

Domain knowledge-driven variational recurrent networks for drought monitoring

Mengxue Zhang^{*}, Miguel-Ángel Fernández-Torres, Gustau Camps-Valls

Image Processing Laboratory (IPL), Universitat de València, Paterna, València 46980, Spain

ARTICLE INFO

Edited by Marie Weiss

Keywords:

Drought monitoring
Hybrid modelling
Domain knowledge
Long short-term memory (LSTM)
Variational encoder-decoder (VED)

ABSTRACT

In the context of climate change, droughts, increasingly frequent and severe, necessitate effective monitoring. Existing methods, such as drought indices and data-driven models, face important limitations. Drought indices are built on prior expert knowledge but lack calibration based on actual drought events, while data-driven models prioritize goodness of fit over real event identification, undermining their credibility and generalization, and also struggling to generalize from regional to large-scale contexts. To address these challenges, here we introduce a hybrid machine learning framework for time series that combines domain knowledge and observational data in a variational recurrent neural network. The network models the joint distribution of total precipitation, air temperature, and real drought events, providing accurate predictions and uncertainty estimates. Extensive experiments focusing on a wide range of European drought events from 2011 to 2018 consistently show that our hybrid model surpasses both drought indices and data-driven models in terms of accuracy in drought detection, underlining its effectiveness, robustness, and stability. Our model achieves the best ROC-AUC (%) results in Afghanistan (79.7 ± 0.5), Italy (84.3 ± 0.6), Russia (89.4 ± 0.2), Europe-0 (84.3 ± 0.1), and Europe-1 (82.8 ± 0.4), effectively capturing the starting and ending times of drought events with lower uncertainty, and also generalizing better for unseen locations.

1. Introduction

The Intergovernmental Panel on Climate Change (IPCC) defines drought as an extended period of arid conditions leading to significant hydrological imbalances (IPCC, 2019). Yet, this definition remains adaptable due to the multifaceted nature of drought, with impacts influenced by social, economic, and environmental factors (Caretta et al., 2022). Effectively, droughts are one of the costliest natural disasters, causing seriously destructive consequences on the ecological environment, agricultural production, and socio-economic impacts (Mishra and Singh, 2010). With global warming, the frequency and intensity of composite dry-heat events, such as heatwaves and droughts, have increased significantly (Vicente-Serrano et al., 2020). In these circumstances, advanced and accurate monitoring techniques, ultimately integrated into early warning systems, are urgently required to efficiently and rapidly detect and response drought events (Gitz et al., 2016), as well as to improve risk management (Kreibich et al., 2022).

Due to the complex physical, chemical, and biological processes involved in drought events, much effort has been devoted to developing

climate indices and (primarily mechanistic) analytical tools. In general, drought indices have become standardized tools for objectively quantifying the characteristics of these episodes in terms of intensity, magnitude, duration, and spatial extent. Three widely-used indices are the Palmer Drought Severity Index (PDSI), which is based on a soil water balance equation (Palmer, 1965); the Standardized Precipitation Index (SPI), which is grounded in a precipitation probabilistic model (McKee et al., 1993); and the Standardized Precipitation-Evapotranspiration Index (SPEI), which combines both approaches (Vicente-Serrano et al., 2010). The run theory bridges climate indices and drought characteristics analysis (Yevjevich, 1967; Panu and Sharma, 2002). Other indices based on comprehensive remote sensing data have been proposed, either related to vegetation greenness, such as the Normalized Difference Vegetation Index (NDVI) and the fraction of absorbed photosynthetically active radiation (FAPAR) (Fensholt et al., 2004), or water-limitation principles, such as the Precipitation Condition Index (PCI) (Hazaymeh and Hassan, 2016) and the Standardized Moisture Anomaly Index (SZI) (Zhang et al., 2019). Furthermore, integrated indices combining multi-sensor remote sensing data present new opportunities

^{*} Corresponding author.

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

<https://doi.org/10.1016/j.rse.2024.114252>

Received 24 October 2023; Received in revised form 30 May 2024; Accepted 4 June 2024

Available online 22 June 2024

0034-4257/© 2024 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC license (<http://creativecommons.org/licenses/by-nc/4.0/>).

for composite drought monitoring (West et al., 2019; Jiao et al., 2021).

Alternatively, data-driven models directly learn from historical drought severity (Guha-Sapir et al., 2015) and take Essential Climate Variables (ECVs) or indices as inputs to estimate drought intensity or drought-relevant variables (Prodhan et al., 2022). Existing models include, but are not limited to, autoregressive integrated moving average (Mishra and Desai, 2005), k -nearest neighbours (Fadaei-Kermani et al., 2017), artificial neural networks (Belayneh et al., 2016), fuzzy logic based models (Agboola et al., 2013), Bayesian networks (Madadgar and Moradkhani, 2014), support vector regression (Belayneh et al., 2016), and random forests (Feng et al., 2019). More recently, deep architectures have been successfully employed, such as deep feed-forward neural networks (Shen et al., 2019), long short-term memory-based models (Dikshit et al., 2021), advanced convolutional neural networks (González-Calabuig et al., 2022), and deep generative models (Ravuri et al., 2021).

However, there are two main shortcomings among existing methods for drought monitoring. On the one hand, drought indices are known as top-down models, built on prior expert knowledge and implemented in the form of rigid mechanistic models, and lack of post-hoc adjustment based on recorded drought events (AghaKouchak et al., 2022). Moreover, effective drought indices are based on a monthly statistical aggregation, which is not applicable in drought monitoring with a finer temporal resolution (Wang et al., 2021). On the other hand, data-driven models pay more attention to the goodness of fit, and their evaluations focus on the estimation error of drought indices rather than real drought-relevant variables or events, consequently both neglecting the uncertainty and undervaluing the credibility which are required for drought analysis (Li et al., 2021; Kikon and Deka, 2022). Furthermore, data-driven models struggle with monitoring generalization from regional to continental scales.

Integrating data-driven and physical models has recently captured the attention and promised a new generation of models for the Earth Sciences (Reichstein et al., 2019). The hybrid framework combines the flexibility and representation capability of machine learning approaches with the consistency of domain knowledge, making it possible to provide more accurate, less uncertain, physically consistent, and credible outcomes, facilitating the understanding of the complex Earth system. The synthesis of physics and data-driven models may achieve improved performance and generalization and, more importantly, may lead to significant consistency and credibility of the machine learning models (Slater et al., 2022). Validated examples include precipitation forecasting (Zantedeschi et al., 2020), seasonal precipitation prediction (Madadgar et al., 2016), terrestrial evaporation (Abu-Raddad et al., 2022), cloud detection (Li et al., 2022), biophysical parameter estimation (Cortés-Andrés et al., 2022), soil moisture forecasting (Xu et al., 2023), and drought index prediction (Wu et al., 2022). The above-related studies based on hybrid modelling exhibit their potential for Earth Sciences.

To address the aforementioned challenges, here we introduce a hybrid machine learning framework that integrates domain knowledge while employing data-driven models to learn from observational data. The work extends our preliminary methodology in (Zhang et al., 2022), where we first attempted a hybrid recurrent network incorporating multi-scalar SPEI as input features and realized effective monitoring of regional drought events. This paper pursues a unified monitoring model for droughts on the continental scale across intensities, times, and locations. We extend the methodology and include many more use cases and analyses, contributing to the field of drought monitoring in several ways:

- First, we introduce a novel Domain Knowledge-driven Variational Recurrent Network (DK-VRN), capable of detecting droughts of varying intensities over diverse time-frames and across continental scales. DK-VRN models the joint distribution of the ECVs and the drought probability, also providing uncertainty estimates.

- DK-VRN receives Essential Climate Variables (ECVs) as input and is end-to-end supervised by historical drought events rather than drought indices. The multi-scalar SPEI is instead incorporated into DK-VRN as a domain knowledge prior, which is beneficial for the model's generalization, reliability, and stability.
- Lastly, evaluations over a wide range of drought events in Europe between 2011 and 2018, consistently labelled by expert knowledge, are conducted to verify the effectiveness of the hybrid framework. DK-VRN outperforms existing models, demonstrating higher accuracy and increased stability, thus establishing it as a trustworthy tool for drought monitoring and promising advances for early warning systems.

The remainder of the paper is organized as follows. After introducing the datasets used in Section 2, we present our methodology in Section 3 and then analyze the experimental results in Section 4. A discussion about the modelling approach, advantages, challenges, and implications follows in Section 5. We conclude the paper in Section 6 with some final remarks and future research directions.

2. Data collection

2.1. Climate variables and drought indices

Generally, when rainy days are missing, moisture depletion occurs from the upper soil layers, and a reduction of relative humidity happens eventually in the atmosphere, establishing the drought conditions (Bravar and Kavvas, 1991). Besides, high temperatures can exacerbate drought conditions by increasing evaporation and intensifying the impact of water deficits (Sheffield and Wood, 2008). Total precipitation and air temperature at 2 m are accordingly selected as the predictors (see Fig. 1), which are well-known ECVs correlated to droughts, from the re-gridded ERA5-Land data¹ ($0.083^\circ \times 0.083^\circ$, 8km^2), covering the period from Jan 2003 to Dec 2018 with a temporal resolution of 8 days, across an extended European area (Mahecha et al., 2020).

SPEIs computed at different scales (1, 3, and 6 months) are selected as the representative domain knowledge. Multi-scalar SPEI measures drought severity by considering both intensity and duration, allowing us to identify their corresponding onset and offset times. Besides, based on a water balance equation and a precipitation probabilistic model, the explicit equation-based index represents domain knowledge for drought monitoring:

$$D = P - \text{PET}, \quad (1)$$

where D denotes the measure of the water surplus or deficit for the analyzed month, P represents precipitation, and PET stands for potential evapotranspiration. The index estimates D by fitting a three-parameter log-logistic distribution function as follows:

$$f(D) = \frac{\beta}{\alpha} \left(\frac{D - \gamma}{\alpha} \right)^{\beta-1} \left[1 + \left(\frac{D - \gamma}{\alpha} \right)^\beta \right]^{-2}, \quad (2)$$

where α , β , and γ are scale, shape, and origin parameters, respectively. Last but not least, it is a multi-scalar index that identifies different drought types. It should be noted that for measuring the implicit expert knowledge rather than new ECVs gathered in the formulation of the index, PET for the multi-scalar SPEI is computed using an equation that only relies on the ECVs mentioned above (Thornthwaite, 1948):

$$\text{PET} = 16K \left(\frac{10T}{I} \right)^m, \quad (3)$$

where T is the monthly-mean temperature, I is a heat index, which is

¹ <https://cablab.readthedocs.io/>

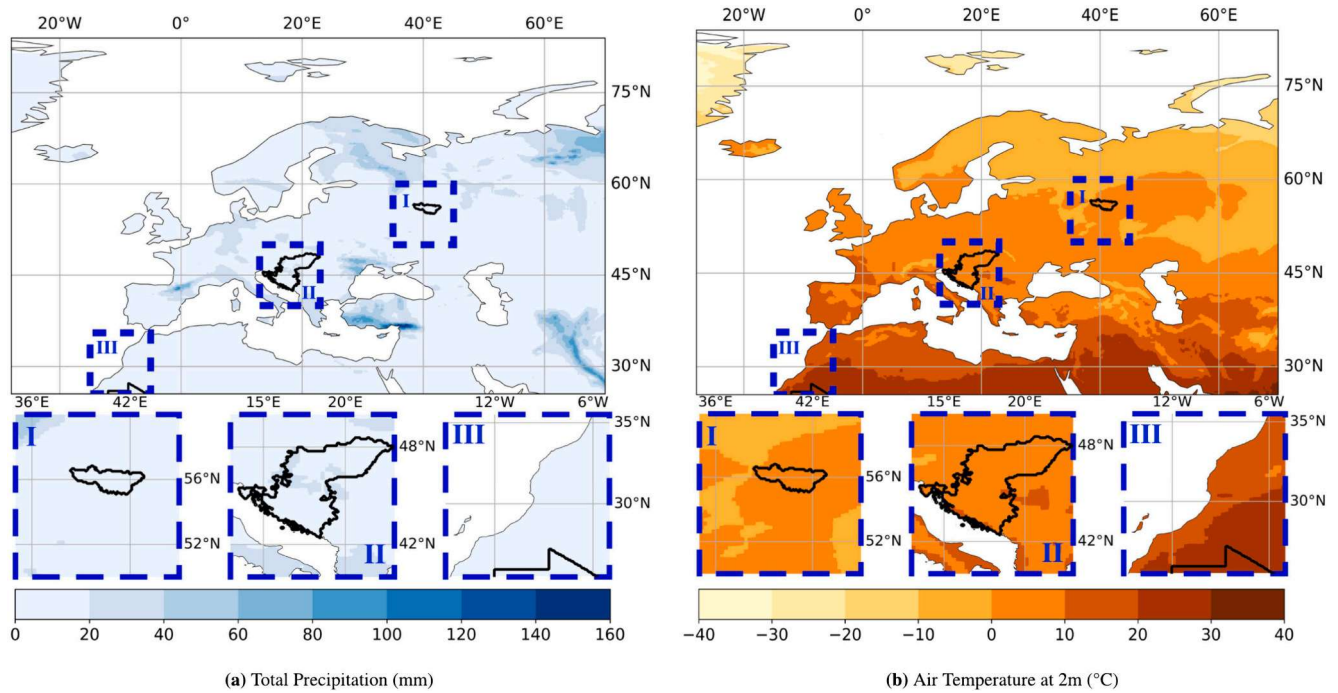


Fig. 1. ECVs and target drought events for an example time step. Maps of Total Precipitation and Air Temperature at 2 m on 2003-04-03 are depicted in (a) and (b), respectively. The three accompanying drought events on 2003-04-03 are marked with black lines: I (Russia), II (Bosnia; Croatia; Hungary), and III (Mauritania). Regional maps of drought events are highlighted with blue dashed lines, zooming in the ECVs for each drought location. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

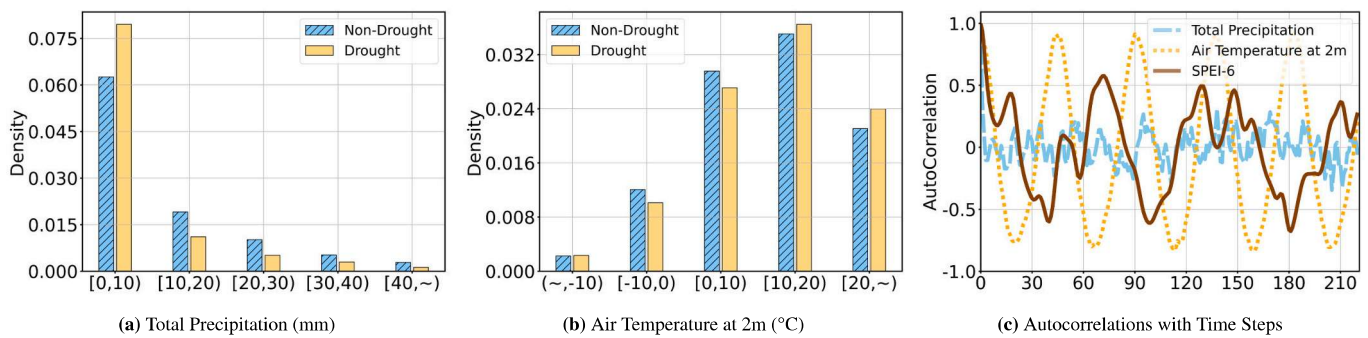


Fig. 2. Statistical analysis of the input data. Histogram of (a) Total Precipitation and (b) Air Temperature at 2 m, together with the (c) time-lagged Spearman autocorrelations for these ECVs and SPEI-6. The latter suggests the complementary temporal information between ECVs and SPEI for drought detection.

calculated as the sum of 12 monthly index values, m is a coefficient depending on I , and K is a correction coefficient computed as a function of the latitude and month. However, more accurate methods based on additional variables are available for this estimation. The comparisons between different estimation methods of PET for SPEI are investigated at (Stagge et al., 2014). In this work, ECVs data available between 1986 and 2008 is used for SPEI computation (the period between 2009 and 2018 is left for validation and test purposes), because SPEI generally relies on stable statistics of long-term reference data, and the three-parameter Pearson Type III distribution is chosen for its calculation.²

An statistical analysis of the input training data is depicted in Fig. 2. As can be seen, ECVs present distinct histogram shapes for drought and non-drought. For example, droughts are frequently accompanied by high air temperatures and low precipitation. Besides, different time-lagged Spearman autocorrelations given ECVs and SPEI suggest their complementary temporal information for drought detection.

2.2. Drought annotations

The EM-DAT³ is a global natural disaster database recording the occurrence and effects of more than 21,000 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. The criteria guarantee that all recorded disaster events are objective and influential (Guha-Sapir et al., 2015). A time series of the European drought events from Jan 2003 to Dec 2018 is considered for the experiments in this paper. To this end, we use the Geocoded Disasters (GDIS) database (Rosvold and Buhaug, 2021), an extension of EM-DAT where the drought event data was converted from tabular data to the raster data structure.

² https://github.com/monocongo/climate_indices

³ <https://public.emdat.be/>

2.3. Data preprocessing

All ECVs are denoted as x_0 and are preprocessed based on the computation of monthly anomalies, as well as rescaled to $[-1, 1]$.

$$\tilde{x}_{0m,n,t} = \min \left[\max \left[\frac{x_{0m,n,t} - \mu_{m,n}^i}{3\sigma_{m,n}^i}, -1 \right], 1 \right], \quad (4)$$

where m, n, t denote the latitude, longitude, and time coordinates, respectively; and μ^i, σ^i denote the monthly mean and standard deviation for the month i , respectively.

SPEI constitutes an indicator of the water deficit. The lower the SPEI value, the higher the water deficit and, therefore, the higher the drought probability. Thus, we employ the complement of the SPEI values, denoted as x_1 , to measure their corresponding drought probabilities, which are given as follows:

$$\tilde{x}_1 = 1 - F^{-1}(x_1), \quad (5)$$

where F^{-1} denotes the inverse of the cumulative distribution function. SPEI values are then temporally resampled to the 8-day scale by setting their values to the one of the nearest, not reached month.

As for the EM-DAT, during the study period, 14 drought events belonging to 11 regions are involved (constrained by the resolutions of ECVs, drought events with a scale smaller than 1° are ignored). An exemplified illustration of the ECVs and the drought events on 2003-04-03 is visualized in Fig. 1, Fig. A.1, and Fig. A.2, respectively, where the geospatial coordinates (NSWE) are 59.0°N , 43.0°S , 30.0°W , and 64.0°E , respectively.

We then split the data into the training, validation, and testing data with different periods. We make sure that each subset has no overlapping times between them. 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. In this context, monitoring models must be applied to various unobserved geospatial regions, e.g., Italy and Moldova (i.e., no drought events in these places are neither in the training nor the validation subsets). The details of the drought events considered in this study are described in Table 1.

To mitigate the imbalanced problem given the drought events, which constitute anomalies and therefore a very small percentage of the total number of samples in the database, we adopt an undersampling strategy for the training samples: only European geographical locations where at least a drought event happened between 2003–2018 are selected for training purposes, being discarded those where droughts never occurred. The label distributions or ratios before and after this strategy are reported in Table 1. Ratio (all) denotes the percentage of the drought samples over all samples in the extended European region considered in

Table 1

Drought event subsets considered in this study. Events with * denote a single drought event across multiple countries. They are defined according to the EM-DAT database (see Table A.1 for a more complete spatial description of them).

	Training	Validation	Test
Afghanistan	2006.07–12	–	–
	2008.10–11	–	2011.01–08
Bosnia*	2003.02–05	–	–
Croatia*	2003.02–05	–	–
Hungary*	2003.02–05	–	–
Italy	–	–	2012.06–10
Lithuania	2006.08–12	–	–
Mauritania	2003.01–12	–	–
Moldova	–	–	2012.11–12
Russia	2003.01–12	2010.04–08	2012.06–08
Syria	2008.01–12	2009–2010	–
Tajikistan	2008.10–11	–	–
Period	2003–2008	2009–2010	2011–2018
Ratio (all)	0.13%	0.94%	0.15%
Ratio (happened)	9.19%	24.05%	3.80%

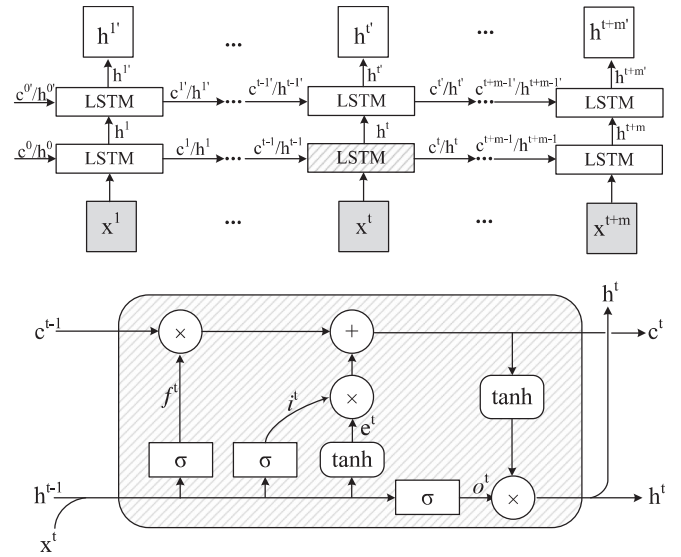


Fig. 3. Graphical illustration of a two-layer LSTM. The computations involved by the first LSTM layer (cell) at time step t are detailed at the bottom, where x denotes input data at the different time steps $1, \dots, t, \dots, t+m$.

this study, while Ratio (happened) denotes the percentage of the drought samples over the subset of European locations where at least a drought event happened during the corresponding training/validation/test period.

3. Methodology

3.1. Long short-term memory networks

Long Short-Term Memory-based (LSTM-based) models (Hochreiter and Schmidhuber, 1997) are recurrent neural networks which introduce self-loop gates to define pathways where the gradient can flow for long durations, better learning long-term dependencies than simple recurrent architectures (Goodfellow et al., 2016). LSTMs have been widely used in Earth science applications for the estimation of biophysical parameters (Camps-Valls et al., 2021), land cover classification (Campos-Taberner et al., 2023; Campos-Taberner et al., 2020) and crop yield forecasting (Mateo et al., 2023), to name a few.

For drought monitoring, the input variables of LSTMs are time-series of ECVs, and the output variables are time-series of hidden variables, which are followed by a hyperbolic tangent activation and a linear layer that maps the LSTM output at each time step to a single value: the drought probability.

A two-layer LSTM is illustrated in Fig. 3. At each time step, input x_t and the cell/hidden states (c_{t-1} and h_{t-1}) of the previous time step together contribute to the cell/hidden states of the current time step within each LSTM layer. Existing studies have verified the capability of LSTM-based models to learn dynamic temporal information from time-series drought-relevant data (Zhong et al., 2019; Yuan et al., 2020; Camps-Valls et al., 2021). However, traditional LSTMs are deterministic neural networks that do not support uncertainty estimation. Furthermore, LSTMs need to give consideration to domain knowledge, which may improve the generalization and stability for drought monitoring at continental scales.

3.2. Variational encoder-decoders

Variational Encoder-Decoders (VEDs) are probabilistic latent variable models proposed by (Kingma and Welling, 2013). Applications of VED to Earth Science include but are not limited to desertification detection (Zerrouki et al., 2020), extreme event detection (Růuvživčká

et al., 2022), and cyclone prediction (Kapoor et al., 2023). Generally, VED consists of a variational encoder that maps data to the approximated posterior distribution over latent variables and a decoder that generates data from the latent variables, being also able to provide uncertainty estimates. In this context, we introduce a factorized VED for modelling drought processes.

Let us assume that drought processes are driven by latent (unseen, unobserved) variables z , and the evidence of drought events can be characterized as the joint distribution of known variables, e.g., ECVs (x_0), and interested but unknown variables, e.g., the probability of drought events y . Under this assumption, the drought dynamics are given as:

$$\begin{aligned} P(y, x_0) &= \int P(y, x_0 | z) P(z) dz \\ &= \int P(y | x_0, z) P(x_0 | z) P(z) dz. \end{aligned} \quad (6)$$

The joint distribution (behavior) of drought-relevant variables can be regarded as a measure of the drought condition (Beersma and Buis-hand, 2004; Hao and Singh, 2015). Thus, the objective is to maximize the logarithm of the joint distribution $\log P(y, x_0)$ over D_0 , where D_0 denotes the training subset $\{y^i, x_0^i\}$ and $i = \{1, 2, \dots, N\}$, being N the total number of training samples.

With the factorized VED at hand, the desired drought probability and the uncertainty can be estimated via the likelihood distribution $P(y | x_0, z^{(k)})$ and the posterior distribution $Q(z^{(k)})$ when input variables x_0 are known.

However, the factorized VED is not enough for time-series drought monitoring, where two challenges arise: 1) How to learn the $Q(z)$ effectively and 2) How to implement high-performance encoder-decoder architectures for processing time-series drought-relevant variables. The following subsections introduce and describe the proposed DK-VRN to address these questions.

3.3. DK-VRN for drought monitoring

In theory, $Q(z)$ is in the form of $Q(z|x)$ ($Q(z|y, x_0)$ in our case), and approximates $P(z|y, x_0)$. However, y is typically imbalanced. Comparatively, drought indices x_1 built on prior expert knowledge are much easier to fetch, distributed normally, and are commonly used as alternatives when drought labels are noisy. Therefore, we resort to x_1 to help learn $Q(z)$.

We assume $P(z|x_1)$ and $P(z|y, x_0)$ are closer than $P(z)$ and $P(z|y, x_0)$. Intuitively, this assumption about $P(z|y, x_0)$ and $P(z)$ is feasible because of the existence of $P(x_1|z)$, which improves the closeness between two posterior distributions when indices x_1 are prior estimations to y . Then, $\mathcal{L}(x_1)$ is computed as:

$$\mathcal{L}(x_1) = -D_{KL}(Q(z|x_1) || P(z)) + \frac{1}{K} \sum_{k=1}^K \log P(x_1 | z^{(k)}). \quad (7)$$

The KL divergence loss term can be interpreted as regulating Q , encouraging $Q(z_1)$ becoming closer to the prior $P(z)$, approximating $P(z|x_1)$, and the $\log P(x_1 | z^{(k)})$ loss term can be interpreted as the reconstruction error of x_1 .

To shift domain knowledge from x_1 , we take $Q(z|x_1)$ in $\mathcal{L}(x_1)$ as the approximation to the intractable $Q(z|y, x_0)$ of $\mathcal{L}(y, x_0)$. The prior distribution $P(z)$ in $\mathcal{L}(y, x_0)$ is now replaced by the learnable posterior distribution $Q(z|x_1)$ in $\mathcal{L}(x_1)$, encouraging $Q(z|x_0)$ to be close to $Q(z|x_1)$, i.e., indirectly close to the true but intractable $P(z|y, x_0)$. Finally, $\mathcal{L}(y, x_0)$ is computed as:

$$\begin{aligned} \mathcal{L}(y, x_0) &= -\lambda D_{KL}(Q(z|x_0) || Q(z|x_1)) \\ &+ \frac{1}{K} \sum_{k=1}^K \log P(y | x_0, z^{(k)}) + \frac{1}{K} \sum_{k=1}^K \log P(x_0 | z^{(k)}), \end{aligned} \quad (8)$$

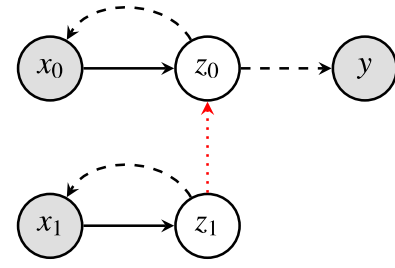


Fig. 4. Directed graphical model of the proposed DK-VRN. Shaded circles represent observations (x_0, x_1, y), white circles indicate hidden variables to be inferred (z_0, z_1). x_0 denotes the ECVs input data, x_1 denotes the SPEI input data, z_0 denotes the latent variables given x_0 , z_1 denotes the latent variables given x_1 , and y denotes the drought labels. The solid lines denote the inference process, the dashed lines denote the generative process, and the red dotted lines denote the domain-knowledge regularization process. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

where λ is added to control the strength of the KL divergence between the posterior distributions for the ECVs and the drought indices, i.e., as a way of providing the expert knowledge required as guidance for more accurate drought monitoring from observational data. $\mathcal{L}$ for the whole DK-VRN system is as follows:

$$\mathcal{L} = \mathcal{L}(y, x_0) + \mathcal{L}(x_1). \quad (9)$$

Finally, the directed graphical model of DK-VRN is shown in Fig. 4, where their dependencies are revealed.

3.4. Architecture design

The graphical illustration of the proposed DK-VRN for drought monitoring is shown in Fig. 5. As can be seen, LSTM layers are integrated into its different encoders and decoders. Specifically, each variational encoder Q is sequentially composed of two LSTM layers, one batch normalization layer, and two linear layers. In each encoder, x is transformed into the stochastic variables of the latent space, i.e., $Q_\theta(z_0|x_0)$ represents the distribution of z_0 given x_0 :

$$Q_\theta(z_0|x_0) = N(\mu_\theta(x_0), \sigma_\theta(x_0)), \quad (10)$$

where a multivariate Gaussian distribution is chosen, being its parameters μ and σ the outputs of two linear layers, and $\sigma(x)$ constrained to be diagonal. Then, by applying the reparameterization trick, which initiates ϵ_0 from $N(0, 1)$, and after the multiplication and summation operations, samples from the distribution $Q_\theta(z_0|x_0)$ can be obtained:

$$\begin{aligned} z_0^{(k)} &= g(x_0, \epsilon_0^{(k)}) = \mu_\theta(x_0) + \epsilon_0^{(k)} \sigma_\theta(x_0), \\ \text{where } z_0^{(k)} &\sim Q_\theta(z_0|x_0). \end{aligned} \quad (11)$$

Next, each decoder P is sequentially composed of one optional LSTM layer, one batch normalization layer, and one linear layer. In each decoder, the latent variables are transformed into the target variables, e.g., $P_\lambda(y|z_0)$ represents the likelihood given z_0 :

$$\begin{aligned} P_\lambda(y|z_0^{(k)}) &= \text{Bern}(f_\lambda(z_0^{(k)})), \\ \text{where } z_0^{(k)} &\sim Q_\theta(z_0|x_0), \end{aligned} \quad (12)$$

where a Bernoulli distribution is chosen, f_λ produces the parameter p for value 1 (i.e., the drought probability), and $1 - p$ for value 0, implemented as a sigmoid function.

3.5. Training and test stages

On the one hand, the training stage is illustrated in Fig. 6(a). First, values for drought indices x_1 given all training time steps constitute the input data to the encoder Q_x , which produces the distribution of latent

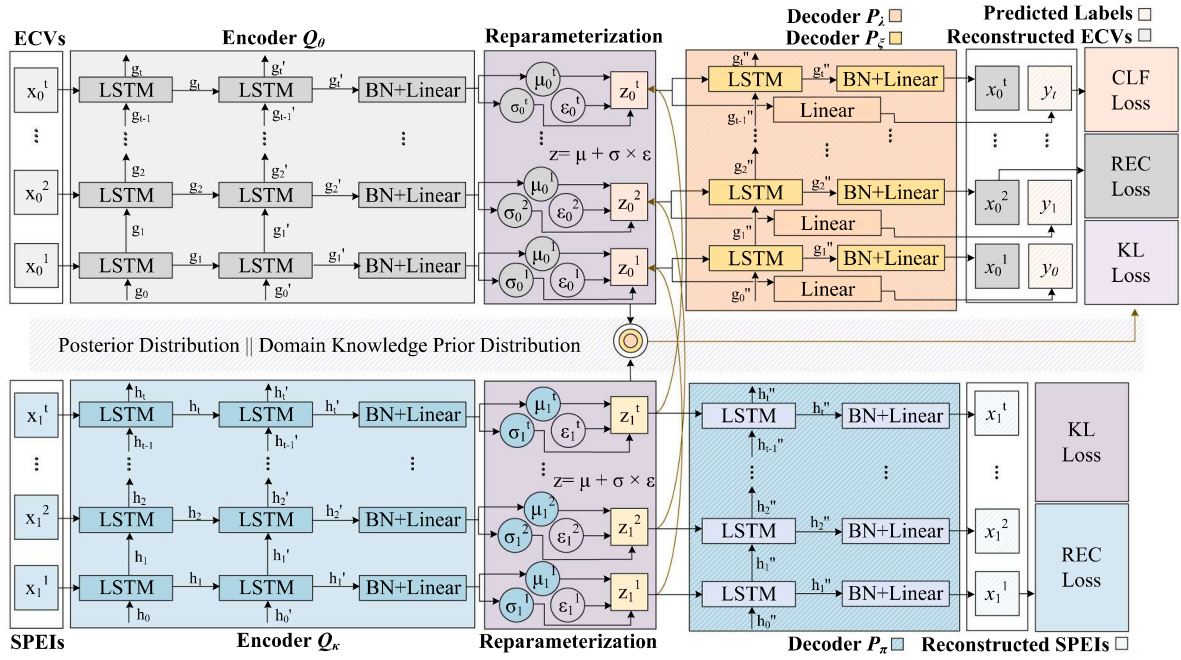


Fig. 5. Graphical illustration of DK-VRN for drought monitoring. At time step t , x_0^t and x_1^t are the ECVs and the SPEI data, respectively; μ^t , σ^t are the mean and standard deviation parameters of the distribution of the latent variables, respectively; random variables are denoted by ϵ^t ; z_0^t , z_1^t are the samples from the latent distribution $Q_\theta(x_0)$ and $Q_\kappa(x_1)$, respectively; and y^t are the drought labels. Note that the above variables are vectors instead of scalars. The CLF loss is the classification loss given the drought labels, REC Loss is the reconstruction loss given the input data x_0, x_1 , and KL Loss is the Kullback-Leibler (KL) divergence between the posterior distribution and the prior distribution. Ultimately, the proposed DK-VRN is composed of two variational encoders Q_θ, Q_κ and three generative decoders $P_\lambda, P_\epsilon, P_\pi$.

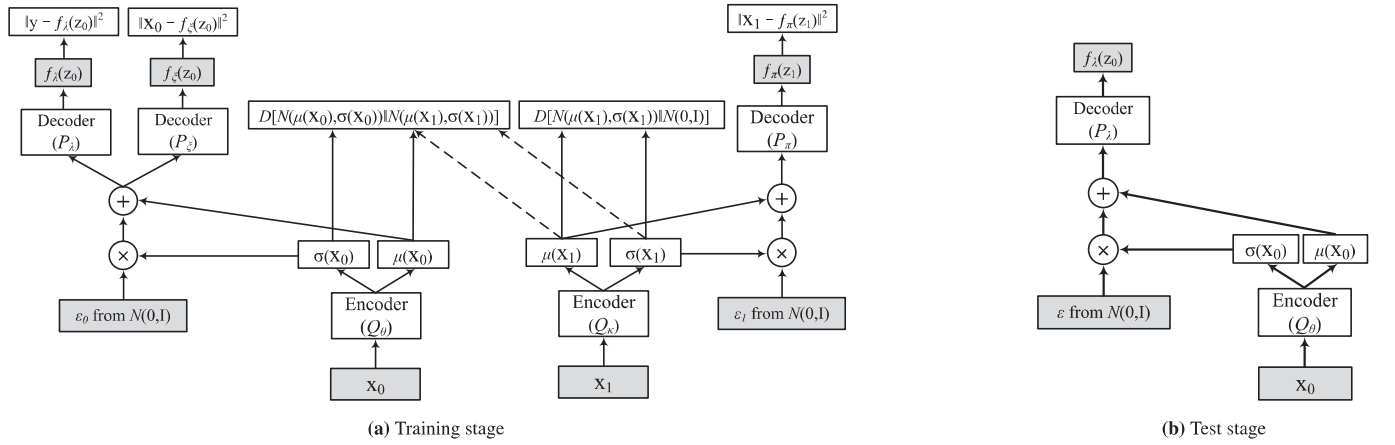


Fig. 6. Training and test stages of the proposed DK-VRN. x_0 denotes the ECVs input data, x_1 denotes the SPEI input data, ϵ denotes the random variables, y denotes the drought labels, and D denotes the KL divergence loss function.

variables $Q_\kappa(z_1|x_1)$. Second, the reparameterization trick yields samples $z_1^{(k)} \sim Q_\kappa$. Third, the decoder P_π generates the input $f_\pi(z_1)$. Fourth, values for ECVs x_0 given all training time steps constitute the input data to the encoder Q_θ , which produces the distribution of latent variables $Q_\theta(z_0|x_0)$. Note that $Q_\kappa(z_0|x_0)$ is set to the prior distribution in the KL divergence term of $Q_\theta(z_0|x_0)$. Fifth, the reparameterization trick yields samples $z_0^{(k)} \sim Q_\theta$. Sixth, the decoder P_λ and P_ϵ generate the drought probability $f_\lambda(z_0)$ and the input $f_\epsilon(z_0)$, respectively. Finally, $Q_\theta, Q_\kappa, P_\lambda, P_\epsilon, P_\pi$ are updated end-to-end via the mini-batch stochastic gradient descent for $\mathcal{L}$:

$$\begin{aligned} \{\kappa, \pi\} &\leftarrow \{\kappa, \pi\} + \{\tilde{\nabla}_\kappa, \tilde{\nabla}_\pi\} \sum_{x_1} \mathcal{L}(x_1; \kappa, \pi) \\ \{\theta, \lambda, \xi\} &\leftarrow \{\theta, \lambda, \xi\} + \{\tilde{\nabla}_\theta, \tilde{\nabla}_\lambda, \tilde{\nabla}_\xi\} \sum_{y, x_0} \mathcal{L}(y, x_0; \theta, \lambda, \xi), \end{aligned} \quad (13)$$

where the classification loss term between y and $f_\lambda(z_0^{(l)})$ is set to the Binary Cross-Entropy (BCE), the reconstruction loss terms between x_0 and $f_\epsilon(z_0^{(l)})$, and x_1 and $f_\pi(z_0^{(l)})$, are both set to the Mean Squared Error (MSE) and the KL divergence loss term is computed between the posterior distribution and the domain knowledge prior distribution, as mentioned above.

On the other hand, the test stage is shown in Fig. 6(b). First, x_0 and ϵ_0 of all time steps (till the current time step) are the input data, and the

encoder Q_θ produces the distribution of latent variables $Q_\theta(z_0|x_0)$. Second, the reparameterization trick yields samples $z_0^{(l)} \sim Q_\theta$. Third, the decoder P_λ produces the distribution of the drought probability. Finally, the estimated expectation of the drought probability and of the uncertainty are computed by means of importance sampling, being S the number of generated outputs:

$$\mathbb{E}_Q[f_i(z_0)] \approx \frac{1}{S} \sum_{s=1}^S f_i(z_0^{(s)}), \quad (14)$$

$$\text{Var}_Q[f_i(z_0)] \approx \frac{1}{S} \sum_{s=1}^S [f_i(z_0^{(k)}) - \mathbb{E}_Q[f_i(z_0)]]^2,$$

where $z_0^{(s)} \sim Q_\theta(z_0|x_0)$. It should be noted that the uncertainty essentially derives from the uncertain nature of the data and how well the model is fitted (Gawlikowski et al., 2023a). For drought monitoring, the observational errors (aleatoric uncertainty) represent the noise of the processed input data and the noise of the recorded drought events. About the model, the predictive errors (epistemic uncertainty) include errors in the model structure, errors in the training process, and so forth. In this context, we perform variational inference on the latent variables z_0 to estimate the predictive errors, while the uncertainty from the data and from the model parameters are ignored here.

Furthermore, it is noteworthy that time-series SPEI x_1 is only required at the training stage of DK-VRN, not during the test stage, and the monitoring resolution is only constrained to the resolution of the ECVs x_0 , which is therefore applicable for drought monitoring with a finer temporal resolution or short-term drought prediction when monthly SPEI is not available.

4. Experimental results

4.1. Experimental setup

A) *Evaluation metrics.* Standard quantitative metrics based on recorded drought events, including the Area Under the Receiver Operating Characteristic Curve (ROC-AUC), the Area Under the Precision-Recall Curve (PR-AUC), Macro F1-score (Macro F1), and the Intersection over Union (IoU), also known as Ruzicka similarity (Cha, 2007), are adopted for evaluation purposes. Additionally, training and test running times are computed to quantitatively analyze the computational efficiency of the different methods considered in this study, while the number of parameters (Params) for each model is reported as a complexity measure. Then, we establish three categories for evaluation purposes: the event evaluation, the climate classification evaluation, and the overall evaluation. In the event evaluation, the quantitative metrics are computed by only considering specific locations where droughts happened (Afghanistan, Italy, Moldova, Russia). In the climate classification evaluation, the quantitative metrics are computed by considering samples belonging to different Köppen-Geiger climate zones (Beck et al., 2018). In the overall evaluation, the quantitative metrics are computed by considering all locations of the entire European drought-affected region (Europe-0) and all locations of the entire European region (Europe-1). The final test results are reported by averaging the outputs of 5 independent runs under different random seeds.

B) *Model baselines.* We compare the proposed DK-VRN with the SPEI-6, as well as three deterministic neural networks: Multi-Layer Perceptron (MLP), Recurrent Neural Network (RNN) and Long Short-Term Memory Network (LSTM). We also consider three variational neural networks: variational MLP (VMLP), variational RNN (VRNN) and variational LSTM (VLSTM). The prior distribution for these variational models is set by default, i.e., as a multivariate Gaussian distribution with zero mean vector and identity covariance matrix, and

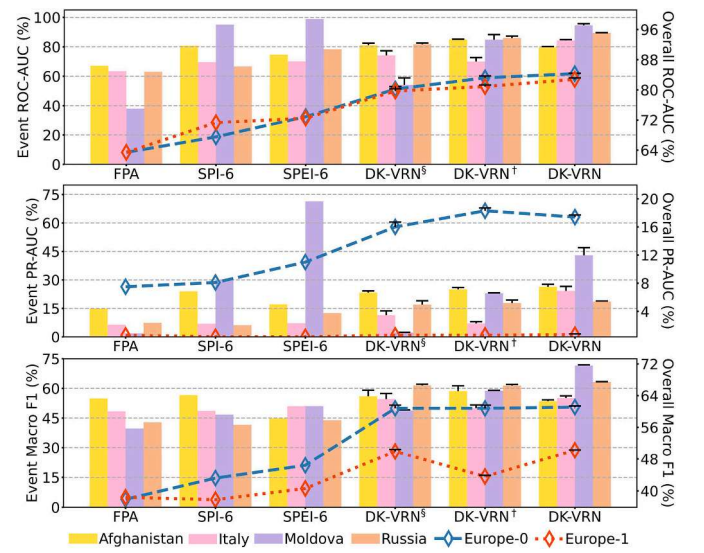


Fig. 7. Performance of DK-VRN for different prior indices in this study. Final results are reported by averaging the outputs of 5 independent runs under different random seeds. The error bars denote the one standard deviation. 'DK-VRN^I' denotes the variant of DK-VRN (SPEI-6 prior). 'DK-VRN^S' denotes the variant of DK-VRN (FAPAR anomaly, FPA prior). For each method, bars for all event regions (Afghanistan, Italy, Moldova, Russia) are sorted following the order of the legend of the figure. Dots are used to show the overall (Europe-0, Europe-1) scores.

no reconstruction loss is used to train these models (only classification loss). Specifically, their architectures are detailed in Table A.2.

C) *Hyperparameters.* The network optimizer is set to Adam (Kingma and Ba, 2014), being the learning rate equal to $1e-4$ and the gradient clip set to $1e-5$. The batch size is 32, the number of epochs is equal to 200, and the early stopping technique is applied to select the best model based on the validation ROC-AUC. The above setting is kept the same for all data-driven methods.

D) *Implementation details.* Ubuntu 20.04.4 LTS, AMD Ryzen 75,800 h with Radeon graphics $\times 16$, 64GiB Memory, 2.5 TB Disk Memory, NVIDIA GeForce RTX 3070 Laptop GPU (8192MiB) with PyTorch 1.10.2 (Paszke et al., 2019), Python 3.9, CUDA 11.3, and cuDNN 8.2.0. The source code for DK-VRN is publicly available on GitHub.⁴

4.2. Analysis of drought indices

First, we investigate the impact of different x_1 as prior domain knowledge (see Eq. (7)) on the performance of DK-VRN. SPI, FAPAR anomaly (FPA) (European Commission, 2021), and SPEI are selected as prior domain knowledge candidates for drought monitoring. Then, three variants of DK-VRN are designed and evaluated. The results are given in Fig. 7. Overall, all DK-VRN models significantly outperform pure drought indices. Among all DK-VRNs, DK-VRN with SPEI prior (our proposal) achieves the best performance, which suggests its effectiveness for meteorological drought monitoring. Nonetheless, a few exceptions can be observed in Afghanistan, which indicates the choice of drought indices may be location-dependent. Similar observations can be found when comparing pure indices. In this way, SPEI is set as the prior domain knowledge for the rest of the experiments.

⁴ <https://github.com/mengxue-rs/DK-VRN>

4.3. Analysis of λ Hyperparameter

Second, we analyze the impact of the λ parameter (see Eq. (8)) on the performance of DK-VRN. To this end, λ is set to logarithmic values lying in the range $[10^{-5}, 10^0]$. The experimental results in terms of ROC-AUC are reported in Fig. 8. As can be appreciated, a best λ value for all drought events does not exist, whereas a moderate λ is appropriate for all situations. For instance, DK-VRN with $\lambda = 10^0$ achieves the best ROC-AUC in Moldova, while performing worst in other regions. Similarly, DK-VRN with $\lambda = 10^{-5}$ does not perform well in Afghanistan and Moldova. Similar results for PR-AUC and F1 metrics are shown in Fig. A.3. As for the overall evaluation, DK-VRN with $\lambda = 10^{-3}$ perform best in Europe-0, whereas DK-VRN with $\lambda = 10^{-2}$ perform best in Europe-1. The differences are related to the distribution of drought samples: Europe-1 has much more non-drought samples than Europe-0, which affects the results and reduces the difference between methods. Although a flexible λ may bring better results, our goal in this paper is to provide a generic model that is suitable for different regions. In this context, we set a global λ to 10^{-2} for the rest of the experiments.

4.4. Results on drought monitoring

Then we report the Drought monitoring performance obtained by DK-VRN and the baseline methods mentioned above. The ROC-AUC, the PR-AUC, and the F1 scores of the different methods can be found in Fig. 9 and Fig. 10. ROC and PR curves are shown in Fig. A.4 and Fig. A.5, respectively.

As for the event evaluations, DK-VRN achieves the highest performance in Afghanistan, Italy, and Russia (ROC-AUC of 79.7 ± 0.5 , 84.3 ± 0.6 , and 89.4 ± 0.2 , respectively). However, a few exceptions can be observed in Moldova, where RNN achieves the highest ROC-AUC, and also SPEI-6 and LSTM outperform variational models, including DK-VRN. This can be explained by the potential blurs caused by these models, which may hamper the recognition performance, particularly when attempting to capture drought events of a small extent like the one happening in Moldova.

As for the climate classification evaluation (see Fig. A.6 for the visualization of the climate classification map), DK-VRN achieves the best results over 11 out of 15 climate zones across the whole of Europe, occupying over 74% of the land. Some exceptions can be noticed: SPEI-6 performs better in BWk (cold desert), BSk (cold semi-arid), Dsc (dry-summer subarctic), and ET (tundra). This can be partly explained by the imbalance when given the drought events in the training, validation, and test subsets (see climate zone proportions in Table A.3), which impacts the performance of data-driven models and hybrid models on the BWk (3.46% train, 3.88% val, 9.22% test), BSk (15.40% train, 15.18% val, 10.54% test), Dsc (5.21% train, no locations in val, 0.23% test), and ET (0.88% train, 0.02% val, 5.48% test).

Overall, DK-VRN achieves the best ROC-AUC, PR-AUC, and Macro F1 in both Europe-0 and Europe-1. Moreover, the standard deviations of DK-VRN are remarkably lower than others. In contrast, MLP and VMLP

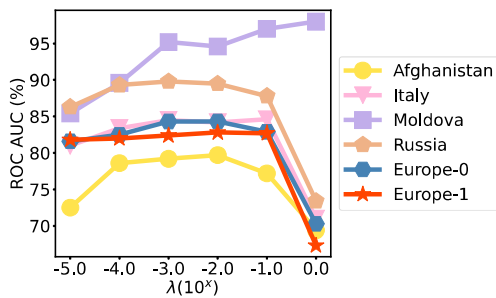


Fig. 8. Performance of DK-VRN for different λ values and the drought events considered in this study.

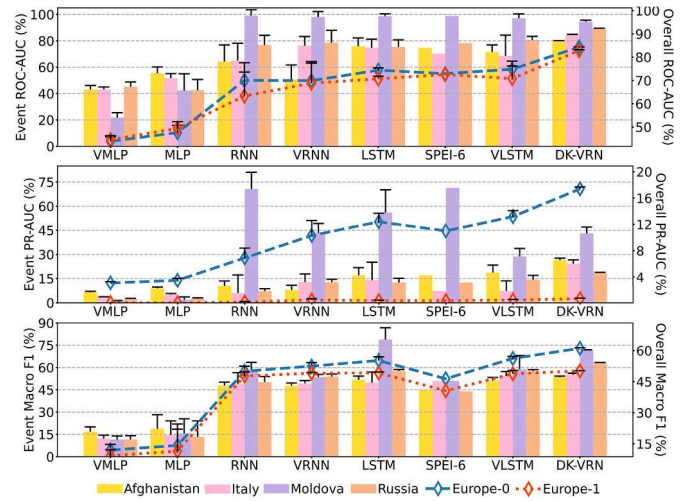


Fig. 9. Drought monitoring performance obtained for the test subset. Final results are reported by averaging the outputs of five independent runs under different random seeds. The error bars denote the one standard deviation. For each method, bars for all event regions (Afghanistan, Italy, Moldova, Russia) are sorted following the order of the legend of the figure. Dots are used to show the overall (Europe-0, Europe-1) scores.

are inferior in all evaluations, demonstrating the advantages of a recurrent architecture for time series modelling. RNN versus VRNN and LSTM versus VLSTM achieve comparable results, indicating the improvement of DK-VRN is little related to the VAE architecture. Comparably, the significant improvement achieved by DK-VRN over VLSTM justifies the effectiveness of the proposed domain knowledge prior. In the end, all the results demonstrate the superiority of DK-VRN compared to the other baseline methods, thus DK-VRN is more applicable for drought monitoring.

4.5. Temporal cross-validation test

In this experiment, we shift the order of the training, the validation, and the test subsets to simulate the data uncertainty from the period selection. The final results of the three-fold cross-validation test obtained by the proposed and baseline methods are reported in Fig. 11. Except for MLP and VMLP, the degradation of performance for almost all methods is distinct. This deterioration can be explained: first, the unbalanced spatio-temporal distribution of drought events makes monitoring challenging for models trained on regions of different climate zones; second, the non-drought labels of some locations may be noisy, which hampers the performance of the systems; third, the shift operation may be unbeneficial for processing temporal information by means of recurrent architectures. Although DK-VRN also experiences performance degradation, it still achieves the highest ROC-AUC, PR-AUC, and Macro F1 scores, which are markedly higher than others. As a result, when considering the inherent uncertainty from the data period, which constitutes a challenging scenario to evaluate drought monitoring, the performance of DK-VRN is better than other baseline methods.

4.6. Generalization analysis

Now we analyze the generalization improvement achieved when encoding SPEI as the prior distribution. To this end, we compare the results of the variational approaches with those obtained by using SPEIs directly as input features. Two variants of VRNN and VLSTM (all incorporate SPEIs as input features) are designed to be compared with DK-VRN. Their results are given in Fig. 12. Overall, both the performance of VRNN[§] and VLSTM[§] are inferior to DK-VRN (only VRNN[§] performs better than DK-VRN in Moldova). However, VRNN[§] is even

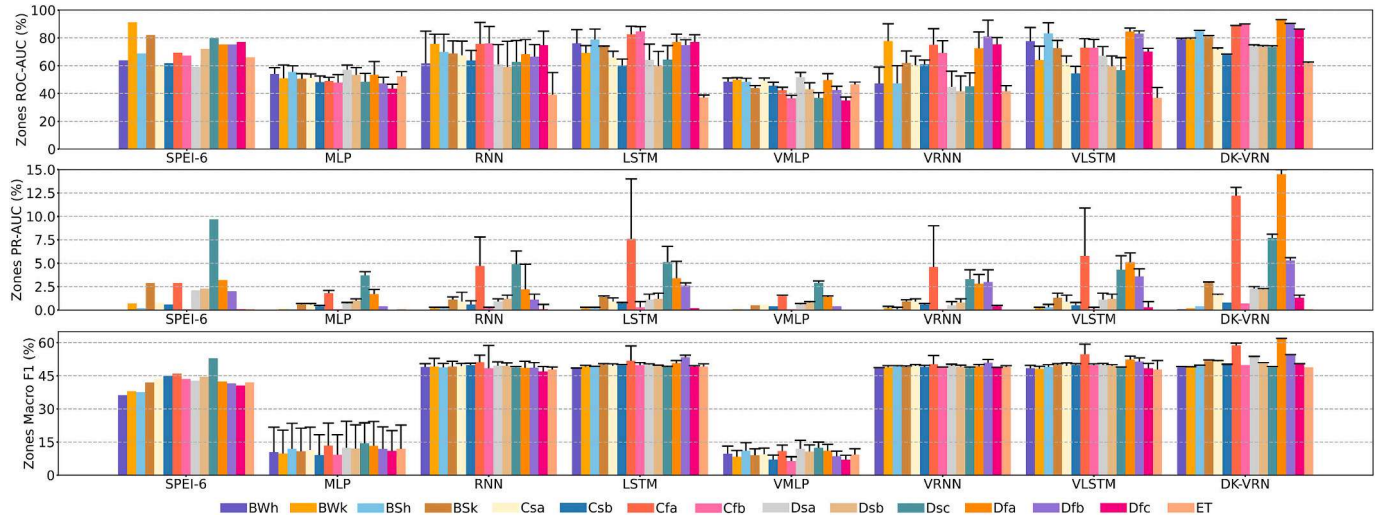


Fig. 10. Drought monitoring performance obtained for the test subset and different climate classifications. Final results are reported by averaging the outputs of five independent runs under different random seeds. The error bars denote the one standard deviation. The full names of all climate classifications can be found in Fig. A.6. For each method, bars for the climate regions are sorted following the order of the legend of the figure.

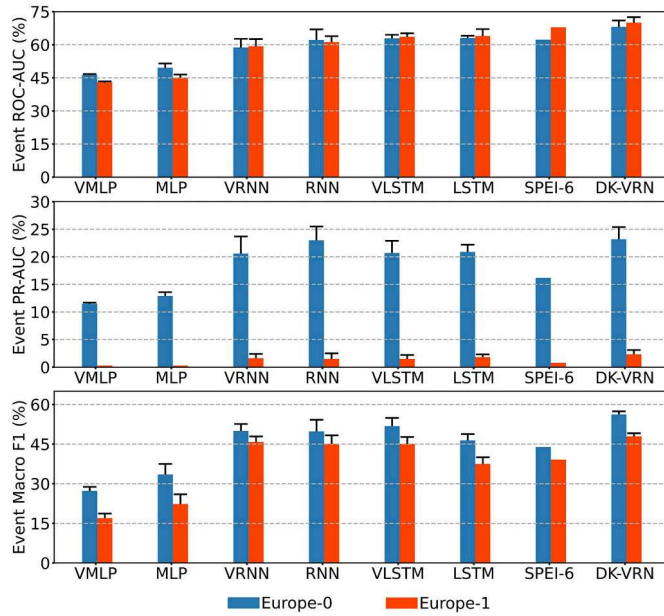


Fig. 11. Temporal cross-validation performance. Final results are reported by averaging the outputs of 5 independent runs under different random seeds. The error bars denote the one standard deviation. For each method, bars for the overall regions (Europe-0, Europe-1) are sorted following the order of the legend of the figure.

inferior to the original VRNN in Russia and Italy. Similar results can be observed for VLSTM[§]. All the results demonstrate that directly using SPEI as input features may be prone to over-fitting and, therefore, is unsuitable for large-scale monitoring over Europe. It should be noted that the network architecture and training process are kept the same for a fair comparison. In these circumstances, VRNN[§] and VLSTM[§] did not provide better performances compared to VRNN and VLSTM. This may be attributed to the different temporal manifolds for ECVs, which

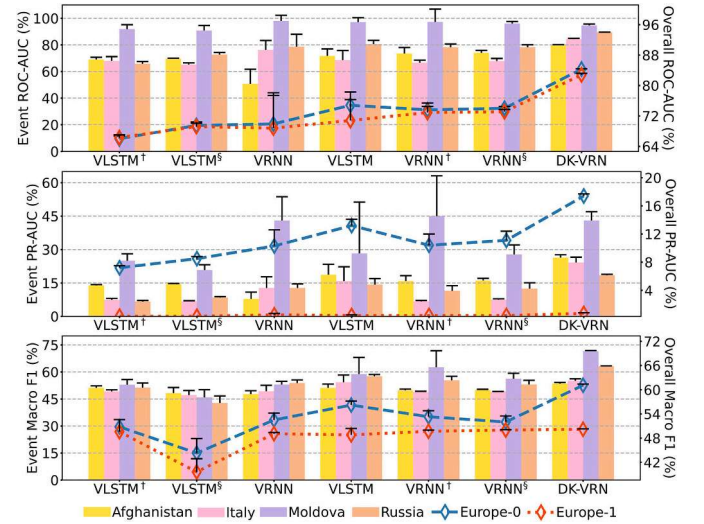


Fig. 12. Performance when using SPEIs as input features versus as prior distribution (DK-VRN, our approach). Final results are reported by averaging the outputs of five independent runs under different random seeds. The error bars denote the one standard deviation. 'VRNN^¹' denotes only using SPEIs as input features in VRNN. 'VRNN^²' uses concatenated ECVs and SPEIs as input features in VRNN. 'VLSTM^¹' only uses SPEIs as input features in VLSTM. 'VLSTM^²' uses concatenated ECVs and SPEIs as input features in VLSTM. For each method, bars for all event regions (Afghanistan, Italy, Moldova, Russia) are sorted following the order of the legend of the figure. Dots are used to show the overall (Europe-0, Europe-1) scores.

constitute raw low-level, local features, and drought indices. They could be understood as high-level features derived from ECVs, able to capture global information. The performance of VRNN[§] and VLSTM[§] can be further improved using a multi-branch architecture (Zhang et al., 2022). In this architecture, two different branches are implemented for ECVs and drought indices, respectively, using convolutional lateral

Table 2

Ablation study on the use of different loss terms when training the proposed DK-VRN (%).

	CLF-Loss	REC-Loss	KL-Loss	ROC-AUC	PR-AUC	Macro F1
Afghanistan	✓	✗	standard	71.6 ± 5.3	18.7 ± 4.7	51.2 ± 2.1
	✓	✗	domain knowledge	77.0 ± 4.3	26.1 ± 3.9	53.5 ± 3.0
	✓	✓	standard	73.4 ± 2.9	19.9 ± 3.5	49.3 ± 3.1
	✓	✓	domain knowledge	79.7 ± 0.5	26.4 ± 1.3	53.5 ± 0.7
Italy	✓	✗	standard	68.5 ± 7.2	15.8 ± 6.5	54.4 ± 4.0
	✓	✗	domain knowledge	74.7 ± 8.3	15.9 ± 9.6	56.7 ± 5.1
	✓	✓	standard	76.0 ± 5.1	12.1 ± 2.6	52.1 ± 1.9
	✓	✓	domain knowledge	84.3 ± 0.6	24.2 ± 2.4	55.1 ± 1.1
Moldova	✓	✗	standard	97.1 ± 3.3	28.6 ± 3.0	59.1 ± 9.1
	✓	✗	domain knowledge	98.2 ± 3.1	41.1 ± 7.2	75.9 ± 9.5
	✓	✓	standard	92.1 ± 9.4	21.5 ± 8.5	64.6 ± 11
	✓	✓	domain knowledge	94.6 ± 1.1	43.0 ± 4.0	71.5 ± 0.4
Russia	✓	✗	standard	80.6 ± 2.8	14.3 ± 2.7	57.6 ± 1.0
	✓	✗	domain knowledge	82.5 ± 2.7	13.7 ± 0.0	57.4 ± 2.3
	✓	✓	standard	86.0 ± 3.0	17.7 ± 2.8	60.5 ± 1.6
	✓	✓	domain knowledge	89.4 ± 0.2	18.5 ± 0.4	63.2 ± 0.2
Europe-0	✓	✗	standard	74.8 ± 3.5	13.2 ± 0.9	56.2 ± 1.0
	✓	✗	domain knowledge	80.2 ± 3.2	14.4 ± 3.9	58.0 ± 1.9
	✓	✓	standard	79.6 ± 0.9	16.7 ± 1.3	58.1 ± 1.0
	✓	✓	domain knowledge	84.3 ± 0.1	17.4 ± 0.3	61.1 ± 0.3
Europe-1	✓	✗	standard	70.9 ± 5.5	0.5 ± 0.2	48.8 ± 1.6
	✓	✗	domain knowledge	76.9 ± 4.7	0.5 ± 0.2	48.4 ± 0.7
	✓	✓	standard	76.5 ± 5.0	0.4 ± 0.1	49.2 ± 0.9
	✓	✓	domain knowledge	82.8 ± 0.4	0.7 ± 0.1	50.2 ± 0.1

connections for blending the data. The improved results are given in Fig. A.7, being the superiority of DK-VRN over these variants still observed. In the end, DK-VRN exploits SPEIs as the prior distribution and the KL term acts as a regularizer in the training process, which is beneficial to model generalization.

4.7. Ablation study

In this ablation study, we evaluate the contribution of the different loss terms to the performance of DK-VRN. As shown in Table 2, the integration of the classification loss, the reconstruction loss, and the domain knowledge-based KL-divergence loss yields the highest ROC-AUC, PR-AUC, and Macro F1 scores for the overall evaluation, which claims the superiority of DK-VRN. Only a few exceptions appear in Moldova, where DK-VRN without the reconstruction loss achieves the highest ROC-AUC and F1 scores. The performance of the models with the domain knowledge-based KL-Loss is always superior to the performance of the models with the standard KL-Loss, which demonstrates the effectiveness of the domain knowledge prior in DK-VRN. The improvement in stability (lower standard deviation) can be partially observed when using the REC-Loss. Accordingly, the standard deviation of DK-VRN is lower than the standard deviations of other variants, indicating the integrated losses are beneficial to the stability of the model. All the results demonstrate that the domain-knowledge KL and REC losses cooperatively contribute to the performance improvements achieved in DK-VRN.

4.8. Architecture complexity analysis

Table 3 shows the running training (in minutes) and test times (in seconds), as well as the number of parameters (Params), of the different drought monitoring models in this study. SPEI-6, whose computation has been accelerated with multi-thread CPU processing, has the shortest test time in Afghanistan, Italy, Moldova, and Russia. Nevertheless, for large-scale data, as for Europe-0 and Europe-1 regions, the inference of deep learning models is certainly improved using GPU processing, which allows for lower computational times when using models such as RNN and LSTM. Among data-driven models, MLP is the quickest at 28.19 min. DK-VRN takes longer at 86.09 min due to its dual-branch operations during training. Regarding test times, RNN is the most efficient, with LSTM slightly slower than VRNN due to the complexity of LSTM cell operations. DK-VRN's test time is comparable to VLSTM as both share the same architecture at this stage. As for model parameters, all methods have a limited parameter size to avoid over-fitting. It should be noted that, although the number of parameters for DK-VRN is higher than in VLSTM, the extra parameters for SPEIs modelling in DK-VRN can be discarded at the test stage, as shown in Fig. 6(b), which demonstrates that DK-VRN achieves superior performance than VLSTM without increasing the test complexity.

5. Discussion and qualitative evaluation

5.1. Visualizations of detection signals

Now we visually analyze the spatially aggregated drought signals through time and their uncertainty for some of the methods: SPEI-6,

Table 3

Computational complexity and number of parameters of the different drought monitoring models in this study.

		SPEI-6	MLP	RNN	LSTM	VMLP	VRNN	VLSTM	DK-VRN
Training time (min)		–	28.19	42.34	42.39	44.98	54.64	57.24	86.09
	Afghanistan	2.16	3.92	3.26	3.28	4.55	4.11	4.21	4.26
Test time (s)	Italy	2.94	4.00	3.21	3.06	4.46	4.04	4.33	5.19
	Moldova	0.18	1.35	1.29	1.64	1.00	1.42	1.55	1.92
	Russia	10.36	12.44	12.04	12.94	17.58	14.97	15.80	15.27
	Europe-0	19.67	22.37	18.37	20.55	26.30	22.67	23.75	23.23
	Europe-1	485.59	489.95	460.79	466.74	815.78	713.31	732.60	736.34
	Params	–	3905	4417	14,209	8721	7715	23,843	56,392

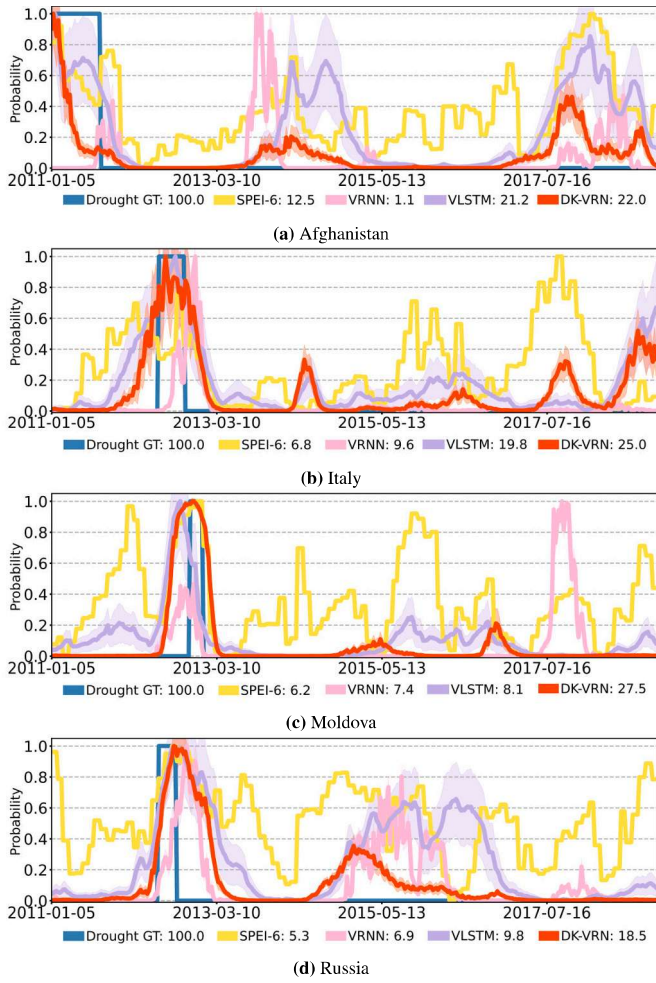


Fig. 13. Spatially aggregated drought detection signals for the different models considered in the experiments. Note that the drought event-based signals are obtained by computing, for each time step, the average value of all pixels belonging to their corresponding regions, being the values later normalized to [0,1]. The shaded area shows the estimated uncertainty (one standard deviation) based on 10 runs under different latent variables. The IoU (see legend) is computed based on the ground truth curves (green) and the prediction signals for the different models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

VRNN, VLSTM, and DK-VRN. Given a drought event, these signals are obtained by computing, for each time step, the average output value of all pixels belonging to its corresponding region, normalized later to a [0, 1] scale. The IoU is computed to measure the similarity between the prediction and the ground truth signals (higher is better). Variational models incorporate uncertainty estimation by learning the distribution of latent variables, allowing for different predictions based on several latent variable samples. As depicted in Fig. 13, DK-VRN excels in recognition and demonstrates lower uncertainty than other methods. It effectively identifies the onset, duration, and conclusion of drought events in Afghanistan, Italy, Moldova, and Russia. In contrast, SPEI-6 tends to yield more temporal false alarms, VRNN encounters challenges in Afghanistan and Italy, and VLSTM exhibits substantial monitoring uncertainty in these regions. SPEI-6, being based on physical PET

estimation and a fixed probabilistic distribution, lacks direct uncertainty estimation, which arises from variations in PET estimates and distributions. In summary, DK-VRN, enhanced by domain knowledge, offers superior regional monitoring performance in terms of accuracy and stability.

5.2. Visualizations of probability maps

We then visualize the drought probability maps together with the uncertainty ones for some methods, including SPEI-6, VLSTM, and DK-VRN. As shown in Fig. 14, SPEI-6 maps exhibit more spatial false alarms than others. In contrast, DK-VRN determines fewer false alarms without sacrificing true positive rates. For instance, SPEI-6 identified influential drought events on 2012-07-07, both in Saudi Arabia, while DK-VRN provides a lower probability, which is more accurate based on drought labels. Another example is given on 2011-12-07, when VLSTM identified a severe drought in Norway whereas DK-VRN did not, also aligned with annotations. Besides, DK-VRN maps present lower uncertainty compared to VLSTM, which suggests the validity and reliability of DK-VRN. Additional example drought probability maps for DK-VRN can be found in Fig. A.8. In summary, DK-VRN provides superior overall drought monitoring performance in terms of generalization and reliability. However, there are some misclassification drought samples around the southeast region of Italy and the northeast region of Russia, suggesting the risk of over-fitting. This may be relevant to the imbalanced positive-negative samples of the training set. Overall, DK-VRN provides superior drought monitoring performance in terms of generalization and reliability.

5.3. Domain knowledge prior analysis

Next, we uncover the role of the proposed SPEI-based domain knowledge prior in DK-VRN. The measurements observed for the validation process are plotted for analysis, and consistent observations can be appreciated during the training process illustrated in Fig. A.9(a)-(f).

First, the domain knowledge prior distribution of the latent variables implies generative knowledge. As shown in Fig. 15(a), the KL divergence term for introducing the domain knowledge prior conveys the information of generating SPEIs to the posterior distribution. When using a large λ , the validation MSE between the SPEIs generated by $P_{\pi}(x_1|z_0)$ and the true SPEIs x_1 is relatively low, and vice versa. These results demonstrate that 1) the encoding preserves the essential information of SPEIs, and 2) the KL term between ECVs and SPEIs effectively transfers that information.

Second, the KL term acts as a regularizer. The effects of this regularization are inspected in Fig. 15(b)-(d). It can be observed that when using a large λ , KL loss is controlled at a low level. Meanwhile, the reconstruction loss of x_0 is kept at a relative level, and the validation classification loss is relatively stable and converges to a low level. Conversely, a low λ results in high KL loss, as well as low reconstruction loss but high and rickety validation classification loss. The results indicate that the new KL term turns into a domain knowledge-based regularization item, which can guide the training process, and consequently strengthen the model generalization and stability.

Finally, $Q_{\kappa}(z|x_1)$ reserves the prior anomalous information presented in time-series SPEIs, and the KL term encourages $Q_{\theta}(z|x_0)$ to perform consistently. The Mean Absolute Error (MAE) of quartiles Q_1 and Q_3 for the mean parameters of the two distributions is shown in Fig. 15(e)-(f). It can be observed that using a small to a large λ , the mean parameters are

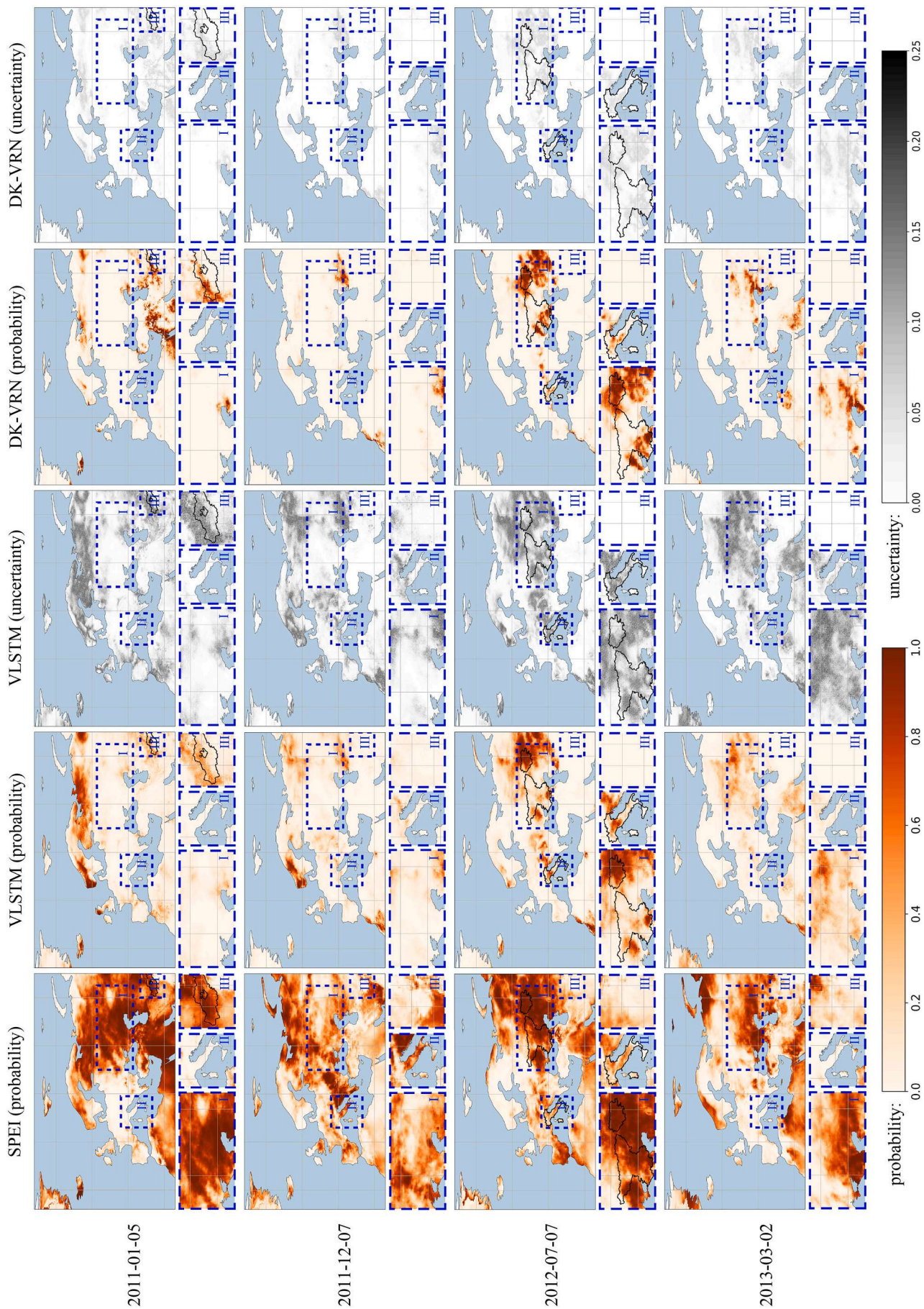


Fig. 14. Time-series drought probability and uncertainty maps provided by different models. These maps are picked at four different time steps (2011-01-05; 2011-12-07; 2012-07-07; 2013-03-02) accompanying three drought events: I (Russia), II (Italy), and III (Afghanistan) spanning from Jan 2011 to Mar 2013. The ground truth is marked with black lines (droughts happen at a particular time step if they are shown), and the regional maps for all drought events are highlighted with blue dashed lines. It should be noted that SPEI-6 does not provide the uncertainty maps. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

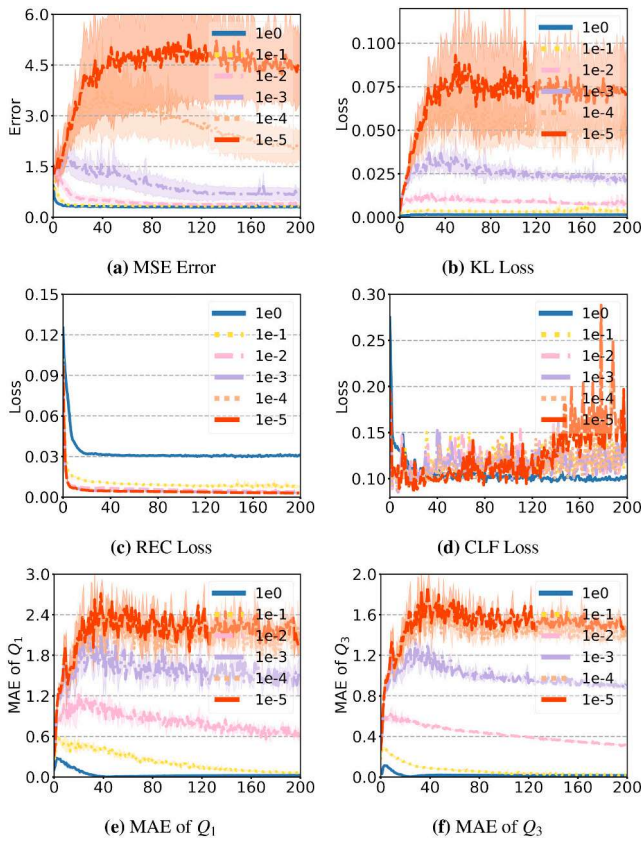


Fig. 15. Evolution of the metrics observed during the validation process for different λ values. The evolution of the validation MSE between the SPEIs generated by $P_\pi(x_1|z_0)$ and the true SPEIs x_1 is depicted in (a), the evolution of the different validation loss terms (KL, reconstruction and classification) is plotted in (b)-(d), and the evolution of the validation MAE of quartiles Q_1 and Q_3 of the mean parameters is shown in (e)-(f).

pulled together for the first and the third quartiles at the beginning steps, varying from a large MAE to a low MAE. For example, with $\lambda = 1e-2$, the MAE of Q_3 is one-third of the MAE of Q_3 when $\lambda = 1e-3$. In other words, when the statistics of $Q_k(z|x_1)$ are abnormal (away from the median), the KL term stimulates the statistics of $Q_\theta(z|x_0)$ to become consistently anomalous, revealing that prior anomalous information provided by multi-scalar SPEIs is learned by $Q_\theta(z|x_0)$.

5.4. Limitations and potentials of DK-VRN

Let us finish commenting on the main potentials and limitations of the proposed hybrid methodology, which is generic and could be transferred to other disasters, such as floods, wildfires, and heatwaves. Particularly for drought events covered in this study, we should first mention the limitations associated with the considered input variables. Droughts are complex phenomena involving numerous geophysical variables. In this research, we used the total precipitation and air temperature of ERA-5 data because of the physical consistency and the long-term records that are required for the computations of multi-scalar SPEI. For instance, a time series over 20 years is recommended for fitting the parameters of SPEI. However, ERA5 is updated daily with a latency of five days, which restricts the real-time applications for drought moni-

toring. Alternatively, satellite data such as GPM (Huffman et al., 2015) could provide real-time large-scale estimations of precipitation within an hour. Accordingly, the incorporation of multi-sensor data can be beneficial for real-time drought monitoring.

Second, regarding the limitations of the target variable, it should be noted the ambiguity of the drought definition itself, which hassles the evaluation of algorithms. Understandably, various stakeholders have different interests relevant to droughts. Meteorological departments are more concerned about the precipitation deficit drought, agricultural sectors care more about the soil moisture deficit drought, and farmers are more worried about the impact on crop production. In this research, we relied on the EM-DAT and the GDIS definitions for meteorological purposes. Compared to SPEI, the EM-DAT and the GDIS provided objective evaluations regarding actual and influential drought events. However, bias and incorrectness may affect the data quality. Moreover, the duration and severity of droughts are lacking in the used data, which are dependent on the atmosphere and the hydrologic processes that feed moisture to the atmosphere (Bravar and Kavvas, 1991). Fortunately, with advanced sensors, more physical-consistent or expert-consistent data can be obtained for drought exploration. Empirically, the performance of DK-VRN could be improved with drought event annotations of higher quality. Furthermore, droughts are extreme climate events that cause imbalanced problems for data-driven models. Operative strategies for data and training processing can alleviate this issue, such as data augmentation, data under-sampling, and balanced loss, to name a few. It is therefore important to tackle the imbalanced problem of drought labels.

Finally, there are limitations associated to the very same methodologies. As a hybrid model, DK-VRN synthesizes the drought index- and deep learning-based features to estimate the joint distribution of climate variables and drought events, which represent the captured latent process of the drought dynamics. However, DK-VRN is a neural network, and thus a black box model difficult to interpret and understand. Enhancing the transparency and interpretability of DK-VRN is an essential step for the deployment of a monitoring system into real applications. Besides, uncertainty analysis and quantification are indispensable for a trustworthy monitoring framework. For instance, different train/val/test splitting affects the monitoring results which is clarified in the temporal cross-validation test. Similarly, different period lengths may discriminate the model training process and the test results. The influences can be analytically estimated with the quantification of the uncertainty. In the future, it would be advisable providing explainability scores and appropriate uncertainty quantification associated to both the hybrid and non-hybrid approaches (Vilone and Longo (2020); Gawlikowski et al. (2023b)).

6. Conclusion

In this paper, we have proposed a novel framework, DK-VRN, which, to our knowledge, is the first hybrid variational model for drought monitoring. DK-VRN realizes the synergy of deep learning models and prior domain knowledge from climatic indices in a probabilistic framework, and the entire architecture is learnable end-to-end. Experimental results on several representative regional droughts and the overall European drought have validated the superiority of DK-VRN compared to SPEI-6 and data-driven deep learning methods of MLP, RNN, LSTM, VMLP, VRNN, and VLSM. DK-VRN achieves the best ROC-AUC (%) results: Afghanistan (79.7 ± 0.5), Italy (84.3 ± 0.6), Russia (89.4 ± 0.2), Europe-0 (84.3 ± 0.1), Europe-1 (82.8 ± 0.4), effectively

capturing the starting and ending times of drought events with higher accuracy and lower uncertainty, and also generalizing better for unseen locations. Moreover, we showcased that the proposed framework is generic and could be easily adjusted to use other remote sensing data sources as input, such as FAPAR anomaly. In future work, we plan to investigate the uncertainty of the input data, model parameters, and labels. Apart from the uncertainty estimates, we plan to explore the inconsistency among multiple drought indicators and integrate satellite observations and physical hydrological models involved in drought formation to build a more credible and accountable system while learning the relative relevance of all the data streams and variables.

CRediT authorship contribution statement

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

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Gustau Camps-Valls reports financial support was provided by University of Valencia. He is part of the invited editorial board for RSE

organizing the special issue he is submitting too. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The source code for the DK-VRN model is publicly available on GitHub.

Acknowledgements

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,' under grant agreement No 101003469, 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. Supplementary figures and tables

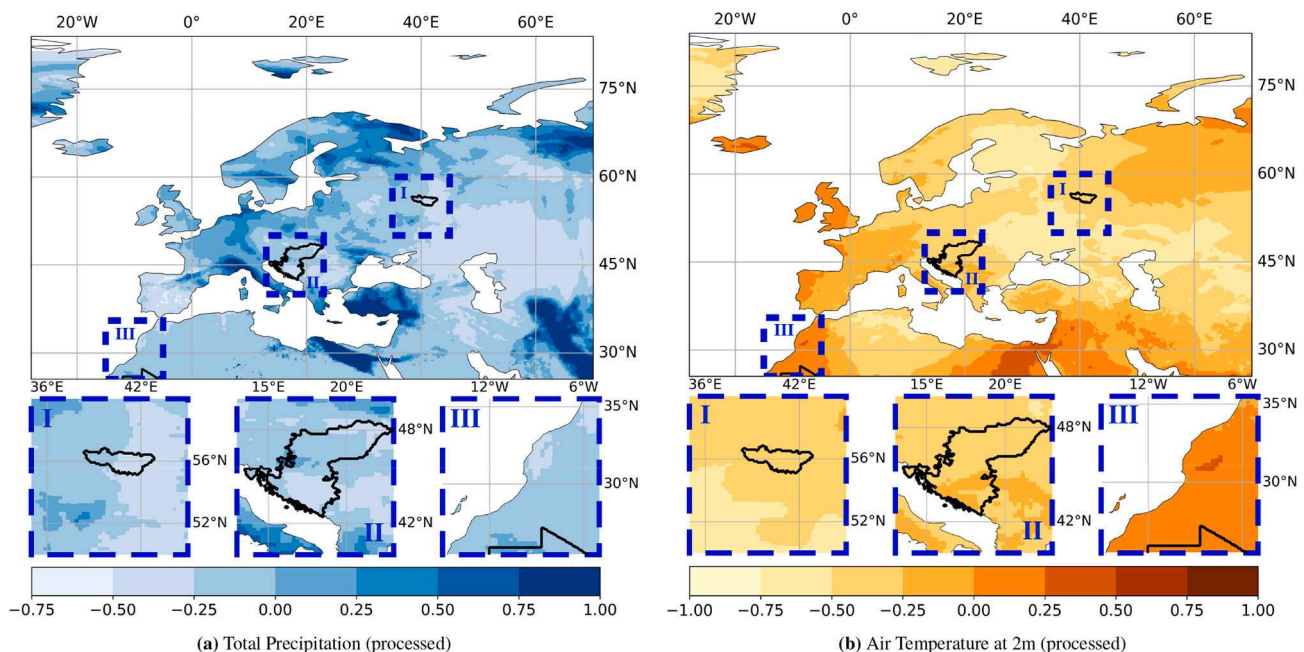


Fig. A.1. ECVs and target drought events for an example time step. Maps of Total Precipitation and Air Temperature at 2 m on 2003-04-03, processed based on the computation of monthly anomalies, are depicted in (a) and (b), respectively. The three accompanying drought events on 2003-04-03 are marked with black lines: I (Russia), II (Bosnia; Croatia; Hungary), and III (Mauritania). Regional maps of drought events are highlighted with blue dashed lines, zooming in the ECVs for each drought location. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

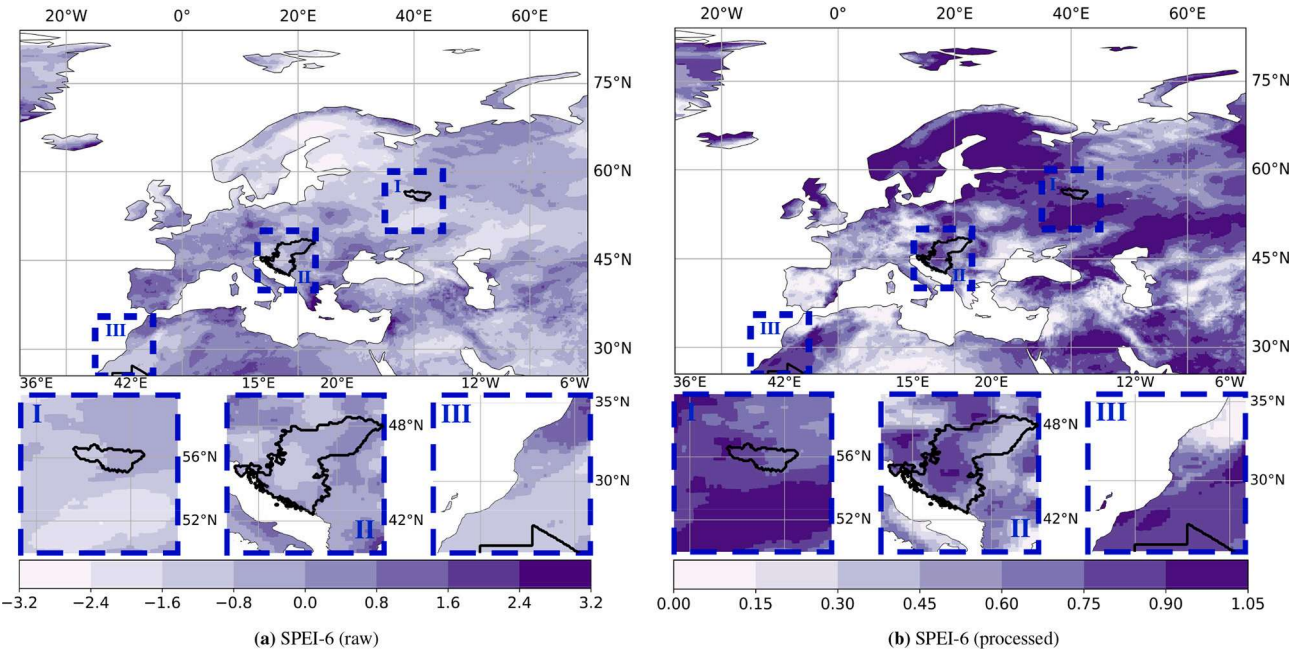


Fig. A.2. SPEI-6 and target drought events for an example time step. Maps of raw SPEI-6 and processed SPEI-6 at 2 m on 2003-04-03 are depicted in (a) and (b), respectively. The three accompanying drought events on 2003-04-03 are marked with black lines: I (Russia), II (Bosnia; Croatia; Hungary), and III (Mauritania). Regional maps of drought events are highlighted with blue dashed lines, zooming in the SPEI-6 values of each drought location. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table A.1
Spatial description of the drought events in the EM-DAT database considered for the experiments.

Country	DisNo.	Locations
Afghanistan	2006–9570	Badakhshan, Badghis, Baghlan, Balkh, Bamyán, Daykundi, Faryab, Ghor, Jawzjan, Kunduz, Samangans, Sar-e-Pul, Takhar, Uruzgan provinces
Afghanistan	2008–9475	Kunduz, Balkh, Faryab, Badghis provinces
Afghanistan	2011–9312	Balkh, Samangan, Takhar, Sar-e-Pul, Hirat, Badghis, Faryab, Jawzjan, Baghlan, Kunduz, Badakhshan, Bamyán, Daykundi, Ghor provinces
Bosnia	2003–9784	Republika Srpska province, Federacija Bosne i Hercegovine province, Federacija Bosne i Hercegovine province
Croatia	2003–9784	Bjelovar-bilogora, Dubrovnik-Neretva, Grad Zagreb, Istra, Karlovac, Kopriivnica-krizevci, Krapina-zagorje, Lika-senj, Medimurje, Kota, Osijek-baranja, Požega-slavonija, Primorje-gorski Šibenik, Sisak-moslavina, Slavonski Brod-posav, Split-dalmatija, Varazdin, Virovitica-podravina, Vukovar-srijem, Zadar-Knin provinces
Hungary	2003–9784	Bacs-kiskun, Baranya, Bekes, Borsod-abauj-zemlen, Budapest, Csongrád, Fejér, Győr-moson-sopron, Hadju-bihar, Heves, Jasz-nagykun-seolnok, Komarom-esztergom, Nógrád, Pest, Somogy, Szabolcs-szatmar-bereg, Tolna, Vas, Veszprém, Zala provinces
Italy	2012–9290	Abruzzi, Basilicata, Calabria, Campania, Emilia-romagna, Friuli-venezia Giulia, Lazio, Liguria, Lombardia, Marche, Molise, Piemonte, Puglia, Sardegna, Sicilia, Toscana, Trentino-alto Adige, Umbria, Valle D'aosta, Veneto provinces
Lithuania	2006–9411	Alytaus, Kauno, Klaipėdos, Marijampolės, Panevezio, Sialiu, Taurages, Telsiu, Utenos, Vilniaus provinces
Mauritania	2010–9292	Assaba, Brakna, Gorgol, Guidimakha, Hodh Ech Gharbi, Nouakchott, Tagant, Trarza provinces; Amourj, Bassikounou, Djigueni, Nema, Timbedra districts (Hodh Ech Chargi province) (Southern Mauritania/Sahel region)
Moldova	2012–9478	Falesti district (Balti province), Vulcanesti district (Cahul province) Chisinau, Criuleni, Ialoveni districts (Chisinau province), Ciadir-Lunga district (Gagauzia province), Basarabeasca district (Lapusna province), Orhei, Rezina, Telenesti districts (Orhei province), Ungheni district (Ungheni province)
Russian	2003–9777	Vladivostok city (Primorskiy Kray province)
Russian	2010–9318	Adygeya Rep., Astrakhanskaya Oblast, Bashkortostan Rep., Belgorodskaya Oblast, Bryanskaya Oblast, Chelyabinskaya Oblast, Ivanovskaya Oblast, Kalmykiya Rep., Kaluzhskaya Oblast, Krasnodarskiy Kray, Kurskaya Oblast, Lipetskaya, Moskovskaya, Moskva, Orenburgskaya Oblast, Orlovskaya Oblast, Penzenskaya, Rostovskaya Oblast, Ryazanskaya, Samarskaya, Saratovskaya Oblast, Smolenskaya Oblast, Stavropolskiy Kray, Tambovskaya Oblast, Tulsckaya, Tverskaya Oblast, Ulyanovskaya Oblast, Vladimirskaya, Volgogradskaya Oblast, Voronezhskaya Oblast, Yaroslavlskaya Oblast provinces
Russian	2012–9395	Adygeya Rep., Astrakhanskaya Oblast, Bashkortostan Rep., Belgorodskaya Oblast, Bryanskaya Oblast, Chelyabinskaya Oblast, Kalmykiya Rep., Krasnodarskiy Kray, Kurskaya Oblast, Lipetskaya Oblast, Name Unknown, Orenburgskaya, Orlovskaya, Penzenskaya, Rostovskaya, Samarskaya Oblast, Saratovskaya Oblast, Stavropolskiy Kray, Tambovskaya, Ulyanovskaya Oblast, Volgogradskaya Oblast, Voronezhskaya Oblast provinces
Syria	2008–9755	Damascus, Homs, Hama, Raqqah, Hassakeh provinces
Tajikistan	2008–9456	Badakhshoni Kuni, Khatlon, Sogd, Tadzhiikistan Territories provinces

Table A.2

Architecture details for the DL models considered in the experiments.

	MLP	LSTM	VMLP	VLSTM	DK-VRN	
Input	T, K_0	T, K_0	T, K_0	T, K_0	T, K_0	T, K_1
Layer 0	MLP, 32	LSTM, 32	MLP, 32	LSTM, 32	LSTM, 32	LSTM, 32
Layer 1	MLP, 32	LSTM, 32	MLP, 32	LSTM, 32	LSTM, 32	LSTM, 32
Layer 2	MLP, 32	MLP, 32	MLP, 32 + 32	MLP, 32 + 32	MLP, 32 + 32	MLP, 32 + 32
Layer 3	MLP, 1	MLP, 1	REP	REP	REP	REP
Layer 4	–	–	MLP, 1	MLP, 1	MLP, 1	LSTM, 32
Layer 5	–	–	–	–	–	MLP, K_0
Output	T, 1	T, 1	T, 1	T, 1	T, 1	T, K_1

LSTM and RNN share the same hyperparameters, while LSTM layers are replaced by RNN ones. VLSTM and VRNN share the same hyperparameters, while LSTM layers are replaced by RNN ones. The number after ‘,’ indicates the output dimension of the layer, e.g., LSTM, 32 + 32 means the output is a tuple with the dimensions of 32 and 32. ‘REP’ means the reparameterization step. The last layers are highlighted in **bold**. T is the length of the input and output time series, K_0 represents the number of climate variables used as input, while K_1 does the same for the number of drought indices ($K_0=2$ and $K_1=3$ in our experiments, respectively). Note that output [T, 1] corresponds to the classification loss, output [T, K_0] corresponds to the reconstruction loss of ECVs, and output [T, K_1] corresponds to the reconstruction loss of the multi-scalar SPEI.

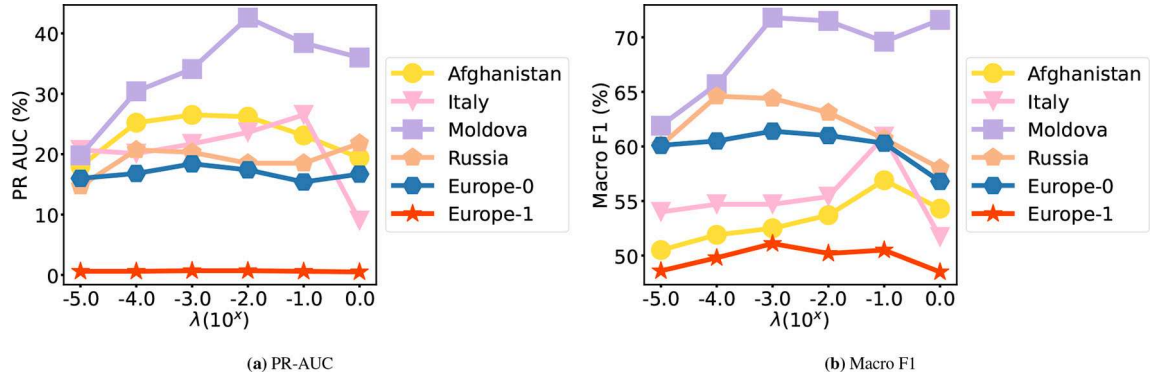


Fig. A.3. Performance of DK-VRN for different λ values and the drought events considered in this study. Afghanistan, Italy, Moldova, and Russia denote four independent drought events in the test subset. Europe-0 denotes the entire European drought-affected region. Europe-1 denotes all locations of the entire European region. (a) PR-AUC and (b) Macro F1-score results.

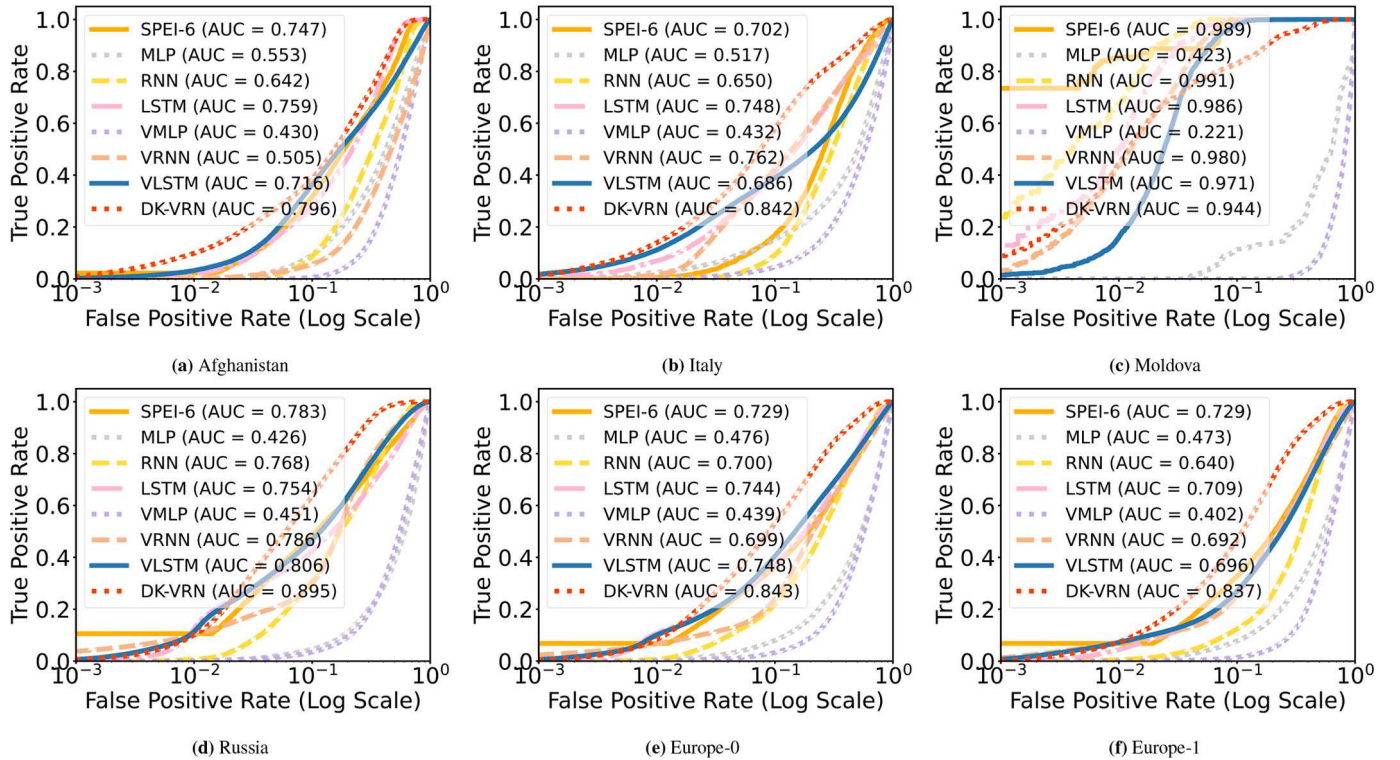


Fig. A.4. ROC curves for the different drought monitoring models in this study and the test subset.

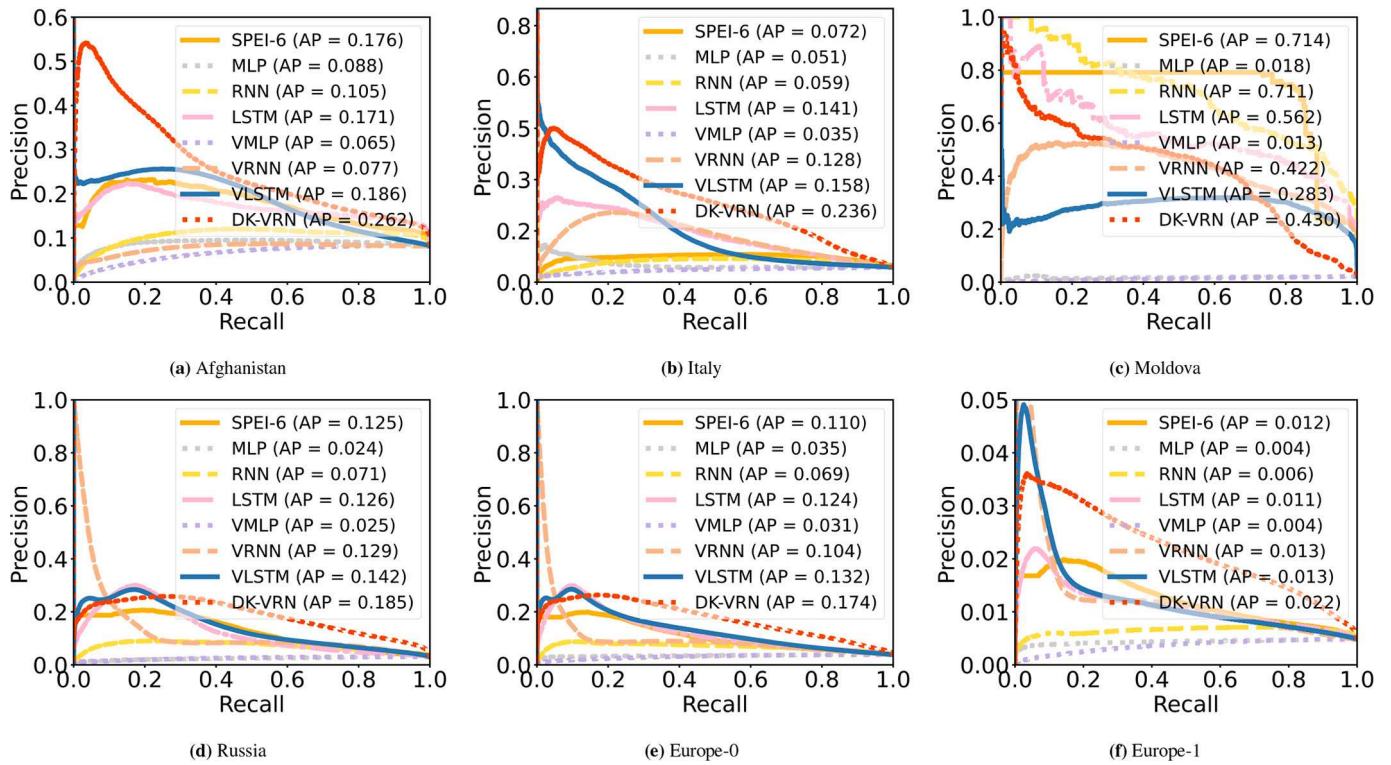


Fig. A.5. PR curves for the different drought monitoring models in this study and the test subset.

Table A.3

Training, validation, and test subsets for the climate classification evaluation. The first number denotes the number of samples, and the second number in parentheses denotes the proportion of each climate zone given all locations.

	Training	Validation	Test
4, BWh	1411 (12.80%)	662 (2.14%)	76,740 (18.51%)
5, BWk	381 (3.46%)	1202 (3.88%)	38,223 (9.22%)
6, BSh	489 (4.43%)	451 (1.46%)	4703 (1.13%)
7, BSk	1698 (15.40%)	4704 (15.18%)	43,690 (10.54%)
8, Csa	639 (5.80%)	229 (0.74%)	18,603 (4.49%)
9, Csb	8 (0.07%)	—	2701 (0.65%)
14, Cfa	335 (3.04%)	243 (0.78%)	3900 (0.94%)
15, Cfb	90 (0.82%)	26 (0.08%)	21,780 (5.25%)
16, Cfc	—	—	—
17, Dsa	330 (2.99%)	—	3419 (0.82%)
18, Dsb	613 (5.56%)	1 (0.003%)	4642 (1.12%)
19, Dsc	574 (5.21%)	—	933 (0.23%)
25, Dfa	227 (2.06%)	3701 (11.94%)	7384 (1.78%)
26, Dfb	4120 (37.37%)	19,407 (62.64%)	82,196 (19.83%)
27, Dfc	14 (0.13%)	352 (1.14%)	82,947 (20.01%)
29, ET	97 (0.88%)	5 (0.02%)	22,732 (5.48%)
30, EF	—	—	—
Total	11,026 (100.00%)	30,983 (100%)	414,593 (100%)

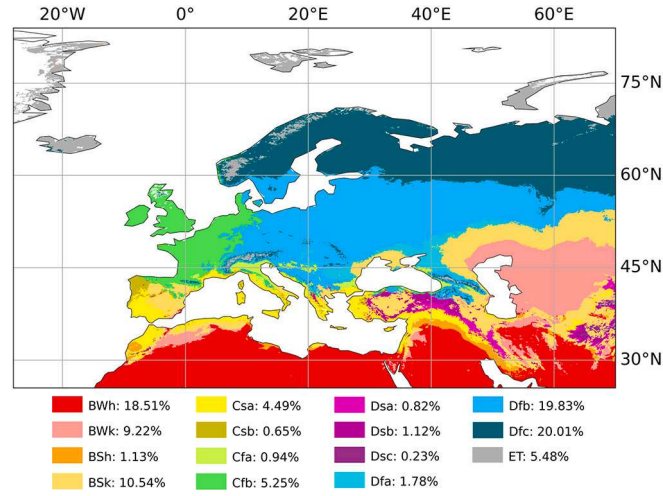


Fig. A.6. Climate classification map for the climate classification evaluation. The map shows all the different climate zones across an extended European area and the proportion of each zone given all locations, where BWh (hot desert), BWk (cold desert), BSh (hot semi-arid), BSk (cold semi-arid), Csa (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), ET (tundra). Note that there are two climate classifications Cfc (subpolar oceanic) and EF (ice cap) that do not have test samples because their corresponding locations are all non-drought. The colour map follows the official settings. A more detailed description of the Köppen-Geiger climate classification can be found in [Beck et al., 2018](#).

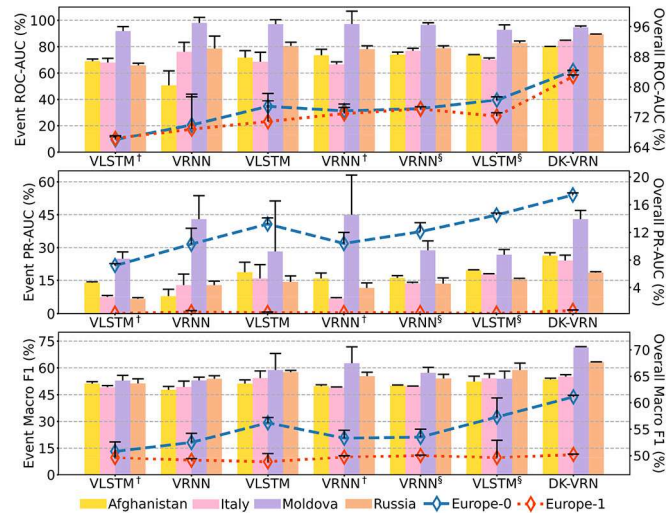


Fig. A.7. Performance when using SPEIs as input features versus as prior distribution (DK-VRN, our approach). Final results are reported by averaging the outputs of 5 independent runs under different random seeds. The error bars denote the one standard deviation. ‘VRNN[†]’ denotes using SPEIs as input features in VRNN. ‘VRNN[§]’ uses concatenated ECVs and SPEIs as input features with a multi-branch VRNN. ‘VLSTM[†]’ uses SPEIs as input features in VLSTM. ‘VLSTM[§]’ uses concatenated ECVs and SPEIs as input features with a multi-branch VLSTM. For each method, bars for all event regions (Afghanistan, Italy, Moldova, Russia) are sorted following the order of the legend of the figure. Dots are used to show the overall (Europe-0, Europe-1) scores.

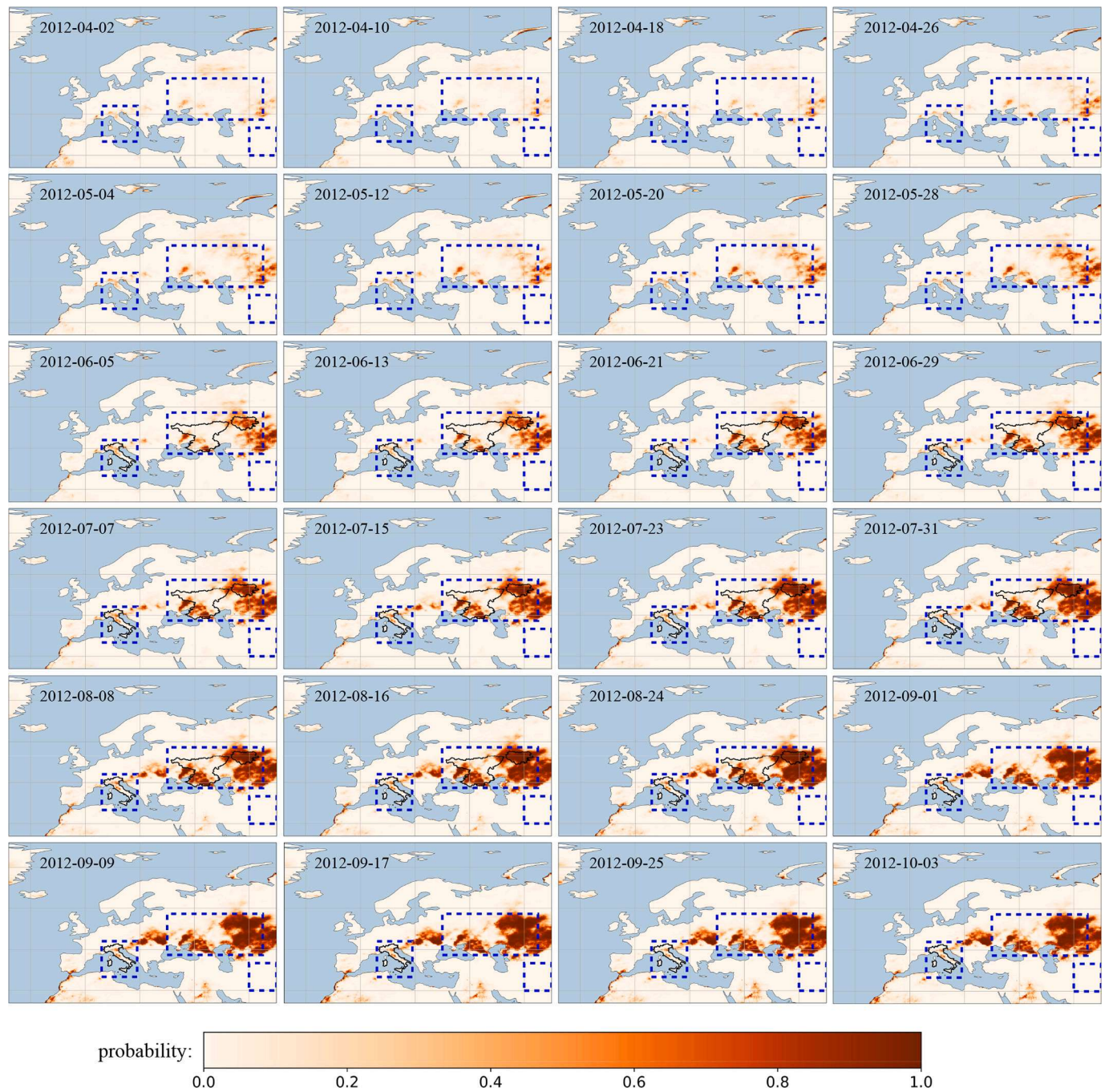


Fig. A.8. Example drought probability maps provided by DK-VRN. These maps are provided with a temporal resolution of 8 days across the extended European area and span from Apr 2012 to Sep 2012. Three drought events: I (Russia), II (Italy), and III (Afghanistan) are included. The ground truth is marked with black lines (droughts happen at a particular time step if they are shown), while regional maps for all drought events are highlighted with blue dashed lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

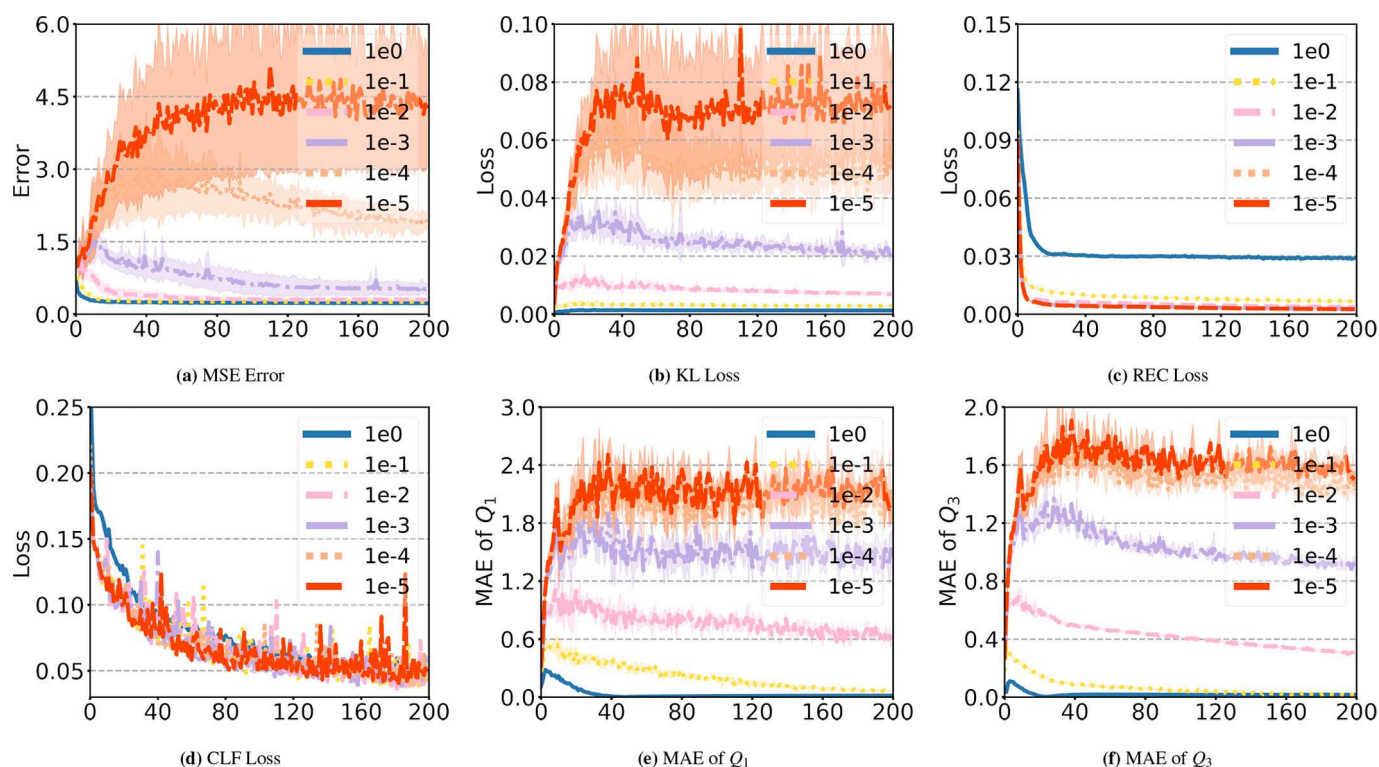


Fig. A.9. Evolution of the measurements observed during the training process for different λ values. The evolution of the validation Mean Square Error (MSE) between the SPEI generated by $P_{\lambda}(x_1|z_0)$ and the true SPEI x_1 is depicted in (a), the evolution of the different validation loss terms (KL-divergence loss, reconstruction loss, and classification loss) is plotted in (b) - (d), and the evolution of the validation Mean Absolute Error (MAE) of quartiles of the mean parameters is shown in (e) - (f).

References

- Abu-Raddad, L.J., Chemaitelly, H., Ayoub, H.H., Tang, P., Coyle, P., Hasan, M.R., Yassine, H.M., Benslimane, F.M., Al-Khatib, H.A., Al-Kanaani, Z., et al., 2022. Relative infectiousness of sars-cov-2 vaccine breakthrough infections, reinfections, and primary infections. *Nat. Commun.* 13, 1–11.
- Agboola, A., Gabriel, A., Aliyu, E., Aleso, B., 2013. Development of a fuzzy logic based rainfall prediction model. *Int. J. Eng. Technol.* 3, 427–435.
- AghaKouchak, A., Pan, B., Mazdiyasni, O., Sadegh, M., Jiwa, S., Zhang, W., Love, C., Madadgar, S., Papalexio, S., Davis, S., et al., 2022. Status and prospects for drought forecasting: opportunities in artificial intelligence and hybrid physical-statistical forecasting. *Phil. Trans. R. Soc. A* 380, 20210288.
- 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, 1–12. URL: <https://www.eea.europa.eu/data-and-maps/data/external/koppen-geiger-climate-classification>.
- Beersma, J.J., Buishand, T.A., 2004. Joint probability of precipitation and discharge deficits in the Netherlands. *Water Resour. Res.* 40.
- Belayneh, A., Adamowski, J., Khalil, B., Quilty, J., 2016. Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction. *Atmos. Res.* 172, 37–47.
- Bravar, L., Kavvas, M., 1991. On the physics of droughts. I. A conceptual framework. *J. Hydrol.* 129, 281–297.
- Campos-Taberner, M., García-Haro, J.F., Martínez, B., Izquierdo-Verdiguier, E., Atzberger, C., Camps-Valls, G., Gilbert, M.A., 2020. Understanding deep learning in land use classification from Sentinel-2 time series. *Sci. Rep.* 12 <https://doi.org/10.1038/s41598-020-74215-5>.
- Campos-Taberner, M., García-Haro, F.J., Martínez, B., Sanchez-Ruiz, S., Moreno-Martínez, A., Camps-Valls, G., Gilbert, M.A., 2023. Land use classification over smallholding areas in the European common agricultural policy framework. *ISPRS J. Photogramm. Remote Sens.* 197, 320–334. URL: <https://www.sciencedirect.com/science/article/pii/S092427162300045X>. <https://doi.org/10.1016/j.isprsjrs.2023.02.005>.
- 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.
- Caretta, M.A., Mukherji, A., Arfanuzzaman, M., Betts, R.A., Gelfan, A., Hirabayashi, Y., Lissner, T.K., Liu, J., Lopez Gunn, E., Morgan, R., Mwanga, S., Supratid, S., 2022. Water. In: Pörtner, H.O., Roberts, D.C., Tignor, M., Poloczanska, E.S., Mintenbeck, K., Alegria, A., Craig, M., Langsdorf, S., Löschke, S., Möller, V., Okem, A., Rama, B. (Eds.), *Climate Change 2022: Impacts, Adaptation and Vulnerability*. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK and New York, NY, USA, pp. 551–712.
- Cha, S.H., 2007. Comprehensive survey on distance/similarity measures between probability density functions. *City* 1, 1.
- Cortés-Andrés, J., Camps-Valls, G., Sippel, S., Székely, E., Sejdinovic, D., Diaz, E., Pérez-Suay, A., Li, Z., Mahecha, M., Reichstein, M., 2022. Physics-aware nonparametric regression models for earth data analysis. *Environ. Res. Lett.* 17, 054034.
- Dikshit, A., Pradhan, B., Huete, A., 2021. An improved spei drought forecasting approach using the long short-term memory neural network. *J. Environ. Manag.* 283, 111979.
- European Commission, 2021. Edo fraction of absorbed photosynthetically active radiation (fapar) anomaly (modis) (version 1.3.2). URL: PID: <http://data.europa.eu/89h/91a222a0-74fe-468f-b53a-b622aa1161cf>.
- Fadaei-Kermani, E., Barani, G., Ghaeini-Hessaroeyeh, M., 2017. Drought monitoring and prediction using k-nearest neighbor algorithm. *J. AI Data Min.* 5, 319–325.
- Feng, P., Wang, B., Li Liu, D., Yu, Q., 2019. Machine learning-based integration of remotely-sensed drought factors can improve the estimation of agricultural drought in South-Eastern Australia. *Agric. Syst.* 173, 303–316.
- Fensholt, R., Sandholt, I., Rasmussen, M.S., 2004. Evaluation of modis lai, fapar and the relation between fapar and ndvi in a semi-arid environment using in situ measurements. *Remote Sens. Environ.* 91, 490–507.
- Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al., 2023a. A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* 1–77.
- Gawlikowski, J., Tassi, C.R.N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al., 2023b. A survey of uncertainty in deep neural networks. *Artif. Intell. Rev.* 56, 1513–1589.
- Gitz, V., Meybeck, A., Lipper, L., Young, C.D., Braatz, S., 2016. *Climate Change and Food Security: Risks and Responses*. Food and Agriculture Organization of the United Nations (FAO). Report 110.
- González-Calabuig, M., Cortés-Andrés, J., Fernández-Torres, M.-Á., and Camps-Valls, G.: Recent Advances in Deep Learning for Spatio-Temporal Drought Monitoring, Forecasting and Model Understanding, EGU General Assembly 2022, Vienna, Austria, 23–27 May 2022, EGU22-11872, <https://doi.org/10.5194/egusphere-eg22-11872>, 2022.
- Goodfellow, I., Bengio, Y., Courville, A., 2016. *Deep Learning*. MIT Press.
- Guha-Sapir, D., Below, R., Hoyois, P., 2015. Em-dat: International Disaster Database. Université Catholique de Louvain, Brussels, Belgium.
- Hao, Z., Singh, V.P., 2015. Drought characterization from a multivariate perspective: a review. *J. Hydrol.* 527, 668–678.
- Hazaymeh, K., Hassan, Q.K., 2016. Remote sensing of agricultural drought monitoring: a state of art review. *AIMS Environ. Sci.* 3, 604–630.

- Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. *Neural Comput.* 9, 1735–1780.
- 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*, 2019.
- IPCC, 2019. Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems [P.R. Shukla, J. Skea, E. Calvo Buendia, V. Masson-Delmotte, H.-O. Pörtner, D. C. Roberts, P. Zhai, R. Slade, S. Connors, R. van Diemen, M. Ferrat, E. Haughey, S. Luz, S. Neogi, M. Pathak, J. Petzold, J. Portugal Pereira, P. Vyas, E. Huntley, K. Kissick, M. Belkacemi, J. Malley, (eds.)]. In press.
- Jiao, W., Wang, L., McCabe, M.F., 2021. Multi-sensor remote sensing for drought characterization: current status, opportunities and a roadmap for the future. *Remote Sens. Environ.* 256, 112313.
- Kapoor, A., Negi, A., Marshall, L., Chandra, R., 2023. Cyclone trajectory and intensity characterization with uncertainty quantification using variational recurrent neural networks. *Environ. Model. Softw.* 162, 105654.
- Kikon, A., Deka, P.C., 2022. Artificial intelligence application in drought assessment, monitoring and forecasting: a review. *Stoch. Env. Res. Risk A.* 36, 1197–1214.
- Kingma, D.P., Ba, J., 2014. Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kingma, D.P., Welling, M., 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.
- Kreibich, H., Van Loon, A.F., Schröter, K., Ward, P.J., Mazzoleni, M., Sairam, N., Abeshu, G.W., Agafonova, S., AghaKouchak, A., Aksoy, H., et al., 2022. The challenge of unprecedented floods and droughts in risk management. *Nature* 1–7.
- Li, J., Wang, Z., Wu, X., Xu, C.Y., Guo, S., Chen, X., Zhang, Z., 2021. Robust meteorological drought prediction using antecedent SST fluctuations and machine learning. *Water Resour. Res.* 57, e2020WR029413.
- Li, J., Wu, Z., Sheng, Q., Wang, B., Hu, Z., Zheng, S., Camps-Valls, G., Molinier, M., 2022. A hybrid generative adversarial network for weakly-supervised cloud detection in multispectral images. *Remote Sens. Environ.* 280, 113197.
- Madadgar, S., Moradkhani, H., 2014. Spatio-temporal drought forecasting within bayesian networks. *J. Hydrol.* 512, 134–146.
- Madadgar, S., AghaKouchak, A., Shukla, S., Wood, A.W., Cheng, L., Hsu, K.L., Svoboda, M., 2016. A hybrid statistical-dynamical framework for meteorological drought prediction: application to the southwestern United States. *Water Resour. Res.* 52, 5095–5110.
- 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. Dynam.* 11, 201–234.
- Mateo, A., Aduara, J., Piles, M., Muñoz-Marí, J., Pérez-Suay, A., Camps-Valls, G., 2023. Interpretable long-short term memory networks for crop yield estimation. *IEEE Geosci. Remote Sens. Lett.* 20, 2501105. <https://doi.org/10.1109/LGRS.2023.3244064>.
- McKee, T.B., Doesken, N.J., Kleist, J., et al., 1993. The relationship of drought frequency and duration to time scales. In: *Proceedings of the 8th Conference on Applied Climatology*, Boston, MA, USA, pp. 179–183.
- Mishra, A., Desai, V., 2005. Drought forecasting using stochastic models. *Stoch. Env. Res. Risk A.* 19, 326–339.
- Mishra, A.K., Singh, V.P., 2010. A review of drought concepts. *J. Hydrol.* 391, 202–216.
- Palmer, W.C., 1965. *Meteorological Drought*, vol. 30. US Department of Commerce, Weather Bureau.
- Panu, U., Sharma, T., 2002. Challenges in drought research: some perspectives and future directions. *Hydrol. Sci. J.* 47, S19–S30.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., 2019. Pytorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* 32.
- Prodhan, F.A., Zhang, J., Hasan, S.S., Sharma, T.P.P., Mohana, H.P., 2022. A review of machine learning methods for drought hazard monitoring and forecasting: current research trends, challenges, and future research directions. *Environ. Model. Softw.* 105327.
- 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. *Nature* 597, 672–677.
- Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al., 2019. Deep learning and process understanding for data-driven earth system science. *Nature* 566, 195–204.
- Rosvold, E.L., Buhaug, H., 2021. Gdis, a global dataset of geocoded disaster locations. *Sci. Data* 8, 61.
- Růuvřivka, V., Vaughan, A., De Martini, D., Fulton, J., Salvatelli, V., Bridges, C., Mateo-Garcia, G., Zantedeschi, V., 2022. Rav/En: unsupervised change detection of extreme events using ml on-board satellites. *Sci. Rep.* 12, 16939.
- Sheffield, J., Wood, E.F., 2008. Projected changes in drought occurrence under future global warming from multi-model, multi-scenario, ipcc ar4 simulations. *Clim. Dyn.* 31, 79–105.
- 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. Geoinf.* 79, 48–57.
- Slater, L., Arnal, L., Boucher, M.A., Chang, A.Y.Y., Moulds, S., Murphy, C., Nearing, G., Shalev, G., Shen, C., Speight, L., et al., 2022. Hybrid forecasting: using statistics and machine learning to integrate predictions from dynamical models. *Hydrol. Earth Syst. Sci. Discuss.* <https://doi.org/10.5194/hess-2022-334> (in review).
- Stagge, J.H., Tallaksen, L.M., Xu, C.Y., Van Lanen, H.A., 2014. Standardized precipitation-evapotranspiration index (spei): Sensitivity to potential evapotranspiration model and parameters. *Hydrology in a changing world* 367–373.
- Thornthwaite, C.W., 1948. An approach toward a rational classification of climate. *Geogr. Rev.* 38, 55–94.
- Vicente-Serrano, S.M., Beguería, S., López-Moreno, J.L., 2010. A multiscalar drought index sensitive to global warming: the standardized precipitation evapotranspiration index. *J. Clim.* 23, 1696–1718.
- Vicente-Serrano, S.M., Quiring, S.M., Pena-Gallardo, M., Yuan, S., Dominguez-Castro, F., 2020. A review of environmental droughts: increased risk under global warming? *Earth Sci. Rev.* 201, 102953.
- Vilone, G., Longo, L., 2020. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*.
- Wang, Q., Zeng, J., Qi, J., Zhang, X., Zeng, Y., Shui, W., Xu, Z., Zhang, R., Wu, X., Cong, J., 2021. A multi-scale daily spei dataset for drought characterization at observation stations over mainland China from 1961 to 2018. *Earth Syst. Sci. Data* 13, 331–341.
- West, H., Quinn, N., Horswell, M., 2019. Remote sensing for drought monitoring & impact assessment: Progress, past challenges and future opportunities. *Remote Sens. Environ.* 232, 111291.
- Wu, Z., Yin, H., He, H., Li, Y., 2022. Dynamic-lstm hybrid models to improve seasonal drought predictions over China. *J. Hydrol.* 615, 128706.
- Xu, L., Yu, H., Chen, Z., Du, W., Chen, N., Huang, M., 2023. Hybrid deep learning and s2s model for improved sub-seasonal surface and root-zone soil moisture forecasting. *Remote Sens.* 15, 3410.
- Yevjevich, V.M., 1967. Objective Approach to Definitions and Investigations of Continental Hydrologic Droughts. Ph.D. thesis. Colorado State University. Libraries.
- Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., Xu, H., Tan, W., Yang, Q., Wang, J., et al., 2020. Deep learning in environmental remote sensing: achievements and challenges. *Remote Sens. Environ.* 241, 111716.
- Zantedeschi, V., De Martini, D., Tong, C., de Witt, C.S., Kalaitzis, A., Chantry, M., Watson-Parris, D., 2020. Towards data-driven physics-informed global precipitation forecasting from satellite imagery. In: *Proceedings of the AI for Earth Sciences Workshop at NeurIPS*.
- Zerrouki, Y., Harrou, F., Zerrouki, N., Dairi, A., Sun, Y., 2020. Desertification detection using an improved variational autoencoder-based approach through etm-landsat satellite data. *IEEE J. Select. Top. Appl. Earth Observ. Rem. Sens.* 14, 202–213.
- Zhang, B., AghaKouchak, A., Yang, Y., Wei, J., Wang, G., 2019. A water-energy balance approach for multi-category drought assessment across globally diverse hydrological basins. *Agric. For. Meteorol.* 264, 247–265.
- Zhang, M., Fernández-Torres, M.A., Camps-Valls, G., 2022. Hybrid recurrent neural network for drought monitoring. In: *NeurIPS 2022 Workshop on Tackling Climate Change with Machine Learning*. URL: <https://www.climatechange.ai/papers/neurips2022/51>.
- Zhong, L., Hu, L., Zhou, H., 2019. Deep learning based multi-temporal crop classification. *Remote Sens. Environ.* 221, 430–443.