Multi-model ensemble benchmark data for hydrological modeling in Japanese river basins

Yohei Sawada¹, Shinichi Okugawa¹ and Takayuki Kimizuka²
¹Institute of Engineering Innovation, the University of Tokyo, Japan
²Department of Civil Engineering, the University of Tokyo, Japan

Abstract:
Verification processes of rainfall-runoff modeling are important to improve the skill of hydrological models to reproduce water cycles in river basins. It is ideal that newly developed models are compared with many benchmarking conventional models in many river basins as part of the verification process. However, this robust verification is time-consuming if model developers prepare data and models from scratch. Here we present a useful dataset which can accelerate the robust verification of hydrological models. Our newly developed dataset, Multi-model Ensemble for Robust Verification of hydrological modeling in Japan (MERV-Jp), provides runoff simulation by 44 calibrated conceptual hydrological models in 135 Japanese river basins as well as meteorological forcing which is necessary to drive conceptual hydrological models. By comparing simulated runoff with river discharge observations which are not used for the calibration of hydrological models, we find that the best models in the 44 models can reproduce observed river runoff with KGE larger than 0.6 in most of the 135 river basins, so that the runoff simulation of MERV-Jp is reasonably accurate. MERV-Jp is publicly available to support all hydrological model developers to robustly verify their model improvement.

KEYWORDS hydrological modeling; model verification; open-source dataset; multi-model ensemble

INTRODUCTION
The accurate simulation of rainfall-runoff processes in river basins is critically important in hydrology. Many methods have been proposed to accurately estimate runoff from meteorological forcing and geomorphological information of river basins. These methods can be classified into (1) fully data-driven methods which train statistical models such as neural networks by observed relationship between meteorological conditions and runoff (e.g. Hsu et al., 1995), (2) conceptual hydrological models in which rainfall-runoff dynamics is mimicked by several connected water tanks (e.g. Ishihara and Kobatake, 1979) with parameter optimization algorithms (e.g. Duan et al., 1992), (3) physically-based distributed hydrological models which are often formulated as a coupled model of a gridded land surface model and a river routing model (e.g. Wang et al., 2009).

The skill of these methods has been evaluated by comparing simulated and observed river discharge. The skill metrics such as Nash-Sutcliffe Efficiency (NSE: Nash and Sutcliffe, 1970) and Kling-Gupta Efficiency (KGE: Gupta et al., 2009) have been widely used to evaluate hydrological models. Many studies concluded that their models were improved if NSE and/or KGE were increased by their proposed modifications (e.g. Hanazaki et al., 2022; Lees et al., 2021; Gou et al., 2020; Beck et al., 2020).

It is not straightforward to verify that the improvement of a rainfall-runoff model was significant and meaningful. As shown in Figure 1a, the simplest verification is to compare the performance metrics before and after proposed modifications in one river basin. However, the performance improvement found in a single river basin is hard to be generalized to a wide variety of river basins with different geology, morphology, and climate. Many recent works verified their performance improvement in many river basins by the intensive work to compile many types of global and regional data (e.g. Hanazaki et al., 2022; Gou et al., 2020; Beck et al., 2020). In addition to the experiments in many river basins, the interpretation of the verification results can be significantly improved if models can be compared with many other conventional models. Note that modelers would not be satisfied just with the increase of skill scores. Their purpose is to attribute the model’s improvement to some scientific knowledge. This interpretation of a model’s improvement is often not straightforward. Although many model improvement works used their original model as a single benchmark, this choice of a benchmark is often problematic for the interpretation of a model’s improvement. One may suspect that the original model is so bad and uncalibrated that the revised model can gain substantial improvement. One may also be able to criticize that even a revised model does not have enough accuracy. It is difficult for model developers to objectively respond to these criticisms when both benchmark and proposed models are developed by themselves. They need an independent benchmark which is generated by a transparent process to show that their original and/or revised models are reasonably accurate considering the current standards of hydrological modeling. Lees et al. (2021) performed the simulation of 4 conceptual hydrological models as a benchmark to...
comprehensively evaluate their newly developed neural network-based rainfall-runoff model. In addition to their successful comparison of several Long-Short-Term-Memory (LSTM)-based methods, they clearly showed that these LSTM-based hydrological models outperformed the conventional hydrological models in most of the 669 river basins. They successfully demonstrated the usefulness of their LSTM methods as well as the implications to improve the existing conventional hydrological models by providing independent benchmark data.

The evaluation with many benchmarking conventional models in many river basins is ideal (Figure 1b), and datasets which can accelerate the model evaluation in this manner are useful. There have been many useful datasets which offer the potential for large sample modeling studies. For example, the series of Catchment Attributes and MEteorology for Large-Sample studies (CAMELS) provide daily datasets which are primarily necessary to drive conceptual hydrological models in the United States (Addor et al., 2017), Chile (Alvarez-Garreton et al., 2018), Brazil (Chagas et al., 2020), and Great Britain (Coxon et al., 2020). However, no multi-model ensemble simulation of hydrological models is provided in these datasets. As Lees et al. (2021) did, users need to run conventional models independent of their own developed models to perform the robust verification shown in Figure 1b.

Here we aim at providing a useful dataset to help model developers who would like to perform the model verification shown in Figure 1b. We provide the dataset of meteorological forcing and river discharge in 135 Japanese river basins. In addition, we provide the results of runoff simulation by 44 calibrated conceptual hydrological models in these 135 river basins. All the data are publicly available. This paper is a data descriptor of our dataset called Multi-model Ensemble for Robust Verification of hydrological modeling in Japan (MERV-Jp). In the next section, we present the method to generate MERV-Jp. In the DATA RECORDS section, the detailed structure of MERV-Jp is explained. The validation of MERV-Jp can be found in the VALIDATION section. In the last section, conclusions and discussion can be found.

METHODS

Data

For precipitation and temperature, Asian Precipitation Highly-Resolved Observational Data Integration Towards Evaluation of Extreme Events (APHRODITE) was used. Gridded daily precipitation of APHRO_JP V1207 (Kamiguchi et al., 2010) was used to estimate basin-averaged daily precipitation. Original spatial and temporal resolutions of APHRO_JP V1207 were 0.05 degree and daily, respectively. All gridded meteorological data are spatially averaged to get basin-averaged data in this study since our conceptual hydrological models calculate basin-averaged runoff from basin-averaged meteorological forcing. Basin-averaged precipitation was calculated as a weighted average of the gridded data. Gridded daily air temperature of APHRO_MA V1808 (Yasutomi et al., 2011) was used to estimate basin-averaged daily temperature. Original spatial and temporal resolutions of APHRO_MA V1808 were 0.25 degree and daily, respectively. Basin-averaged temperature was calculated as a weighted average of the gridded data. The series of APHRODITE can be downloaded at http://aphrodite.st.hirosaki-u.ac.jp/index.html.

For potential evapotranspiration, hourly Potential Evapotranspiration data (hPET; Singer et al., 2021) were used. hPET provides hourly and daily potential evapotranspiration at 0.1-degree grid resolution for the global land surface. We used daily data and calculated basin-averaged potential evapotranspiration by performing the weighted average of the gridded data. hPET can be downloaded at https://data.bris.ac.uk/data/dataset/qb8ujazzda0s2aykkv0oq0ctp (Singer et al., 2020).

River discharge observations in 135 Japanese river basins were extracted from the Global Runoff Data Centre (GRDC: https://www.bafg.de/GRDC/EN/Home/homepage_node.html). Daily river discharge data from 1993 to 2003 were used to calibrate and validate hydrological models. The river basin mask data in GRDC were also used to calculate basin-averaged meteorological forcing.

Model

We used Modular Assessment of Rainfall-Runoff Models Toolbox (MARRMoT; Knoben et al., 2019a). MARRMoT is a modular open-source toolbox containing 46 existing conceptual hydrological models. We excluded SMAR (ID: 40) and PRMS (ID: 45) models from 46 models in MARRMoT since their implicit Euler scheme became unstable and could not complete the computation in some of the river basins in our computational environment. The remaining 44 models were used to generate a benchmark multi-model runoff simulation in 135 river basins. MARRMoT considers a wide variety of existing conceptual hydrological models, so that the benchmark generated by this tool can be expected to represent the diversity of the performance. In our 44 models, they use between 1 and 8
stores and between 1 and 23 parameters. Every model has soil moisture stores. Routing stores are included in 17 models, groundwater stores in 11 models, snow stores in 11 models, and interception stores in 9 models. The details of MARRMoT can be found in Knoben et al. (2019a). The source code can be downloaded at https://github.com/wknoben/MARRMoT. The conceptual models are simple and easy to interpret and can be driven by transparent processes. Therefore, they are appropriate to generate the open-source benchmark in this study. Despite the limited output types compared to physical-based distributed hydrological models, all conceptual models in MARRMoT can generate daily river discharge, which is the primary output of all types of hydrological models. Physical-based distributed hydrological models are also useful to generate a benchmark. Although it is practically difficult to provide many multi-model ensembles of physical-based models, the inclusion of some physical-based hydrological models to MERV-Jp may contribute to maximizing the potential of this dataset, which will be the future work of this study.

We drove the 44 models in MARRMoT using daily precipitation, temperature, and potential evapotranspiration data in 135 river basins. In each river basin, we optimized unknown parameters of each model by comparing the differences between simulated and observed runoff from 1993 to 2000. The differences between simulation and observation were evaluated by KGE. We searched the set of parameters which can maximize KGE by the downhill simplex method (Nelder and Mead, 1965). We run models from 1993 to 2000 as a spin-up and then run them again for the same period to evaluate KGE in every iteration of the parameter optimization. After this calibration was completed, we drove the models once again from 1993 to 2003. The time series of runoff from this final simulation were provided as a dataset and can be used as a benchmark. Note that the period of 2001–2003 can be recognized as a validation period of the calibrated hydrological models since observation data in this period were not used to tune model parameters. We will show the performance of MERV-Jp in this validation period. We did not use NSE for the parameter optimization. It has been found that the optimization based on mean-square-error based metrics such as NSE is vulnerable to outliers and tends to overfit to calibration data (Gupta et al., 2009), although the selection of appropriate metrics is still controversial (e.g. Knoben et al., 2019b). We showed NSE in the validation period as the supporting information although NSE was not used for the model calibration.

Study area

Figure 2 shows the distribution of 135 river basins in which we provide basin-averaged meteorological forcing and runoff simulation. Since conceptual hydrological models in MARRMoT calculate basin-averaged runoff, it should be interpreted that we estimate river discharge at the downstream end of each river basin. We chose all river basins in which river discharge observations in GRDC were available from 1993 to 2003. Note that many river basins are fully included in larger river basins, and we recognized these river basins as independent river basins in MERV-Jp. In other words, when there are multiple observation points in a river, we solved runoff at all the observation points by individually building conceptual hydrological models of each catchment for observed points. When users solve a whole river basin by distributed hydrological models and simultaneously get runoff results at many observation points, they can simultaneously evaluate all of them using MERV-Jp.

DATA RECORDS

MERV-Jp can be downloaded at Zenodo (https://zenodo.org/record/6626268; Sawada and Okugawa, 2022). MERV-Jp contains 135 csv files. Each one of these files corresponds to data of each one of 135 river basins. Each csv file includes daily time series of:

(i) flag of period: ‘0’ for the calibration period and ‘1’ for the validation period
(ii) date (Year/Month/Day)
(iii) watershed-averaged precipitation (mm/day)
(iv) watershed-averaged air temperature (degree)
(v) watershed-averaged potential evapotranspiration (mm/day)
(vi) Observed runoff rate (mm/day)
(vii) Simulated runoff rates (mm/day) of 44 models

In addition, we have supporting information for the list of river basins, the river basin’s map, and a readme file.

VALIDATION

In this section, we show the overall performance of MERV-Jp in the validation period to demonstrate that MERV-Jp is appropriate to be used as a benchmark for comparison of hydrological models. Figure 3 shows maximum, median, and minimum KGE in the 44 models for the
135 river basins. Figure 3a indicates that the best models in the 44 models can reproduce observed river runoff with KGE larger than 0.6 in most of the river basins. Although there is no absolute KGE criteria for “good” or “bad” models, KGE larger than 0.6 has been recognized as a signal of skillful models in the literature (see Knoben et al., 2019b for the comprehensive review of the usage of KGE to determine if hydrological models are accurate or not). Therefore, MERV-Jp includes at least one skillful model in most of the river basins. Figures 3b and 3c show that there is a diversity of skills in the 44 models. This diversity is useful to interpret benchmarking of a user’s model. According to the median of KGE shown in Figure 3b, the ensemble of the 44 models tends to have poor skill in snowy river basins. This is because many conceptual hydrological models in MARRMoT lack or overly simplify modeling of snow processes. Figure S1 in the supporting information shows the distribution of KGE in individual models. Many models without snow modules such as IHACRES (ID: 5) perform poorly in the snowy river basins while they are reasonably accurate in the snow-free river basins.

Figure 4a shows the performance of each hydrological model for all 135 river basins. Variances expressed by the boxplot reveal the difference of the performance of a single model in different river basins. Users can perform model verification shown in Figure 1b by adding performance metrics of their own model to this group of boxplots. Figure 4a indicates that in most of the models KGE is larger than 0.6 in the majority of the river basins. Some models such as FLEX-IS (ID: 34) perform exceptionally better with smaller variances of KGE than the others. Rela-
tively poorly performing models such as Traditional bucket model (ID: 1), TOPMODEL (ID: 14) and GSFB (ID: 20) still achieve reasonably high KGE considering the standards proposed in the published literature (e.g. Knoben et al., 2019b). All the data in MERV-Jp can be used as a reasonable benchmark.

In Figure 4b, each boxplot shows KGE in each river basin, and variances expressed by the boxplot reveal the difference of the performance in different models. As we discussed in Figure 3, it is relatively difficult for most of the models to accurately simulate runoff in snowy river basins such as Kurobe (no. 76) and Sho (no. 77). The findings shown in this section do not change when the performance is evaluated by NSE (Figure S2).

CONCLUSIONS AND DISCUSSION

Here we developed a new dataset which supports developers of hydrological models to robustly evaluate the skill of their models to simulate river runoff. Our MERV-Jp is the publicly available datasets of meteorological forcing and simulated runoff by 44 calibrated hydrological models in 135 Japanese river basins. Here we summarize the potential of MERV-Jp to contribute to hydrological modeling.

First, MERV-Jp provides basin-averaged meteorological forcing data to drive conceptual hydrological models. Users of MERV-Jp can omit many parts of data preprocessing to develop hydrological models in their research activities, which may be beneficial especially for those who are not trained as hydrologists. While our forcing data (precipitation, temperature, and potential evapotranspiration) are insufficient to drive many physically based distributed hydrological models, MERV-Jp’s meteorological data can also be used to develop fully data-driven machine learning-based hydrological models. MERV-Jp makes it easier to prepare input data for rainfall-runoff models in many Japanese river basins. Although CAMELS datasets have already realized such a large sample study (Addor et al., 2017; Alvarez-Garreton et al., 2018; Chagas et al., 2020; Coxon et al., 2020), to the best of our knowledge, there has been no such kind of datasets in Japan or Asian monsoon regions.

Second, the multi-model ensemble of simulated runoff by 44 different models is a useful benchmark to robustly evaluate the improvement of hydrological models. By comparing users’ new models and MERV-Jp, users may clarify the significance of their proposed improvement. In addition, the interpretation of the evaluation can be improved by comparing the proposed model with a wide variety of conventional hydrological models in a wide variety of river basins in Japan. For instance, if model developers use a single hydrological model as a benchmark, one may suspect that the selection of a benchmark is biased to exaggerate the performance of their new model. Users of MERV-Jp can easily respond to this criticism by simultaneously using a wide variety of conceptual hydrological models as a benchmark. When users of MERV-Jp would like to show that their developed models outperform conventional models, they need to just show the better scores of their model than the best model in MERV-Jp. In the case that their model is not much better than the existing models but their model improvement is scientifically meaningful, they may be able to show that the performance scores of their

Figure 4. Performance of 44 models in 135 river basins. (a) Boxplots of KGE classified by 44 models. Variances expressed by the boxplots reveal the difference of the performance of a single model in different river basins. (b) Boxplots of KGE classified by 135 river basins. Variances expressed by the boxplots reveal the difference of the performance in different models. Note that extremely small KGE outliers are not shown.
model is within the range of scores of all models in MERV-Jp to reveal that their model is at least reasonably accurate. Therefore, MERV-Jp can support a wide variety of model evaluation activities.

Third, MERV-Jp can be used to train data-driven rainfall-runoff predictions. Recently, combining process-based models and machine-learning models has been actively investigated. For example, Pathak et al. (2018) applied reservoir computing, one of the variants of neural networks, to predict non-linear systems. They found that the skill of reservoir computing was significantly improved if they included the output of process-based models as input data of reservoir computing even though the process-based model is imperfect. This method has been applied to numerical weather prediction (Arcomano et al., 2022). Tomizawa and Sawada (2021) applied sequential data assimilation to a process-based model and this output of data assimilation was learnt by reservoir computing. The skill of this reservoir computing to predict non-linear systems is much better than the purely data-driven reservoir computing even though the process-model is imperfect. The hybrid of conceptual hydrological models in MERV-Jp and machine-learning models may be able to improve the skill to estimate river discharge. Simulated and observed river discharge as well as meteorological forcing in MERV-Jp may be used to innovate rainfall-runoff modeling based on the emerging technique of machine learning.

There are four major limitations in our MERV-Jp. First, the data period of MERV-Jp (i.e. 1993–2003) is short. This is because GRDC does not archive long-term river discharge observations in Japan. Future work will extend the period using river discharge observation archived by the Japanese government. Second, since MERV-Jp is generated by conceptual hydrological models, we cannot explicitly consider the effects of artificial flow regulation, which may degrade the performance of MERV-Jp. Third, users of MERV-Jp need to use the same meteorological forcing data as MERV-Jp to isolate the difference between hydrological models’ performances from the difference of simulated river discharge between a user’s model and MERV-Jp. This issue gets severe when users’ models are physical-based distributed hydrological models whose input data cannot be perfectly the same as MERV-Jp. In this case, it is necessary to compare models as the coupled systems of meteorological data and a hydrological model. Even in the case that users drive their hydrological model by different input data from MERV-Jp, they may have the original model (“your previous model” in Figure 1b) as a benchmark whose meteorological forcing is shared with their evaluated model and our MERV-Jp works as a supporting dataset to verify that their original benchmark is appropriate (see the Introduction). Fourth, the daily temporal resolution of MERV-Jp is inappropriate for detailed event-based evaluation of flood simulation since it is necessary to obtain hourly runoff simulation for this purpose. Note that most of the large-sample model evaluation activities used daily runoff even when they used a physical-based distributed hydrological model (e.g. Hanazaki et al., 2022), and as such our daily data of MERV-Jp can be used for a wide variety of studies on hydrological model evaluation.

ACKNOWLEDGMENTS

We thank two anonymous reviewers for their constructive comments. This study was supported by Foundation of River & basin Integrated CommunicationS (FRICS) and JST FOREST program (grant no. JPMJFR205Q).

SUPPLEMENTS

Figure S1. Distribution of KGE in the validation period for all 44 conceptual hydrological models Figure S2. (a) Boxplots of NSE classified by 44 models. Variance expressed by the boxplots reveal the difference of the performance of a single model in different river basins. (b) Boxplots of NSE classified by 135 river basins. Variance expressed by the boxplots reveal the difference of the performance in different models. Note that extremely small NSE outliers are not shown.

REFERENCES

Addor N, Newman AJ, Mizukami N, Clark MP. 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. Hydrology and Earth System Sciences 21: 5293–5313. DOI: 10.5194/hess-21-5293-2017.

Alvarez-Garreton C, Mendoza PA, Bosier JP, Addor N, Galleguillos M, Zambrano-Bigiarini M, Lara A, Puelma C, Cortes G, Garreaud R, McPhee J, Ayala A. 2018. The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies – Chile dataset. Hydrology Earth System Sciences 22: 5817–5846. DOI: 10.5194/hess-22-5817-2018.

Arcomano T, Szunyogh I, Wikner A, Pathak J, Hunt BR, Ott E. 2022. A hybrid approach to atmospheric modeling that combines machine learning with a physics-based numerical model. Journal of Advances in Modeling Earth Systems 14: e2021MS002712. DOI: 10.1029/2021MS002712.

Beck HE, Pan M, Lin P, Seibert J, van Dijk ADJ, Wood EF. 2020. Global fully distributed parameter regionalization based on observed streamflow from 4,229 headwater catchments. Journal of Geophysical Research: Atmospheres 125: e2019JD031485. DOI: 10.1029/2019JD031485.

Chagas VBP, Chaffe PLB, Addor N, Fan FM, Fleischmann AS, Paiva RCD, Siqueira VA. 2020. CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil. Earth System Sciences Data 12: 2075–2096. DOI: 10.5194/essd-12-2075-2020.

Coxon G, Addor N, Bloomfield J, Freer J, Fry M, Hannaford J, Howden N, Lane R, Lewis M, Robinson E, Wagener T, Woods R. 2020. Catchment attributes and hydro-meteorological timeseries for 671 catchments across Great Britain (CAMELS-GB), NERC Environmental Information Data Centre [data set]. DOI: 10.5285/8344e4f3-d2ea-44f5-8afa-86d2987545a9.

Duan QY, Sorooshian S, Gupta V. 1992. Effective and efficient global optimization for conceptual rainfall-runoff models. Water Resources Research 28: 1015–1031. DOI: 10.1029/91wr02985.

Gou J, Miao C, Duan Q, Tang Q, Di Z, Liao W, Wu J, Zhou R.
2020. Sensitivity analysis-based automatic parameter calibration of the VIC model for streamflow simulations over China. *Water Resources Research* **56**: 1–19. DOI: 10.1029/2019WR025968.

Gupta HV, Kling H, Yilmaz KK, Martinez GF. 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology* **377**: 80–91. DOI: 10.1016/j.jhydrol.2009.08.003.

Hanazaki R, Yamazaki D, Yoshimura K. 2022. Development of a reservoir flood control scheme for global flood models. *Journal of Advances in Modeling Earth Systems* **14**: e2021MS002944. DOI: 10.1029/2021MS002944.

Hsu K, Gupta HV, Sorooshian S. 1995. Artificial neural network modeling of the rainfall-runoff process. *Water Resources Research* **31**: 2517–2530. DOI: 10.1029/95WR01955.

Kamiguchi K, Arakawa O, Kitoh A, Yatagai A, Hamada A, Yasutomi N. 2010. Development of APHRO JP, the first Japanese high-resolution daily precipitation product for more than 100 years. *Hydrological Research Letters* **4**: 60–64. DOI: 10.3178/hrl.4.60.

Knoben WJM, Freer JE, Fowler KJA, Peel MC, Woods RA. 2019a. Modular Assessment of Rainfall-Runoff Models Toolbox (MARRMoT) v1.2: An open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations. *Geoscientific Model Development* **12**: 2463–2480. DOI: 10.5194/gmd-12-2463-2019.

Knoben WJM, Freer JE, Woods RA. 2019b. Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences* **23**: 4323–4331. DOI: 10.5194/hess-23-4323-2019.

Lees T, Buechel M, Anderson B, Slater L, Reece S, Coxon G, Dadson SJ. 2021. Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models. *Hydrology and Earth System Sciences* **25**: 5517–5534. DOI: 10.5194/hess-25-5517-2021.

Nash JE, Sutcliffe JV. 1970. River flow forecasting through conceptual models part I – A discussion of principles. *Journal of Hydrology* **10**: 282–290. DOI: 10.1016/0022-1694(70)90255-6.

Nelder JA, Mead R. 1965. A simplex method for function minimization. *The Computer Journal* **7**: 308–313. DOI: 10.1093/comjnl/7.4.308.

Pathak J,Wikner A, Fussell R, Chandra S, Hunt BR, Girvan M, Ott E. 2018. Hybrid forecasting of chaotic processes: Using machine learning in conjunction with a knowledge-based model. *Chaos* **28**: 1–9. DOI: 10.1063/1.5028373.

Sawada Y, Okugawa S. 2022. Multi-model Ensemble for Robust Verification of hydrological modeling in Japan (MERV-Jp) (1.0) [Data set]. Zenodo. DOI: 10.5281/zenodo.6626268.

Singer MB, Asfaw DT, Rosolem R, Cuthbert MO, Miralles DG, Migaitama DA, MacLeod D, Michaelides K. 2020. Hourly potential evapotranspiration (hPET) at 0.1° grid resolution for the global land surface from 1981-present [Data set]. DOI: 10.5523/bris.qb8uajzzda0s2aykkv0oqkltp.

Singer MB, Asfaw DT, Rosolem R, Cuthbert MO, Miralles DG, Macleod D, Quichimbo EA, Michaelides K. 2021. Hourly potential evapotranspiration at 0.1° resolution for the global land surface from 1981-present. *Scientific Data* **8**: 224. DOI: 10.1038/s41597-021-01003-9.

Tomizawa F, Sawada Y. 2021. Combining ensemble Kalman filter and reservoir computing to predict spatiotemporal chaotic systems from imperfect observations and models. *Geoscientific Model Development* **14**: 5623–5635. DOI: 10.5194/gmd-14-5623-2021.

Wang L, Koike T, Yang K, Yeh PJF. 2009. Assessment of a distributed biosphere hydrological model against streamflow and MODIS land surface temperature in the upper Tone River Basin. *Journal of Hydrology* **377**: 21–34. DOI: 10.1016/j.jhydrol.2009.08.005.

Yasutomi N, Hamada A, Yatagai A. 2011. Development of a long-term daily gridded temperature dataset and its application to rain/snow discrimination of daily precipitation. *Global Environmental Research* **15**: 165–172.