremes in Earth System Models’ (GA 101137682) and the computer resources provided by the Jülich Supercomputing Centre (JSC) (Project No. PRACE-DEV-2022D01-048), the computer resources provided by Artemisa (funded by the European Union ERDF and Comunitat Valenciana), as well as the technical support provided by the Instituto de Física Corpuscular, IFIC (CSIC-UV).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jag.2025.104563.

Data availability

The link to the data and code implementation is provided in the manuscript (see Section 4).

References

Ashouri, H., Hsu, K.-L., Sorooshian, S., Braithwaite, D.K., Knapp, K.R., Cecil, L.D., Nelson, B.R., Prat, O.P., 2015. PERSIANN-CDR: Daily precipitation climate data record from multisatellite observations for hydrological and climate studies. Bull. Am. Meterol. Soc. 96 (1), 69–83. http://dx.doi.org/10.1175/BAMS-D-13-00068.1.

Ault, T.R., 2020. On the essentials of drought in a changing climate. Sci. 368 (6488), 256–260. http://dx.doi.org/10.1126/science.aaz5492.

Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood, E.F., 2018. Present and future Köppen-Geiger climate classification maps at 1-km resolution. Sci. Data 5 (1), 1–12. http://dx.doi.org/10.1038/sdata.2018.214.

Beven, K., Almeida, S., Aspinall, W., Bates, P., Blazkova, S., Borgomeo, E., Goda, K., Phillips, J., Simpson, M., Smith, P., et al., 2016. Epistemic uncertainties and natural hazard risk assessment–part 2: Different natural hazard areas. Nat. Hazards Earth Syst. Sci. Discuss. 2016, 1–74. http://dx.doi.org/10.5194/nhess-18-2769-2018.

Bishop, C.M., Nasrabadi, N.M., 2006. Pattern Recognition and Machine Learning, vol. 4, Springer, URL: https://link.springer.com/book/9780387310732.

Burke, E.J., Brown, S.J., 2008. Evaluating uncertainties in the projection of future drought. J. Hydrometeorol. 9 (2), 292–299. http://dx.doi.org/10.1175/2007JHM929.1.

Camps-Valls, G., Fernández-Torres, M.-Á., Cohrs, K.-H., Höhl, A., Castelletti, A., Pacal, A., Robin, C., Martinuzzi, F., Papoutsis, I., Prapas, I., et al., 2025. Artificial intelligence for modeling and understanding extreme weather and climate events. Nat. Commun. 16 (1), 1919. http://dx.doi.org/10.1038/s41467-025-56573-8.

Camps-Valls, G., Tuia, D., Zhu, X.X., Reichstein, M., 2021. Deep learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science and Geosciences. John Wiley & Sons, http://dx.doi.org/10.1002/9781119646181.

Carrão, H., Russo, S., Sepulcre-Canto, G., Barbosa, P., 2016. An empirical standardized soil moisture index for agricultural drought assessment from remotely sensed data. Int. J. Appl. Earth Obs. Geoinformation 48, 74–84. http://dx.doi.org/10.1016/j.jag.2015.06.011.

Chattopadhyay, A., Mustafa, M., Hassanzadeh, P., Kashinath, K., 2020. Deep spatial transformers for autoregressive data-driven forecasting of geophysical turbulence. In: CI 2020. pp. 106–112. http://dx.doi.org/10.31223/osf.io/cqmb2.

Chen, J., Su, B., 2023. Transfer knowledge from head to tail: Uncertainty calibration under long-tailed distribution. In: IEEE/ CVF CVPR 2023. pp. 19978–19987, URL: https://openaccess.thecvf.com/content/CVPR2023/papers/Chen_Transfer_Knowledge_From_Head_to_Tail_Uncertainty_Calibration_Under_Long-Tailed_CVPR_2023_paper.pdf.

Cortés-Andrés, J., Fernández-Torres, M.-Á., Camps-Valls, G., 2024. Deep learning with noisy labels for spatio-temporal drought detection. IEEE Trans. Geosci. Remote Sens. http://dx.doi.org/10.1109/TGRS.2024.3504340.

Crooks, G.E., 2017. On measures of entropy and information. Tech. Note 9 (4), URL: https://threeplusone.com/pubs/on_information.pdf.

Dal Pozzolo, A., Caelen, O., Johnson, R.A., Bontempi, G., 2015. Calibrating probability with undersampling for unbalanced classification. In: IEEE SSCI 2015. pp. 159–166, URL: https://www3.nd.edu/~dial/publications/dalpozzolo2015calibrating.pdf.

de Andrade, B.C.C., de Andrade Pinto, E.J., Ruhoff, A., Senay, G.B., 2021. Remote sensing-based actual evapotranspiration assessment in a data-scarce area of Brazil: A case study of the urucuia aquifer system. Int. J. Appl. Earth Obs. Geoinformation 98, 102298. http://dx.doi.org/10.1016/j.jag.2021.102298.

Depeweg, S., Hernández-Lobato, J.M., Doshi-Velez, F., Udluft, S., 2017. Uncertainty decomposition in bayesian neural networks with latent variables. http://dx.doi.org/10.48550/arXiv.1706.08495, ArXiv. Prepr.

Depeweg, S., Hernandez-Lobato, J.-M., Doshi-Velez, F., Udluft, S., 2018. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In: ICML 2018. pp. 1184–1193, URL: https://proceedings.mlr.press/v80/depeweg18a/depeweg18a.pdf.

Dong, J., Jiang, Z., Pan, D., Chen, Z., Guan, Q., Zhang, H., Gui, G., Gui, W., 2024. A survey on confidence calibration of deep learning under class imbalance data. Authorea Prepr. http://dx.doi.org/10.36227/techrxiv.171623912.22862601/v1.

Faghani, S., Moassefi, M., Rouzrokh, P., Khosravi, B., Baffour, F.I., Ringler, M.D., Erickson, B.J., 2023. Quantifying uncertainty in deep learning of radiologic images. Radiol. 308 (2), e222217. http://dx.doi.org/10.1148/radiol.222217.

Foroumandi, E., Nourani, V., Huang, J.J., Moradkhani, H., 2023. Drought monitoring by downscaling GRACE-derived terrestrial water storage anomalies: A deep learning approach. J. Hydrol. 616, 128838. http://dx.doi.org/10.1016/j.jhydrol.2022.128838.

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: ICML 2016. pp. 1050–1059, URL: https://proceedings.mlr.press/v48/gal16.pdf.

Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al., 2023. A survey of uncertainty in deep neural networks. Artif. Intell. Rev. 56 (Suppl 1), 1513–1589. http://dx.doi.org/10.1007/s10462-023-10562-9.

Ghosh, S., Mujumdar, P., 2007. Nonparametric methods for modeling GCM and scenario uncertainty in drought assessment. Water Resour. Res. 43 (7), http://dx.doi.org/10.1029/2006WR005351.

Gonzalez-Calabuig, M., Cortes-Andres, J., Williams, T.K.E., Zhang, M., Pellicer-Valero, O.J., Fernández-Torres, M.-A., Camps-Valls, G., 2024. The AIDE toolbox: Artificial intelligence for disentangling extreme events [software and data sets]. IEEE Geosci. Remote. Sens. Mag. 12 (2), 113–118. http://dx.doi.org/10.1109/MGRS.2024.3382544.

Guha-Sapir, D., Below, R., Hoyois, P., 2015. EM-DAT: International Disaster Database. Université Catholique de Louvain, Brussels, Belgium, [dataset], URL: https://www.emdat.be/.

Guilbert, T., Caelen, O., Chirita, A., Saerens, M., 2024. Calibration methods in imbalanced binary classification. Ann. Math. Artif. Intell. http://dx.doi.org/10.1007/s10472-024-09952-8.

Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. On calibration of modern neural networks. In: ICML 2017. pp. 1321–1330, URL: https://proceedings.mlr.press/v70/guo17a/guo17a.pdf.

Haile, G.G., Tang, Q., Li, W., Liu, X., Zhang, X., 2020. Drought: Progress in broadening its understanding. Wiley Interdiscip. Rev.: Water 7 (2), e1407. http://dx.doi.org/10.1002/wat2.1407.

Hall, J.W., 2003. Handling uncertainty in the hydroinformatic process. J. Hydroinformormation 5 (4), 215–232. http://dx.doi.org/10.2166/hydro.2003.0019.

Hoffmann, D., Gallant, A.J., Arblaster, J.M., 2020. Uncertainties in drought from index and data selection. J. Geophys. Res. 125 (18), e2019JD031946. http://dx.doi.org/10.1029/2019JD031946.

Huang, Z., Lam, H., Zhang, H., 2024. Efficient uncertainty quantification and reduction for over-parameterized neural networks. In: NeurIPS 2024. Vol. 36, URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/cb2266111eadcfa2c02187ace64e2183-Paper-Conference.pdf.

Huffman, G.J., Bolvin, D.T., Nelkin, E.J., Tan, J., 2015. Integrated multi-satellite retrievals for GPM (IMERG) technical documentation. Nasa/ Gsfc Code 612 (47), 2019.

Jiang, J., Zhou, T., 2023. Agricultural drought over water-scarce central Asia aggravated by internal climate variability. Nat. Geosci. 16 (2), 154–161. http://dx.doi.org/10.1038/s41561-022-01111-0.

Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. http://dx.doi.org/10.48550/arXiv.1312.6114, ArXiv. Prepr.

Koppa, A., Keune, J., Schumacher, D.L., Michaelides, K., Singer, M., Seneviratne, S.I., Miralles, D.G., 2024. Dryland self-expansion enabled by land–atmosphere feedbacks. Sci. 385 (6712), 967–972. http://dx.doi.org/10.1126/science.adn6833.

Kumar, A., Liang, P.S., Ma, T., 2019. Verified uncertainty calibration. In: NeurIPS 2019. Vol. 32, URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/f8c0c968632845cd133308b1a494967f-Paper.pdf.

Laimighofer, J., Laaha, G., 2022. How standard are standardized drought indices? Uncertainty components for the SPI & SPEI case. J. Hydrol. 613, 128385. http://dx.doi.org/10.1016/j.jhydrol.2022.128385.

Liu, J., Paisley, J., Kioumourtzoglou, M.-A., Coull, B., 2019. Accurate uncertainty estimation and decomposition in ensemble learning. In: NeurIPS 2019. Vol. 32, URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/1cc8a8ea51cd0adddf5dab504a285915-Paper.pdf.

Mahecha, M.D., Gans, F., Brandt, G., Christiansen, R., Cornell, S.E., Fomferra, N., Kraemer, G., Peters, J., Bodesheim, P., Camps-Valls, G., et al., 2020. Earth system data cubes unravel global multivariate dynamics. Earth Syst. Dyn. 11 (1), 201–234. http://dx.doi.org/10.5194/esd-11-201-2020.

Materia, S., García, L.P., van Straaten, C., O, S., Mamalakis, A., Cavicchia, L., Coumou, D., de Luca, P., Kretschmer, M., Donat, M., 2024. Artificial intelligence for climate prediction of extremes: State of the art, challenges, and future perspectives. Wiley Interdiscip. Rev.: Clim. Chang. 15 (6), e914. http://dx.doi.org/10.1002/wcc.914.

McDonald, A., Tan, P.-N., Luo, L., 2022. Comet flows: Towards generative modeling of multivariate extremes and tail dependence. http://dx.doi.org/10.48550/arXiv.2205.01224, ArXiv. Prepr.

McKee, T.B., Doesken, N.J., Kleist, J., et al., 1993. The relationship of drought frequency and duration to time scales. In: CAC 1993. Vol. 17, pp. 179–183, URL: https://climate.colostate.edu/pdfs/relationshipofdroughtfrequency.pdf.

McMillan, H., Seibert, J., Petersen-Overleir, A., Lang, M., White, P., Snelder, T., Rutherford, K., Krueger, T., Mason, R., Kiang, J., 2017. How uncertainty analysis of streamflow data can reduce costs and promote robust decisions in water management applications. Water Resour. Res. 53 (7), 5220–5228. http://dx.doi.org/10.1002/2016WR020328.

McMillan, H.K., Westerberg, I.K., Krueger, T., 2018. Hydrological data uncertainty and its implications. Wiley Interdiscip. Rev.: Water 5 (6), e1319. http://dx.doi.org/10.1002/wat2.1319.

Miralles, D.G., Gentine, P., Seneviratne, S.I., Teuling, A.J., 2019. Land–atmospheric feedbacks during droughts and heatwaves: state of the science and current challenges. Ann. New York Acad. Sci. 1436 (1), 19–35. http://dx.doi.org/10.1111/nyas.13912.

Mishra, A.K., Singh, V.P., 2010. A review of drought concepts. J. Hydrol. 391 (1–2), 202–216. http://dx.doi.org/10.1016/j.jhydrol.2010.07.012.

Mukherjee, S., Mishra, A., Trenberth, K.E., 2018. Climate change and drought: a perspective on drought indices. Curr. Clim. Chang. Rep. 4, 145–163. http://dx.doi.org/10.1007/s40641-018-0098-x.

New, M., Hulme, M., 2000. Representing uncertainty in climate change scenarios: a Monte-Carlo approach. Integr. Assess. 1 (3), 203–213. http://dx.doi.org/10.1023/A:1019144202120.

Orlowsky, B., Seneviratne, S.I., 2013. Elusive drought: uncertainty in observed trends and short-and long-term CMIP5 projections. Hydrol. Earth Syst. Sci. 17 (5), 1765–1781. http://dx.doi.org/10.5194/hess-17-1765-2013.

Palmer, W.C., 1965. Meteorological Drought, vol. 30, US Department of Commerce, Weather Bureau, URL: https://www.droughtmanagement.info/literature/USWB_Meteorological_Drought_1965.pdf.

Platt, J., et al., 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv. Large Margin Classif. 10 (3), 61–74, URL: https://home.cs.colorado.edu/~mozer/Teaching/syllabi/6622/papers/Platt1999.pdf.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., et al., 2021. Skilful precipitation nowcasting using deep generative models of radar. Nat. 597 (7878), 672–677. http://dx.doi.org/10.1038/s41586-021-03854-z.

Rosvold, E.L., Buhaug, H., 2021. GDIS, a global dataset of geocoded disaster locations. Sci. Data 8 (1), 61. http://dx.doi.org/10.1038/s41597-021-00846-6.

Rúvžívčka, V., Vaughan, A., De Martini, D., Fulton, J., Salvatelli, V., Bridges, C., Mateo-Garcia, G., Zantedeschi, V., 2022. RaV´AEn: unsupervised change detection of extreme events using ML on-board satellites. Sci. Rep. 12 (1), 16939. http://dx.doi.org/10.1038/s41598-022-19437-5.

Schumacher, M., Forootan, E., van Dijk, A.I., Schmied, H.M., Crosbie, R.S., Kusche, J., Döll, P., 2018. Improving drought simulations within the Murray-Darling Basin by combined calibration/assimilation of GRACE data into the WaterGAP global hydrology model. Remote Sens. Environ. 204, 212–228. http://dx.doi.org/10.1016/j.rse.2017.10.029.

Schumacher, D.L., Keune, J., Dirmeyer, P., Miralles, D.G., 2022. Drought self-propagation in drylands due to land–atmosphere feedbacks. Nat. Geosci. 15 (4), 262–268. http://dx.doi.org/10.1038/s41561-022-00912-7.

Sepulcre-Canto, G., Horion, S., Singleton, A., Carrao, H., Vogt, J., 2012. Development of a combined drought indicator to detect agricultural drought in Europe. Nat. Hazards Earth Syst. Sci. 12 (11), 3519–3531. http://dx.doi.org/10.5194/nhess-12-3519-2012.

Sheffield, J., Wood, E.F., Roderick, M.L., 2012. Little change in global drought over the past 60 years. Nat. 491 (7424), 435–438. http://dx.doi.org/10.1038/nature11575.

Shen, R., Huang, A., Li, B., Guo, J., 2019. Construction of a drought monitoring model using deep learning based on multi-source remote sensing data. Int. J. Appl. Earth Obs. Geoinformation 79, 48–57. http://dx.doi.org/10.1016/j.jag.2019.03.006.

Smith, L., Gal, Y., 2018. Understanding measures of uncertainty for adversarial example detection. ArXiv. Prepr.

Stocker, B.D., Roth, R., Joos, F., Spahni, R., Steinacher, M., Zaehle, S., Bouwman, L., Prentice, I.C., 2013. Multiple greenhouse-gas feedbacks from the land biosphere under future climate change scenarios. Nat. Clim. Chang. 3 (7), 666–672. http://dx.doi.org/10.1038/nclimate1864.

Tian, J., Liu, Y.-C., Glaser, N., Hsu, Y.-C., Kira, Z., 2020. Posterior re-calibration for imbalanced datasets. In: NeurIPS 2020. Vol. 33, pp. 8101–8113, URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/5ca359ab1e9e3b9c478459944a2d9ca5-Paper.pdf.

Trenberth, K.E., Dai, A., Van Der Schrier, G., Jones, P.D., Barichivich, J., Briffa, K.R., Sheffield, J., 2014. Global warming and changes in drought. Nat. Clim. Chang. 4 (1), 17–22. http://dx.doi.org/10.1038/nclimate2067.

Van Loon, A.F., 2015. Hydrological drought explained. Wiley Interdiscip. Rev.: Water 2 (4), 359–392. http://dx.doi.org/10.1002/wat2.1085.

Vicente-Serrano, S.M., Beguería, S., López-Moreno, J.I., 2010. A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. J. Clim. 23, 1696–1718. http://dx.doi.org/10.1175/2009JCLI2909.1.

Wells, N., Goddard, S., Hayes, M.J., 2004. A self-calibrating palmer drought severity index. J. Clim. 17 (12), 2335–2351. http://dx.doi.org/10.1175/1520-0442(2004)017<2335:ASPDSI>2.0.CO;2.

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 Sens. Environ. 233, 111358. http://dx.doi.org/10.1016/j.rse.2019.111358.

Xu, L., Chen, N., Yang, C., 2021. Quantifying the uncertainty of precipitation forecasting using probabilistic deep learning. Hydrol. Earth Syst. Sci. Discuss. 2021, 1–27. http://dx.doi.org/10.5194/hess-26-2923-2022.

Yosri, A., Ghaith, M., El-Dakhakhni, W., 2024. Deep learning rapid flood risk predictions for climate resilience planning. J. Hydrol. 631, 130817. http://dx.doi.org/10.1016/j.jhydrol.2024.130817.

Zargar, A., Sadiq, R., Khan, F.I., 2014. Uncertainty-driven characterization of climate change effects on drought frequency using enhanced SPI. Water Resour. Manag. 28, 15–40. http://dx.doi.org/10.1007/s11269-013-0467-0.

Zhang, M., Fernández-Torres, M.-Á., Camps-Valls, G., 2024. Domain knowledge-driven variational recurrent networks for drought monitoring. Remote Sens. Environ. 311, 114252. http://dx.doi.org/10.1016/j.rse.2024.114252.

Zhang, Q., Shi, R., Xu, C.-Y., Sun, P., Yu, H., Zhao, J., 2022. Multisource data-based integrated drought monitoring index: Model development and application. J. Hydrol. 615, 128644. http://dx.doi.org/10.1016/j.jhydrol.2022.128644.

Zhao, Q., Liu, K., Sun, T., Yao, Y., Li, Z., 2023. A novel regional drought monitoring method using GNSS-derived ZTD and precipitation. Remote Sens. Environ. 297, 113778. http://dx.doi.org/10.1016/j.rse.2023.113778.