Towards Entity-Aware Conditional Variational Inference for Heterogeneous Time-Series Prediction: An application to Hydrology
Supplementary Material

Rahul Ghosh* Arvind Renganathan* Wallace McAliley† Michael Steinbach*
Christopher Duffy‡ Vipin Kumar*

1 Dataset Details

CAMELS-GB (Catchment Attributes and MEteorology for Large-sample Studies) [1] provides meteorological forcing data (e.g., precipitation, air temperature), streamflow observation, and basin characteristics.

Driver name	Description
precipitation	basin daily averaged precipitation
peti	basin daily averaged potential evapotranspiration
temperature	basin daily averaged temperature

Table 1: *Meteorological drivers used in this experiment.*

1.1 Meteorological Drivers Daily meteorological time series data are provided for the basins as summarised in Table 1.

Attribute class	Attribute name	Description
Location and topography	area	basin area
	elev_mean	basin mean drainage path slope
	dpsbar	basin mean elevation
Climatic indices	p_mean	mean daily precipitation
	pet_mean	mean daily evapotranspiration
	p_seasonality	seasonality of precipitation
	frac_snow	fraction of precipitation falling as snow
	high_prec_freq	frequency of high-precipitation days
	low_prec_freq	frequency of dry days
	high_prec_dur	average duration of high-precipitation events
	low_prec_dur	average duration of dry periods
Soil	sand_perc	percentage sand
	silt_perc	percentage silt
	clay_perc	percentage clay
	porosity_hypres	volumetric soil porosity
	conductivity_hypres	saturated hydraulic conductivity
	soil_depth_pelletier	depth to bedrock
Land cover	dwood_perc	percentage cover of deciduous woodland
	ewood_perc	percentage cover of evergreen woodland
	crop_perc	percentage cover of crops
	urban_perc	percentage cover of suburban and urban
Human influence	reservoir_cap	storage capacity of reservoirs in the basin

Table 2: *Basin attributes (entity characteristics) used in this experiment. Names and descriptions of all the attributes are available in [1].*

1.2 Basin Characteristics Basin characteristics describing the location and topography, climatic indices, soil properties, land-cover properties, and human signatures are provided for each basin as shown in Table 2.

Hyperparameter	Value
Latent dimension	16, **32**, 64, 128
Dimension of hidden state	64, **128**, 256
Batch size	100, 200
Learning rate	0.0005, 0.001, **0.003**, 0.005, 0.05
Update Learning rate	0.0005, **0.001**, 0.003, 0.005, 0.05
Update steps	1, **5**, 10, 15

Table 3: *Hyperparameter values tried with the setting denoted in **bold**.*

1.3 Hyperparameter Tuning We used grid search over a range of parameter values to find the best hyperparameters. Table 3 shows the possible parameter values that were considered. We chose the parameter set with the smallest average root mean square error (RMSE) in the training basins during the validation years as the final parameter configuration.

2 Experiment Results

2.1 Evaluation on In-sample Basins We evaluate the performance in In-Sample test, i.e., the training and testing data are from the same basins but different years. Testing data are exclusively from 1999-2009. As expected, Model Agnostic Meta-Learning using a long short-term memory (MAML_{LSTM}) does not perform well because it does not use entity characteristics or

	RMSE		R^2	
	Mean	Median	Mean	Median
MAML_{LSTM}	1.1768	0.8867	0.4893	0.7460
KGSSL	0.9278	0.6965	0.6957	0.8232
EA-CVI	0.9217	0.7077	0.7271	0.8222
CTLSTM	0.9552	0.7301	0.5674	0.8069
MAML_{CTLSTM}	0.8732	0.6297	0.7725	0.8336

Table 4: *Mean and Median RMSE and R^2 values for streamflow modeling on the benchmark datasets for EA-CVI and the baselines for the In-Sample Basins (282).*

*University of Minnesota. {ghosh128, renga016, stei0062, kumar001}@umn.edu
†U.S. Geological Survey. wmcaliley@usgs.gov
‡Penn State University. cxd111@psu.edu

prior observed streamflow data available for the diverse set of entities. Next, a long short-term memory network whose dynamic inputs are augmented with basin characteristics (CTLSTM) uses the given entity characteristics to modulate the driver-response relationship within the model. A network using MAML with CTLSTM (MAML_{CTLSTM}) further finetunes CTLSTM separately for each entity. Note that the CTLSTM models have the most amount of information provided to them. Table 4 shows that this model performs best in the In-Sample test. Both inverse modeling approaches, Entity-Aware Conditional Variational Inference (EA-CVI) and Knowledge-Guided Self-supervised Learning (KGSSL), do not use the basin characteristics that may not be completely available or may be uncertain and noisy. However, despite not utilizing the entity characteristics, these inverse modeling methods achieve comparable performance to the CTLSTM models which do have access to the entity characteristics. Overall, these findings highlight the importance of incorporating entity characteristics and utilizing the available information to improve predictive performance. However, inverse modeling methods demonstrate the potential to achieve competitive performance even without access to complete or reliable entity characteristics.

3 Reproducibility

CAMELS input data are freely available on the website of UK Centre for Ecology & Hydrology[1]. The code is available at Google Drive[2].

References

[1] Gemma Coxon et al. CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain. *Earth System Science Data*, 2020.

[1] `https://doi.org/10.5285/8344e4f3-d2ea-44f5-8afa-86d2987543a9`

[2] `https://github.com/2021rahul/Towards-Entity-Aware-Conditional-Variational-Inference-for-Heterogeneous-Time-Series-Prediction`