Calibration and Uncertainty for DL-Based Drought Modeling

Mengxue Zhang, PhD student

Image Processing Laboratory
Universitat de València

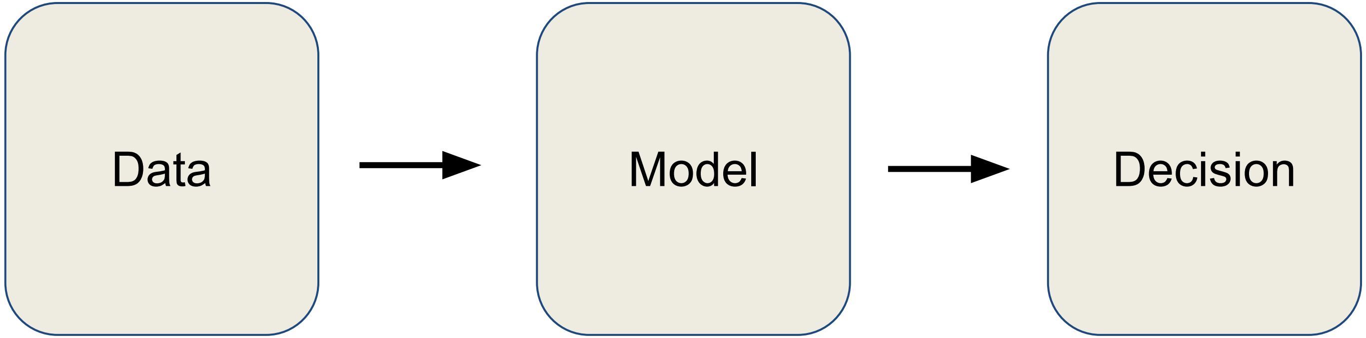
ISP · Image & Signal Processing
Universitat de València

http://isp.uv.es @mengxue_rs mengxue@alumni.uv.es

May 2025

Part I - Background

Drought Modeling (DM)



Drought Modeling (DM)

Data

- Meteorological stations;
- Process simulations;
- Satellite-derived;
- Reanalysis;
- Reports;
- …

Drought Modeling (DM)

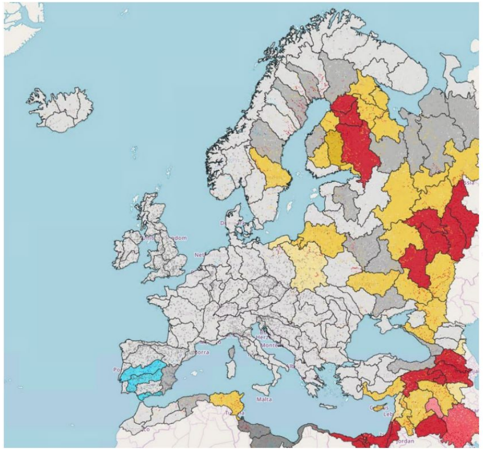
Model

- GCMs/RCMs
- LSMs/HMs
- Impact models
- Drought indices
- ML/DL models
- …

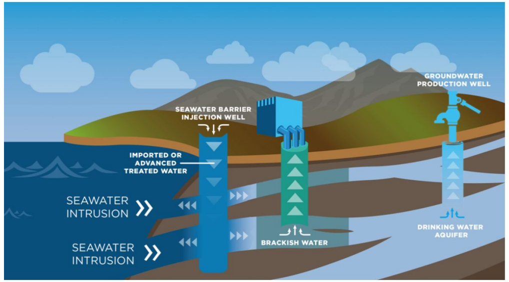
Drought Modeling (DM)

Decision

a) Characterization



b) Mitigation (Water transfers)



Drought Modeling (DM)

Decision

c) Mitigation (Infrastructures)

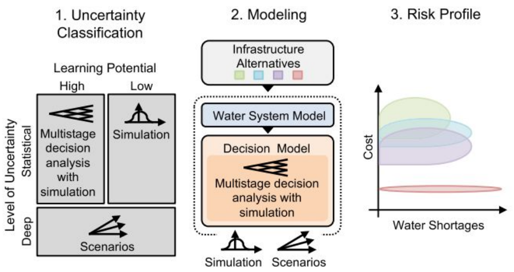
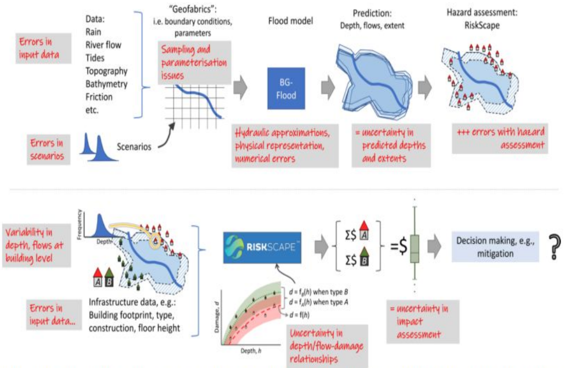


Fig. 1. Method for decision making under multiple uncertainties; Step 1: uncertainties are classified according to whether they are deep or statistical and whether they have high or low potential for learning as uncertainties are realized over time; Step 2: water system and decision models evaluate infrastructure alternatives under these uncertainties; Step 3: performance indicators are displayed in a risk profile for decision support

d) Hazard Assessments



Top: errors from multiple sources "cascades" through the analysis chain, resulting in uncertainty in predicted depths and extents and the derived hazard assessment; Bottom: analysis using the uncertain hazard assessment creates a further cascade, leading to uncertainty in the impact assessment and causing issues in decision making.

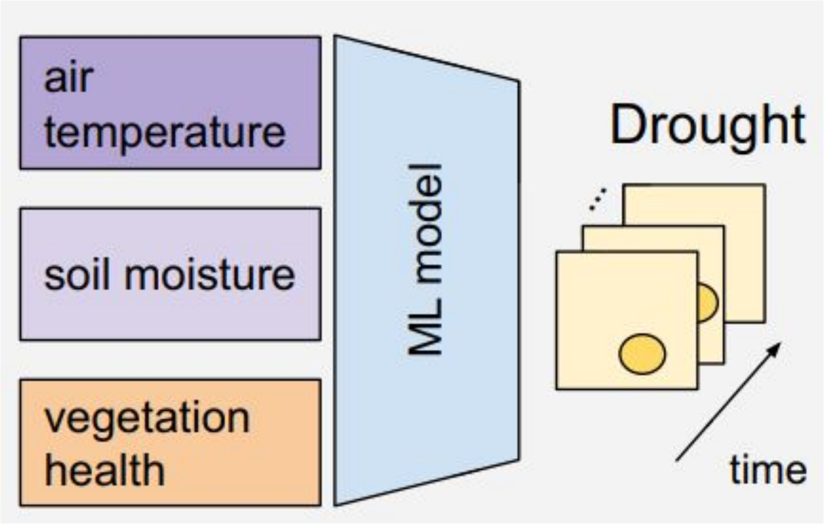
DL-Based Drought Modeling

Motivation

(1) DL models have succeeded in detecting extreme climate events and promise to uncover and understand droughts further;

(2) However, few studies discuss the calibration and UQ for DL-based drought modeling;

(3) improving drought modeling by the identification of spatial and temporal patterns that relate to the true process rather than just bias and uncertainty.



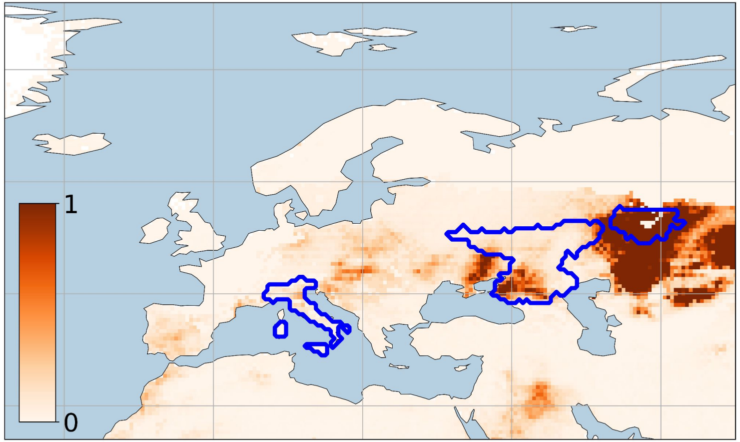
DL-Based Drought Modeling

Table S.1
Architecture details for the DL models considered in this study.

	VLSTM	DK-VRN		
Input	T, 2	T, 2		T, 3
Layer 0	LSTM, 32	LSTM, 32		LSTM, 32
Layer 1	LSTM, 32	LSTM, 32		LSTM, 32
Layer 2	Linear, 32+32	Linear, 32+32		Linear, 32+32
Layer 3	REP	REP		REP
Layer 4	**Linear, 1**	**Linear, 1**	LSTM, 32	LSTM, 32
Layer 5	-	-	**Linear, 2**	**Linear, 3**
Output	T, 1	T, 1	T, 2	T, 3

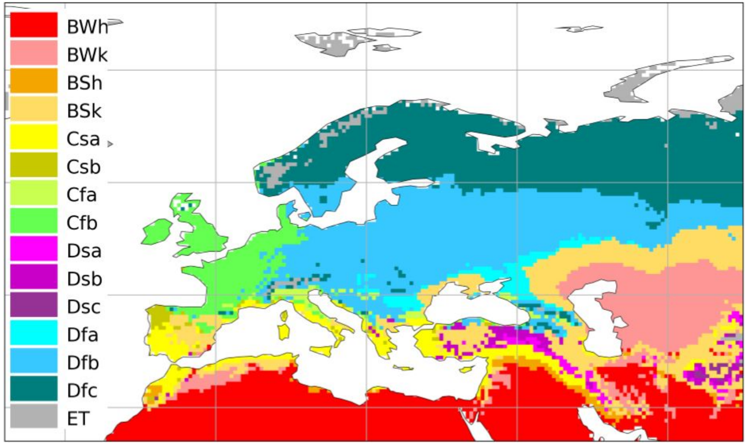
VLSTM is based on a single-branch architecture and DK-VRN is based on a triple-branch architecture. The number after ',' indicates output dimensions, e.g., 'Linear, 32+32' means the output is a tuple with the dimensions of 32 and 32 after a linear layer. 'REP' means the reparameterization step. The last layers are highlighted in **bold**. T is the length of the input and output time series. It should be noted that only the first branch of DK-VRN, whose outpu dimension is T, 1, is activated at the inference stage. More explanations are referred to (Zhang et al., 2024).

ERA5 Reanalysis +



DL-based Pred & EM-DAT

+



Climate classification

DL-Based Drought Modeling

- ➢ Calibration for DL-based DM?
- ➢ UQ for DL-based DM?
- ➢ Decision-making under uncertainty for DL-based DM?

Part II - Calibration for DM

Calibration for DM

Model

- GCMs/RCMs
- LSMs/HMs
- Impact models
- Drought indices
- ML/DL models
- …

Calibration for DM

Model

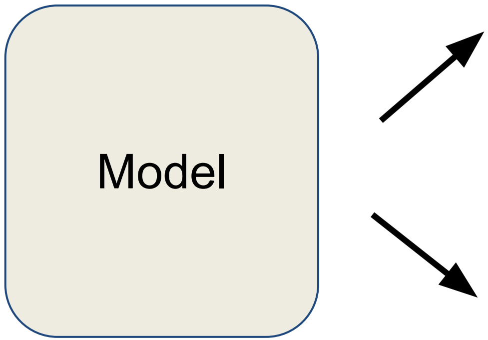
This paper introduces that **climate models** have a limited spatial resolution, their topography is coarse and they will never resolve all processes from planetary waves down to turbulence. Many relevant atmospheric, oceanic and coupled processes are not realistically represented, with knock-on effects on other processes even far away from where the primary biases occur. The bias such as mean and variance are therefore commonplace. To reduce the biases, **bias correction** are proposed from simple adjustments of the mean, multivariate, quantile mapping approaches. As bias correction mere deal with the output, so they are considered as **post-processing calibrations**.

Maraun, Douglas, et al. "Towards process-informed bias correction of climate change simulations." *Nat. Clim. Change.* 2017.

This paper mentions that **LSM**'s output, the soil moisture, is especially sensitive to **soil hydraulic properties**, in particular for the semi-arid sites. The computation of soil hydraulic properties is typically approximated in the vertical dimension by LSMs. The calibration using **in-situ or indirect measurements of soil hydraulic properties** can significantly reduce the simulation errors.

Harrison, Kenneth W., et al. "Quantifying the change in soil moisture modeling uncertainty from remote sensing observations using Bayesian inference techniques." *Water Resour. Res.* 2012

Calibration for DM



Hydrological models describe the natural processes of the water cycle. Due to the large complexity of the natural phenomena, these models contain **substantial simplifications**. They consist of basic equations, often loosely based on physical premises, whose parameters are specific for the selected catchment and problem under study. Some parameters can not be considered as physically measured (or measurable) quantities and have to be estimated on the basis of the available data and information. The calibration of a regionalisation of the hydrological model parameters relies on the **basis of catchment characteristics**.

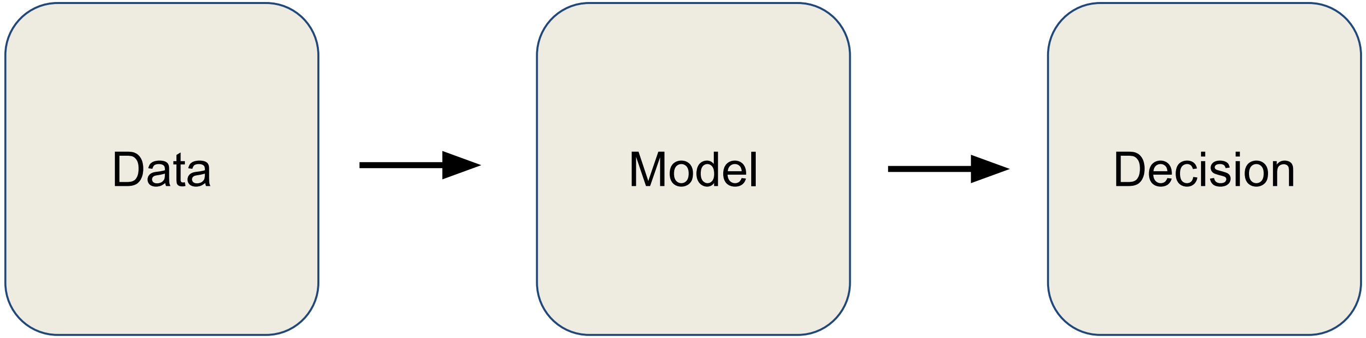
Bárdossy, András. "Calibration of hydrological model parameters for ungauged catchments." *Hydrol. Earth Syst. Sci*. 2007.

PDSI depends on a two-stage bucket model of the soil. The top layer of soil is assumed to hold one inch of moisture. The amount of moisture that can be held by the rest of the underlying soil is a location-dependent value, which must be provided as an input parameter to the program.**The SC-PDSI replaces the empirically derived** climatic characteristics (K) and duration factors (0.897 and 1/3) with **values automatically calculated based upon the historical climatic data** of a location.

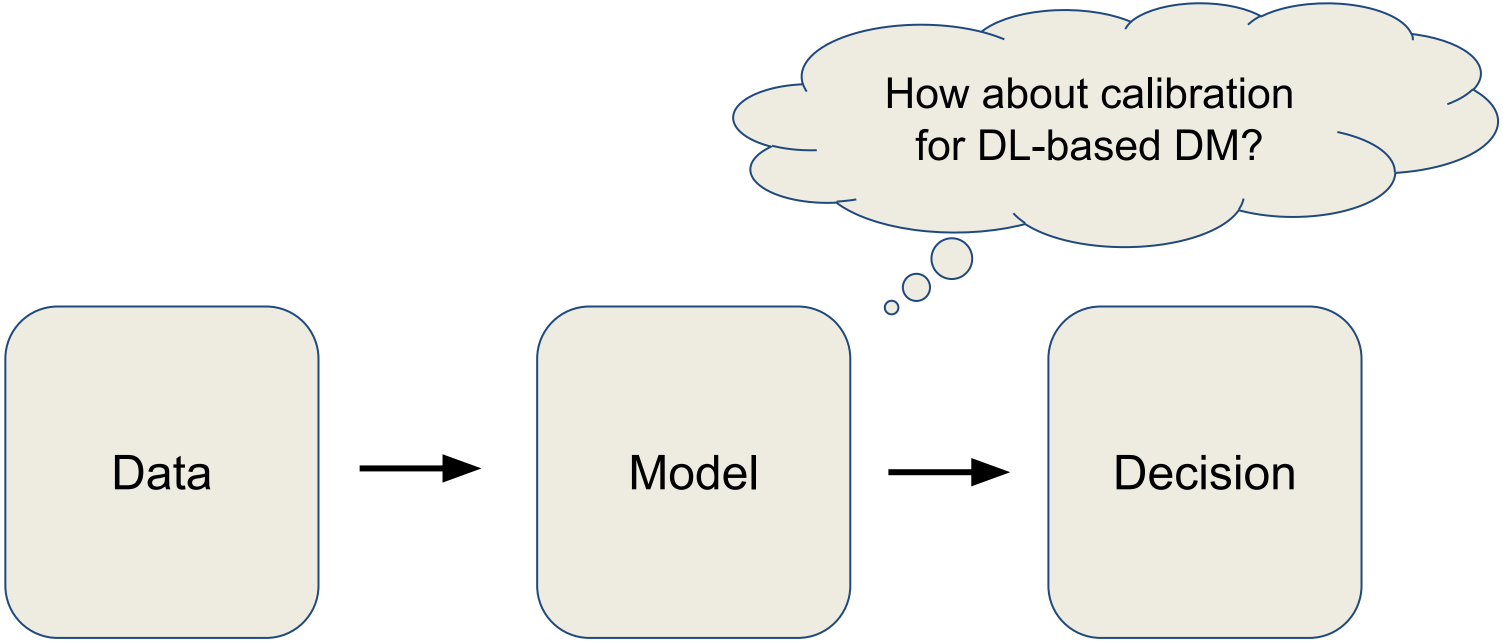
Wells, Nathan, Steve Goddard, and Michael J. Hayes. "A self-calibrating Palmer drought severity index." *J. Clim.* 2004.

Calibration for DM

reasons: process-data mismatch
solutions: observational data rectifies process simplifications



Calibration for DL-based DM

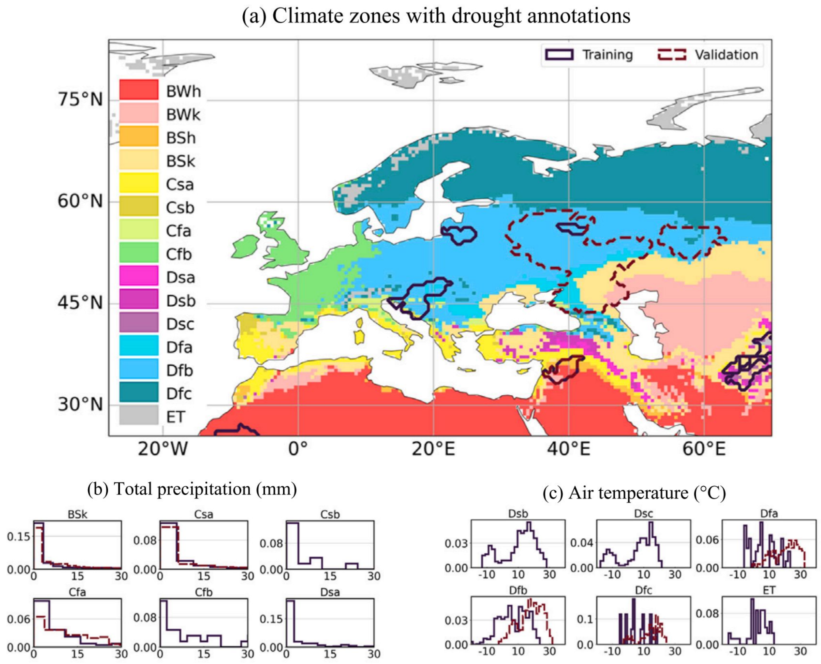


Calibration for DL-based DM

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.



Calibration for DL-based DM

- label shift ✅, likelihood shift ☒
- label shift ☒, likelihood shift ✅ $P_t(Y|X=x) = \dfrac{P_t(X=x|Y)P_t(Y)}{P_t(X=x)}$
- label shift ✅, likelihood shift ✅

Calibration for DL-based DM

- label shift ☑️, likelihood shift ☒

- Suppose that a classifier $\mathcal{K}$ is trained on set T_N which is unbalanced.
- Let $\mathbf{s}$ be a random variable associated to each sample $(x, y) \in T_N$, $\mathbf{s} = 1$ if the point is sampled and $\mathbf{s} = 0$ otherwise.
- Assume that $\mathbf{s}$ is independent of the input x given the class y (*class-dependent selection*):

$$p(\mathbf{s}|y, x) = p(\mathbf{s}|y) \Leftrightarrow p(x|y, \mathbf{s}) = p(x|y)$$

.

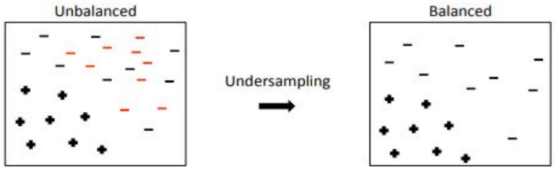


Figure : Undersampling: remove randomly majority class examples. In red samples that are removed from the unbalanced dataset ($\mathbf{s} = 0$).

Calibration for DL-based DM

- label shift ☑️, likelihood shift ☒

Let $p_s = p(+|x, s = 1)$ and $p = p(+|x)$. We can write p_s as a function of p [1]:

$$p_s = \frac{p}{p + \beta(1 - p)} \tag{1}$$

where $\beta = p(s = 1|-)$. Using (1) we can obtain an expression of p as a function of p_s:

$$\boxed{p = \frac{\beta p_s}{\beta p_s - p_s + 1}} \tag{2}$$

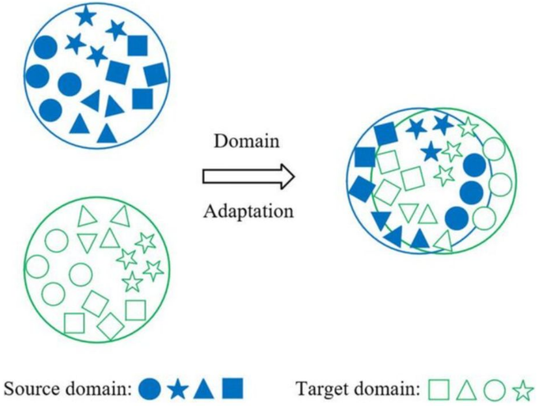
Calibration for DL-based DM

- label shift ☒, likelihood shift ☑️

Reasons
(1) Sensor degradation/ different measurement devices

(2) Synthetic vs Real-World Data/Change in Feature Encoding or Preprocessing

Example:
your model is trained on MODIS data. Later, you apply the same model test on Sentinel-3 data



Calibration for DL-based DM

- label shift ☒, likelihood shift ☑️

$$P_t(Y|X=x) = \frac{P_t(X=x|Y)P_t(Y)}{P_t(X=x)}$$

covariate shift

Covariate shift assumes that the distribution P(x) changes over time, while the function P (y|x) stays the same.
They could be either stationary or non-stationary.

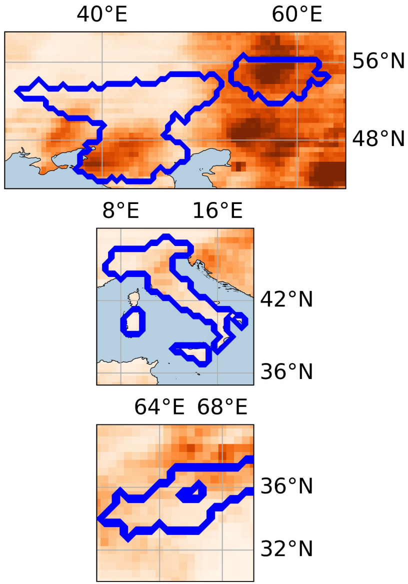
⊗

concept shift

Concept shift refers to a change of P (y|x). This means that the same x (features) may lead to different y (labels) in different situations. Harder to deal with.
They can be either stationary or non-stationary.

Calibration for DL-based DM

- label shift ☑️, likelihood shift ☑️



The **non-random undersampling** <u>causes</u> the label prior shift

Motivation of the climatic calibration
(1) climate zones with varying geographical patterns lead to distinct label priors;

(2) undersampling in the training subset has little impact on the likelihood.

Calibration for DL-based DM

- label shift ☑️, likelihood shift ☑️

subset has little impact on the likelihood. For each climate zone K_i, with $i \in \{1, \dots, N\}$, we aim to alleviate the label prior shift caused by imbalanced classes of spatial–temporal occurrences:

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

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

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).

Algorithm [edit]

Platt scaling is an algorithm to solve the aforementioned problem. It produces probability estimates

$$\mathrm{P}(y = 1 | x) = \frac{1}{1 + \exp(Af(x) + B)},$$

i.e., a logistic transformation of the classifier output $f(x)$, where A and B are two scalar parameters that are learned by the algorithm. After scaling, values

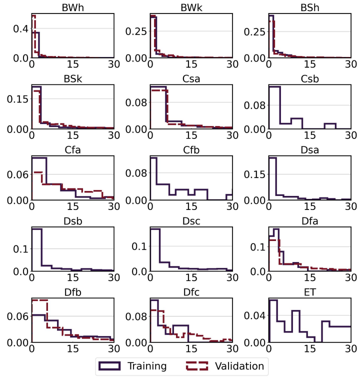
Calibration for DL-based DM

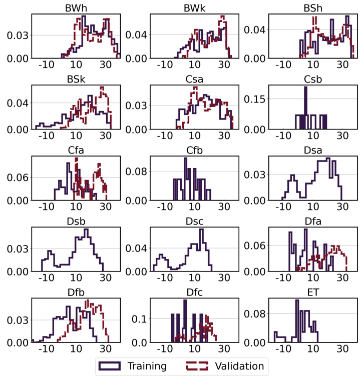
- label shift ☑️, likelihood shift ☑️

The likelihood shift is the primary driver of the discrepancy **between training and validation**

Motivation of the bias-corrected function
(1) assume it is one-parameter fractional transformation (stretch transformation)

(2) it may avoid overfitting on the training subset





Calibration for DL-based DM

- label shift ☑️, likelihood shift ☑️

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:

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

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

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

Evaluation metrics

$$\boxed{\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N_t} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|,} \tag{S.1}$$

$$\mathrm{RMSCE} = \sqrt{\sum_{m=1}^{M} \frac{|B_m|}{N_t} (\mathrm{acc}(B_m) - \mathrm{conf}(B_m))^2}, \tag{S.2}$$

where M is the number of bins, N_t is the total number of samples, B_m is the set of samples whose prediction confidence falls into the interval $I_m = (\frac{m-1}{M}, \frac{m}{M}]$, $\mathrm{acc}(B_m)$ is the accuracy within B_m, and $\mathrm{conf}(B_m)$ is the average confidence within B_m.

Let $p_s = p(+|x, s = 1)$ and $p = p(+|x)$. We can write p_s as a function of p [1]:

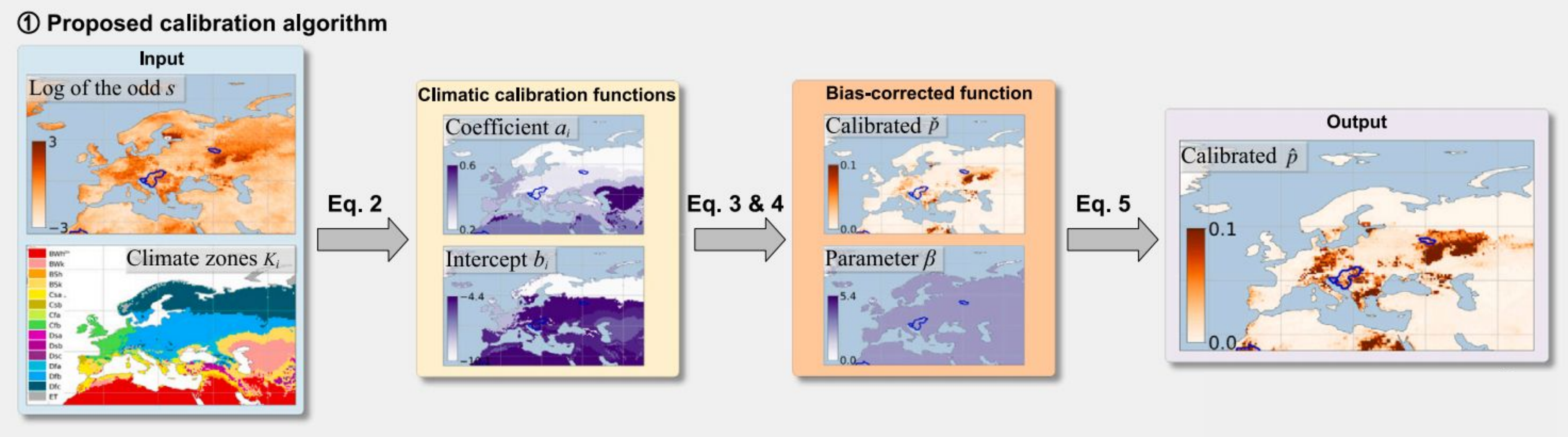
$$p_s = \frac{p}{p + \beta(1-p)} \tag{1}$$

where $\beta = p(s = 1|-)$. Using (1) we can obtain an expression of p as a function of p_s:

$$\boxed{p = \frac{\beta p_s}{\beta p_s - p_s + 1}} \tag{2}$$

Calibration for DL-based DM

- label shift ☑️, likelihood shift ☑️



Results

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

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.

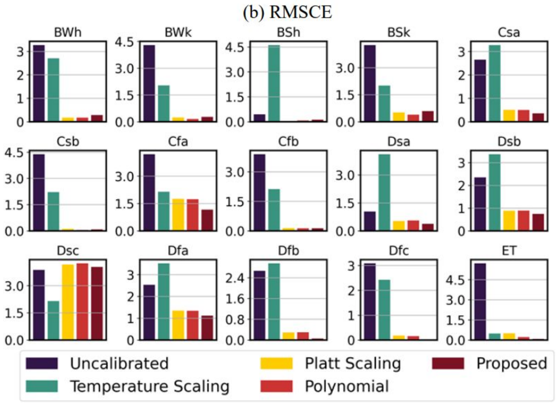


Fig S.1. Calibration errors (%) for the DK-VRN model in the test subset, given the fifteen climate zones considered in this study.

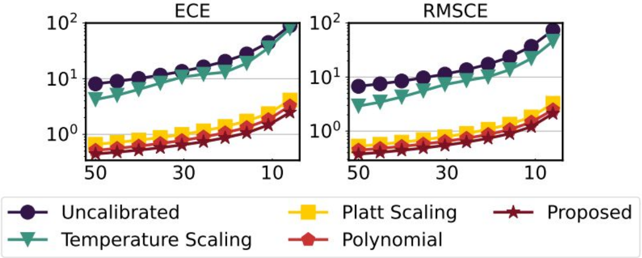


Fig S.2. Calibration errors (%) for the DK-VRN model in the test subset, given the top-predictions, from 50% to 10%.

Results

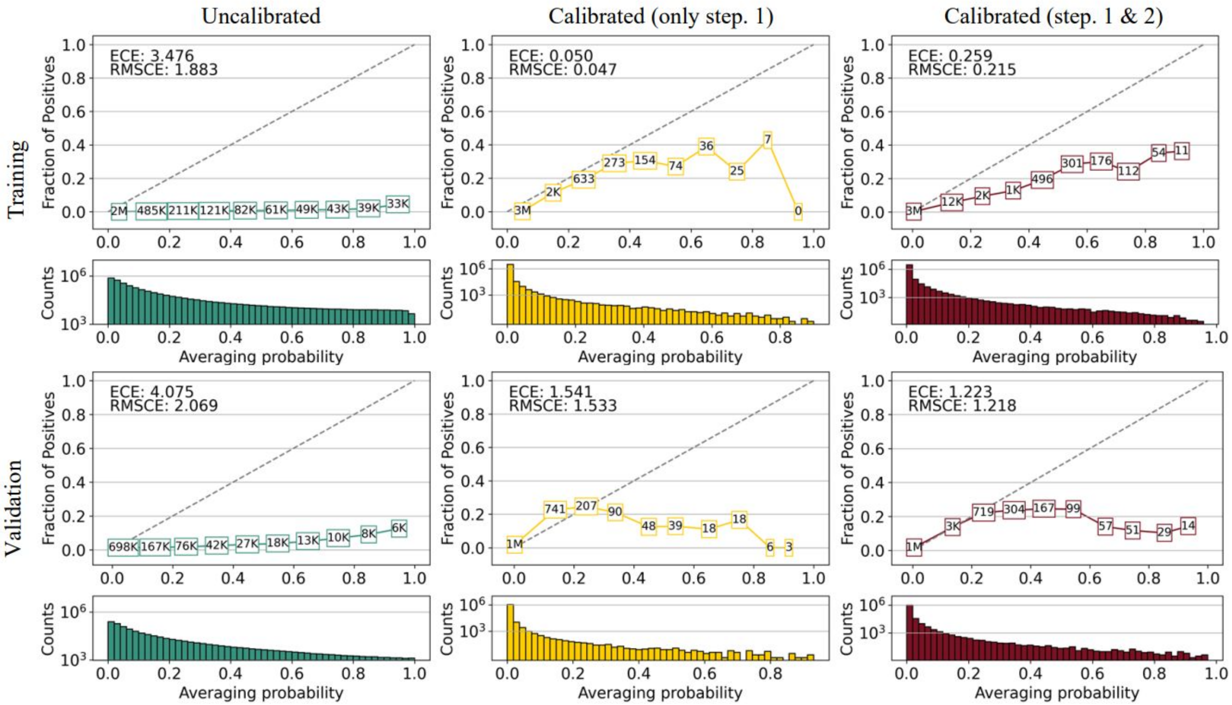


Fig S.3. Reliability diagrams of probabilities provided by the DK-VRN model for the training and validation subsets. Step 1 (climatic calibration functions) and step 2 (bias-corrected function) are described in Section 3.1. ECE and RMSCE are also provided.

Part III - UQ for DM

UQ for DM

Data

- Meteorological stations;
- Process simulations;
- Satellite-derived;
- Reanalysis;
- Reports;
- …

UQ for DM

Data

As ET cannot be directly measured by sensors from space, retrieval algorithms or models are needed to estimate ET from other variables observed by RS. These models estimate ET from visible and/or thermal infrared RS data and include now-well-known models. The diversity of models, input RS data sources, and processing techniques results in a wide range of RS-based ET estimates. **The uncertainty of RS-ET** data products is strongly linked to **model uncertainty** (from model conceptualization and parameters) and **input data uncertainty**.

Tran, Bich Ngoc, et al. "Uncertainty assessment of satellite remote-sensing-based evapotranspiration estimates: a systematic review of methods and gaps." *Hydrol. Earth Syst. Sci*. 2023

The **uncertainty estimation** for **ERA5** is obtained from the Ensemble of Data Assimilations (EDA) system. The EDA addresses some uncertainties of the model and data assimilation system, but not everything. The EDA accounts for uncertainties in **observations**, **sea surface temperature** (SST) and **model physical parametrizations**. Other uncertainties are not accounted for, such as uncertainties in radiative forcing due to greenhouse gases, or systematic errors or the way in which observations are used.

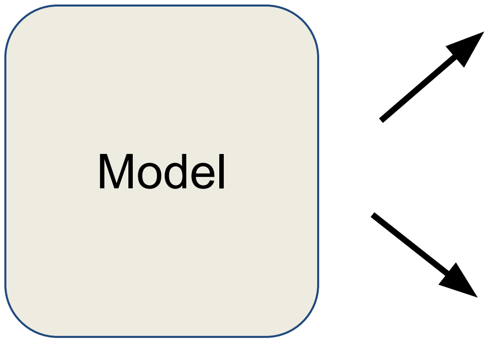
https://confluence.ecmwf.int/display/CKB/ERA5%3A+uncertainty+estimation

UQ for DM

Model

- GCMs/RCMs
- LSMs/HMs
- Impact models
- Drought indices
- ML/DL models
- …

UQ for DM



The projections of changes in the magnitude of meteorological (SPI) and soil moisture (SMA) drought in **CMIP5 display large spreads over all time frames**. A separation of different sources of uncertainty in projections of meteorological and soil moisture drought reveals that for the near term, internal climate variability is the dominant source, while **the formulation of GCMs** generally becomes the dominant source of spread by the end of the 21st century, especially for soil moisture drought. In comparison, the uncertainty from Greenhouse Gas concentrations scenarios is negligible for most regions.

Orlowsky, Boris. "Elusive drought: uncertainty in short-and long-term CMIP5 projections." *Hydrol. Earth Syst. Sci.* 2013.

Numerous sources of uncertainty in **LSMs** have ratified **data assimilation (DA) as a promising procedure to account for inaccuracies (or errors) linked to the model output and the observation data**. However, a crucial task in diagnostic model evaluation is the ability to identify and quantify model uncertainties which are specific to: state variables, parameters, input forcing data, and spatial variations in landscape properties. The model parameters, initial states, and input forcing data have direct impact on the internal dynamics of the model and its simulated output.

Dumedah, Gift, and Jeffrey P. Walker. "Assessment of land surface model uncertainty: A crucial step towards the identification of model weaknesses." *J. Hydrol*. 2014.

UQ for DM

Model

Uncertainty of HMs refers to the potential variability of the results **due to different model structures, input parameters, and forcing variables**. Hence, when models are employed to make forecasts or as a tool for decision-making processes, it is important to quantify and, if possible, reduce the uncertainty of their results. This can be accomplished by:(a) collecting more data, (b) developing more robust models, and (c) using effective techniques to extract and assimilate information from the collected data.

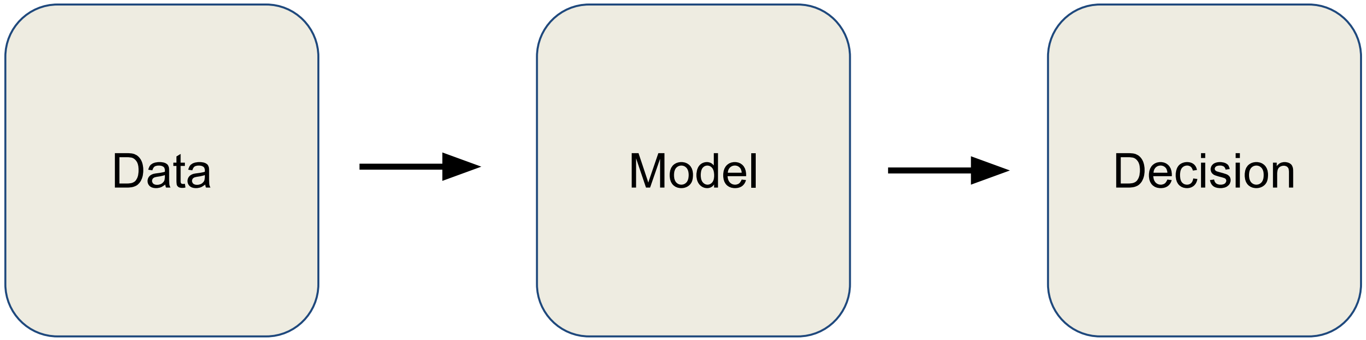
Herrera, Paulo A., Miguel Angel Marazuela, and Thilo Hofmann. "Parameter estimation and uncertainty analysis in hydrological modeling." *Wiley Interdiscip. Rev.: Water*. 2022.

The previously reported increase in global drought is overestimated because the **PDSI** uses a **simplified model of potential evaporation** that responds **only to changes in temperature** and thus responds incorrectly to global warming in recent decades. This can **lead to overestimation of past changes** and conversely **underestimation of recent trends** in the context of the past. Of concern is if the perceived influence of warming on drought as quantified by empirical approaches is extrapolated into the future and predictions of the **impacts of climate change are likely to be overestimated**.

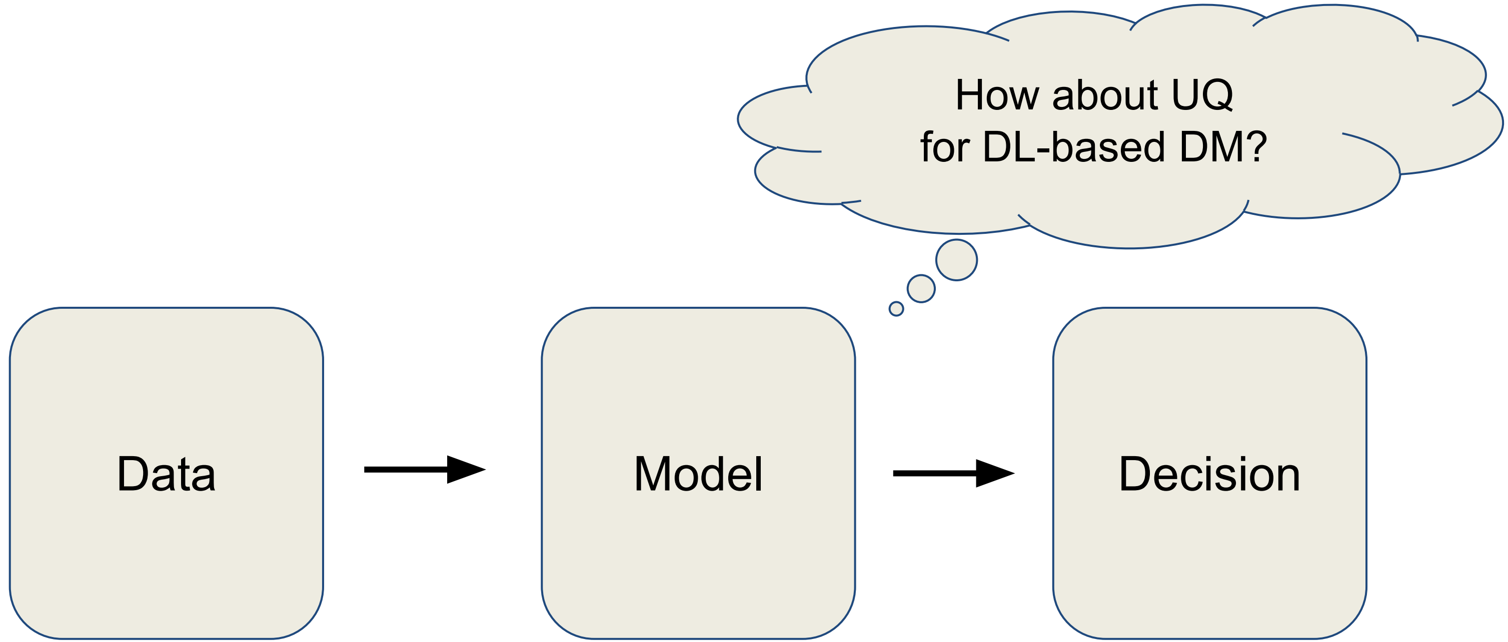
Sheffield, Justin, Eric F. Wood, and Michael L. Roderick. "Little change in global drought over the past 60 years." *Nat*. 2012.

UQ for DM

causes: data variability; intrinsic process/model variability
solutions: data collection; robust model; data preprocessing/assimilation

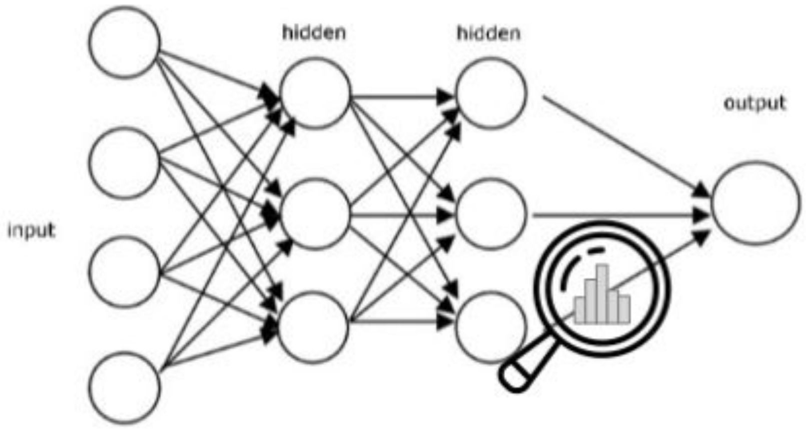


UQ for DL-based DM



UQ for DL-based DM

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.



UQ for DL-based DM

Motivation

(1) Entropy is beneficial when the predictive posterior distribution is Bernoulli or categorical. Entropy is also not sensitive to the scales.

(2) Existing frameworks address only total model uncertainty or rely on specialized Bayesian approaches.

(3) A tractable uncertainty framework that accounts for various uncertainty sources is needed for robust evaluations and comparisons, enabling actionable climate decision-making.

Distribution A (Low variance, concentrated near center)

Support: $\{-1, 0, 1\}$

$$P_A = \begin{cases} P(X=-1) = 0.25 \\ P(X=0) = 0.5 \\ P(X=1) = 0.25 \end{cases}$$

Distribution B (High variance, values farther apart)

Support: $\{-10, 0, 10\}$

$$P_B = \begin{cases} P(Y=-10) = 0.25 \\ P(Y=0) = 0.5 \\ P(Y=10) = 0.25 \end{cases}$$

For both distributions:

$$H = -[0.25\log_2 0.25 + 0.5\log_2 0.5 + 0.25\log_2 0.25] = 1.5 \text{ bits}$$

Variance depends on both the probabilities **and the values**:

- **Distribution A**:

$$\mu = 0, \quad \mathrm{Var}_A = 0.25(-1)^2 + 0.5(0)^2 + 0.25(1)^2 = 0.5$$

- **Distribution B**:

$$\mu = 0, \quad \mathrm{Var}_B = 0.25(-10)^2 + 0.5(0)^2 + 0.25(10)^2 = 50$$

UQ for DL-based DM

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.

UQ for DL-based DM

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.

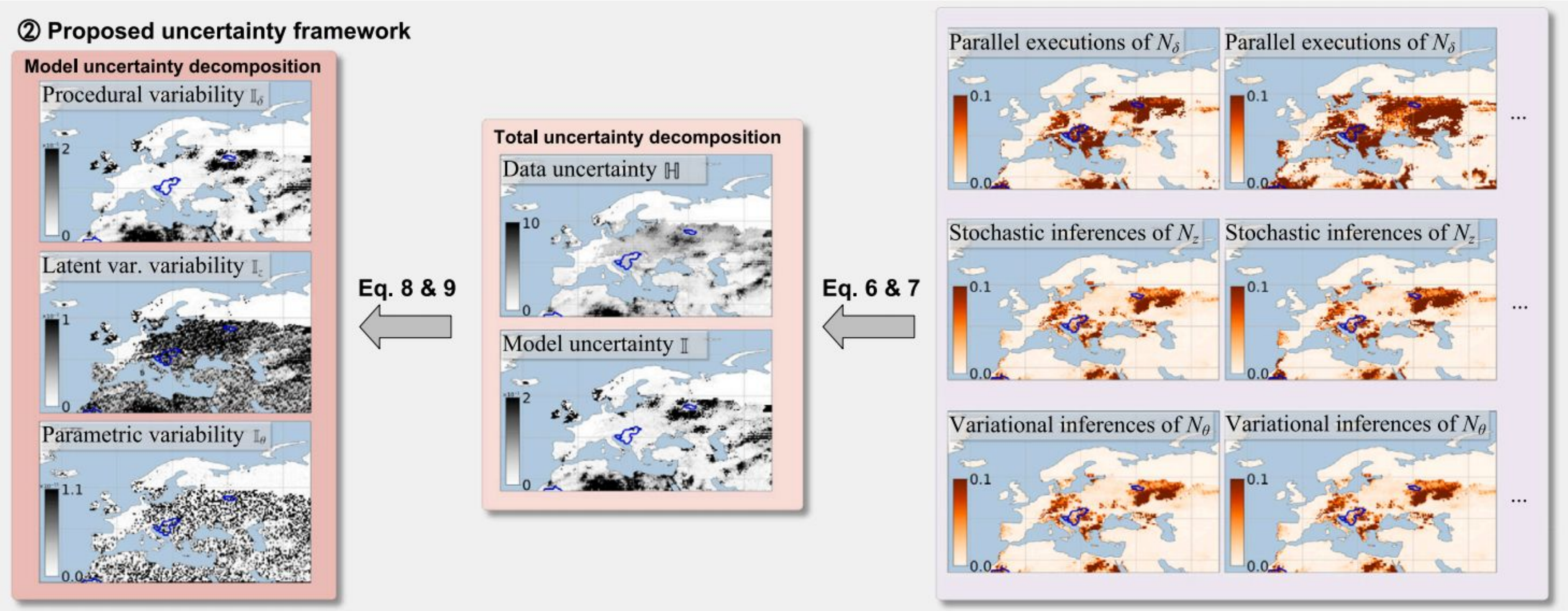
UQ for DL-based DM

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

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:

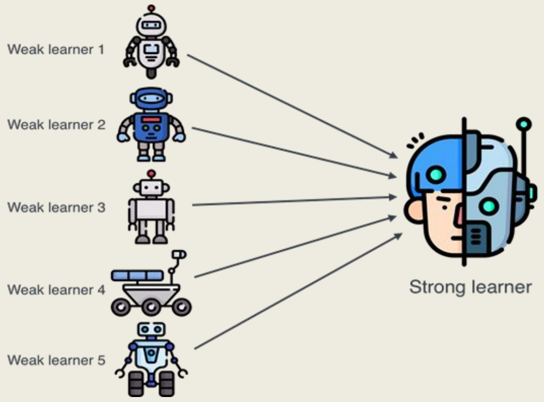
$$\begin{aligned}\mathbb{I}_z &= \mathbb{H}(Y) - \mathbb{H}(Y|z), \\ \mathbb{I}_\theta &= \mathbb{H}(Y) - \mathbb{H}(Y|(\theta,z)) - \mathbb{I}_z, \\ \mathbb{I}_\delta &= \mathbb{H}(Y) - \mathbb{H}(Y|(\delta,\theta,z)) - \mathbb{I}_z - \mathbb{I}_\theta.\end{aligned} \tag{9}$$

UQ for DL-based DM

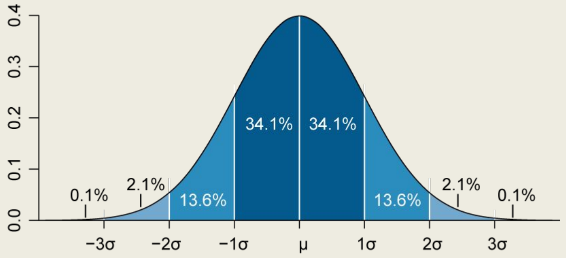


Decision-making for DL-based DM

Ensemble pred.



Rejecting pred.



Results

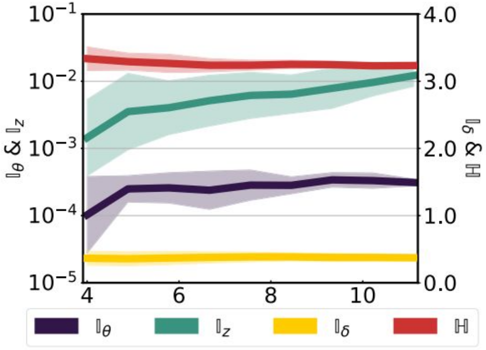


Fig. 3. Impact of the sampling size N (x-axis) on the proposed uncertainty decomposition, given the test subset (%) and the DK-VRN model.

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

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).

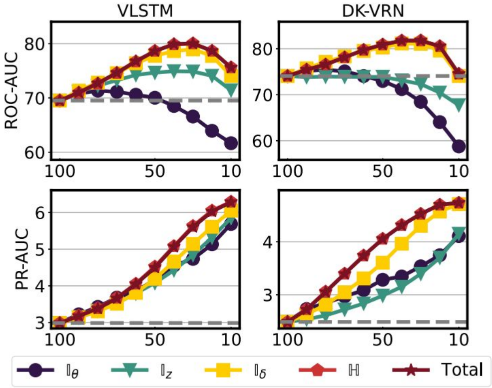


Fig. 4. Accuracy–rejection curves (%) for the VLSTM and DK-VRN models, given uncertain negative samples in the test subset. The x-axis denotes the keeping rate from 100% to 10%. For example, 80% keeping rate here represents rejecting the top 20% negative samples with higher uncertainty. The y-axis denotes 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).

Results

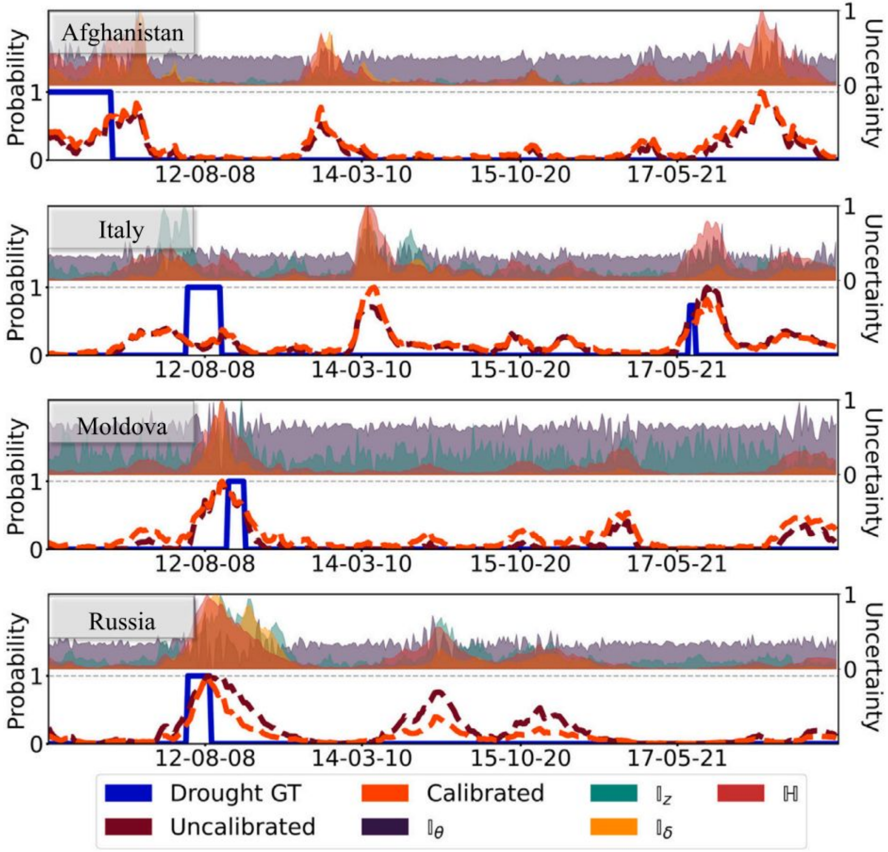


Fig. 5. Spatially aggregated signals of drought probabilities and uncertainties for the DK-VRN model, given the period 2011–2018 (test subset).

Results

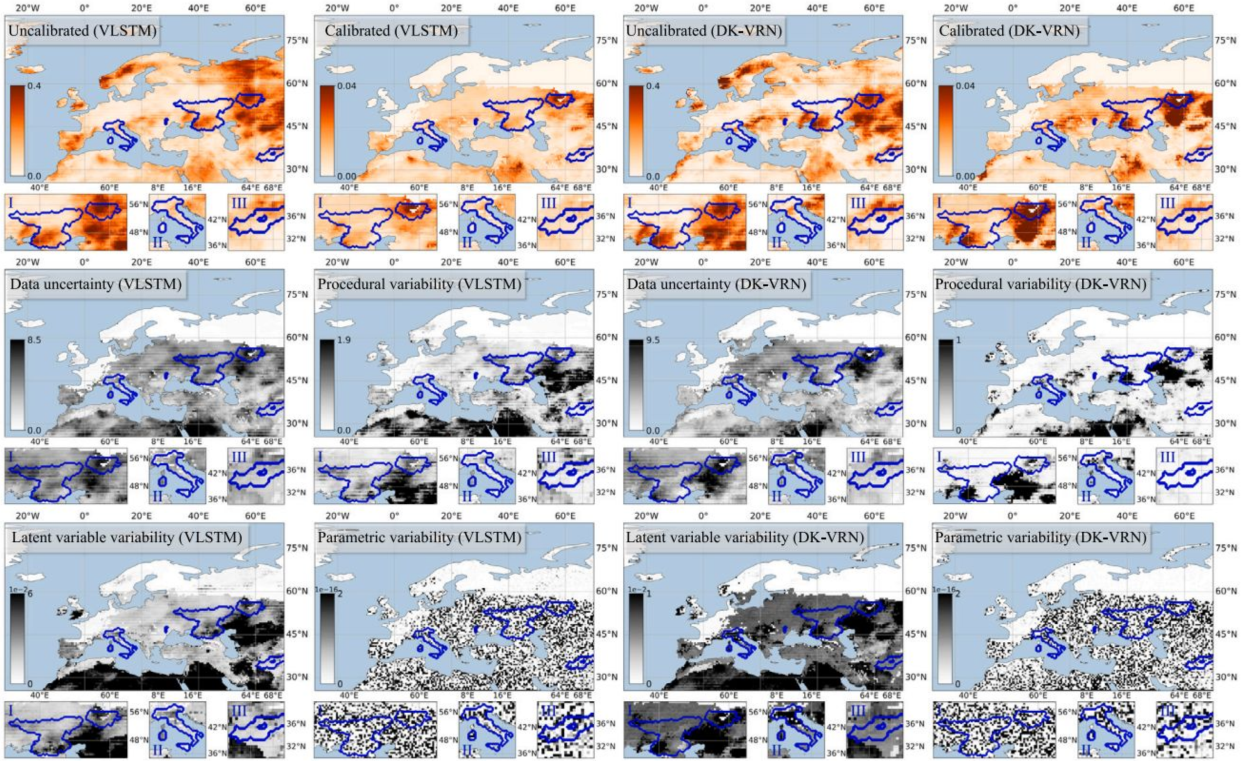


Fig. 6. Temporally aggregated (averaging all time steps) drought probability and uncertainty maps for VLSTM and DK-VRN models, given the period 2011–2018 (test subset). Three drought events occurred in this period: I (Russia), II (Italy), and III (Afghanistan). The ground truth is highlighted with blue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

Part IV - Conclusion

Take-home messages

- Calibration and UQ for DL-based DM;
- Bridge the gap between DL and trustworthy DM;
- Benefit decision-making for DL-based DM.

ISP · Image & Signal Processing
Universitat de València

http://isp.uv.es

@mengxue_rs

mengxue@alumni.uv.es