AGU ADVANCING EARTH AND SPACE SCIENCE

Water Resources Research

RESEARCH ARTICLE
10.1029/2021WR031548

Special Section:
Advancing process representation in hydrologic models: Integrating new concepts, knowledge, and data

Toward a Multi-Representational Approach to Prediction and Understanding, in Support of Discovery in Hydrology

Luis A. De la Fuente[1], Hoshin V. Gupta[1], and Laura E. Condon[1]

[1]Department of Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, AZ, USA

Abstract Key to model development is the selection of an appropriate representational system, including both the representation of what is observed (the data), and the formal mathematical structure used to construct the input-state-output mapping. These choices are critical, because they completely determine the questions we can ask, the nature of the analyses and inferences we can perform, and the answers we can obtain. Accordingly, a representation that is suitable for one kind of investigation might be limited in its ability to support some other kind. Arguably, how different representational approaches affect what we can learn from data is poorly understood. This paper explores three representational strategies as vehicles for understanding how catchment scale hydrological processes vary across hydro-geo-climatologically diverse Chile. Specifically, we test a lumped water-balance model (GR4J), a data-based dynamical systems model (LSTM), and a data-based regression tree model (Random Forest). Insights were obtained regarding system memory encoded in data, spatial transferability by use of surrogate attributes, and informational deficiencies of the data set that limit our ability to learn an adequate input-output relationship. As expected, each approach exhibits specific strengths, with LSTM providing the best characterization of dynamics, GR4J being the most robust under informationally deficient conditions, and Random Forest regression-tree method being most supportive of interpretation. Overall, the contrasting nature of the three approaches suggests the value of adopting a multi-representational framework to more fully extract information from the data and, by doing so, find information that better facilities the goals of robust prediction and improved understanding, ultimately supporting enhanced scientific discovery.

Plain Language Summary The representations we use when analyzing data and modeling systems *completely* determine the questions we can ask, the nature of the analyses and inferences we can perform, and the answers that we can obtain. So, any given modeling approach may be highly suitable for learning certain things about a system but be completely unsuitable for learning other things. To explore how different representational approaches can affect what we can learn from data, we explore how three different modeling approaches (one physical-conceptual lumped water balance method and two machine-learning methods) can support an improved understanding of how catchment-scale hydrological processes vary across the diverse hydro-geo-climatology of Chile. Each approach was found to exhibit specific strengths, and interesting insights were obtained regarding hydrological memory, attributes that correlate with transferability across different regions, and informational deficiencies of the available data set. Overall, this study suggests the value of adopting a general multi-representational framework to facilitate robust prediction and improved understanding, in support of scientific discovery in the Earth and Environmental Sciences.

Key Points:

- The use of multiple representational approaches improves our ability to understand and discover aspects of the Data Generating Process (DGP)
- Evidence was found for short and long hydrological memory, log-linear performance with aridity, and inadequate informativeness of the data
- Multi-representational approaches, with different structures and assumptions, better characterize "*what we know about what we do not know*"

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

L. A. De la Fuente,
ldelafue@email.arizona.edu

Citation:

De la Fuente, L. A., Gupta, H. V., & Condon, L. E. (2023). Toward a multi-representational approach to prediction and understanding, in support of discovery in hydrology. *Water Resources Research*, *59*, e2021WR031548. https://doi.org/10.1029/2021WR031548

Received 4 NOV 2021
Accepted 23 DEC 2022

Author Contributions:

Conceptualization: Luis A. De la Fuente, Hoshin V. Gupta
Formal analysis: Luis A. De la Fuente
Investigation: Luis A. De la Fuente
Methodology: Luis A. De la Fuente, Hoshin V. Gupta
Supervision: Hoshin V. Gupta, Laura E. Condon
Writing – original draft: Luis A. De la Fuente

1. Introduction

"*Chance favors the prepared mind (Louis Pasteur)*"

1.1. The Problem of Selecting an Appropriate Representational System

Key to the development of any dynamical systems model, be it conceptual or data-based, is the selection of an appropriate representational system. This includes two aspects: (a) the choice of relevant inputs (drivers) and boundary conditions, and (b) the formal mathematical/algorithmic structure used to construct the input-state-output mappings that are hypothesized to characterize the system (Gharari et al., 2021; Gupta et al., 2012).

Clearly, the selection of inputs and boundary conditions determines the nature and quality of the information that can be brought to bear on the prediction problem because without adequate and informationally relevant data, the

Writing – review & editing: Luis A. De la Fuente, Hoshin V. Gupta, Laura E. Condon

task of predicting the system outputs is doomed from the outset. Having done so, the mathematical/algorithmic representational system selected for constructing the input-state-output mapping is critical, because it completely determines the questions we can ask, the nature of the analyses and inferences we can perform, and the answers we can obtain.

For example, a theory-based physical-conceptual (PC) type of representation is typically constructed to answer questions such as "what kind and magnitude of streamflow response can we expect to see when a specific catchment system is perturbed by a certain sequence of rainfall (and temperature) inputs?." Within this representational system, whereas a lumped bucket water-balance architecture can be used to obtain insights into aggregate soil moisture storage variations, a spatially distributed architecture is necessary to infer the dynamic evolution of soil moisture (and other state-variables and fluxes) in three dimensional space. Such representations typically focus on preserving and tracking mass and energy (sometimes also momentum) flow through the system.

On the other hand, data-based machine learning (ML) types of representation are focused on preserving and tracking *information* flows through the system (e.g., about previous states of a catchment). Such representations may not be as well suited to directly inferring latent variables such as soil moisture state or fluxes such as percolation, recharge, and interflow that are constrained to obey conservation principles, unless appropriate regularization constraints are also implemented. Further, while recurrent neural network methods facilitate the explicit representation of Markovian-like memory processes, regression tree methods better facilitate an exploration of explanatory variable importance.

Consequently, we can expect, a priori, that alternative representational strategies may provide different perspectives on the factors and processes governing the generation of system behaviors. To avoid semantic confusion, please note that the term "representation" refers to the underlying mathematical and structural principles used in the construction of a "modeling system," whereas several alternative model architectural hypotheses (model structures) can be constructed within a single representational system (e.g., as in the FUSE and SUMMA hypothesis testing approaches developed by Clark et al., 2011; Clark et al., 2015a, 2015b, and 2015c). As such, the choice of representational system is more fundamental than that of the modeling system, and we discuss the implications of this distinction in more detail later.

1.2. Representations as Complementary Perspectives on Reality

For any given application, it can be challenging to determine what the most appropriate representational system for model development. In hydrology, as in other fields, this situation has led to the availability of a very large variety of models, each based on different assumptions (and even philosophies), and often having been tested under different (sometimes very specific) conditions. This diversity of approaches recalls the classic story of the "*blind people and the elephant*" where each person's interpretation is limited, both by their experience being based on some very specific aspect of the animal, and also by their ability to map that experience onto their previous knowledge (i.e., they are limited by what they can recognize).

So, rather than asking which representational approach is (somehow) "the best," one might instead consider whether the multiple perspectives offered by different representations can provide information that can be used to develop a better overall understanding of the system under investigation. By taking a multi-representational perspective, within which each interpretation of the system is deemed to be valuable, we can hope to make progress towards a deeper understanding of the underlying Data Generating Process (DGP); that is, towards ultimately discovering the real nature of the physical process that gives rise to the phenomena that we can observe and obtain observational data about.

1.3. Objectives and Scope of This Paper

The objective of this paper is to explore how a multi-representational approach can help to extract relevant information from a data set, with a view to improving prediction and understanding, and with the ultimate goal of providing support for the discovery process. Our specific goal is to develop a better understanding of the nature of catchment hydrology across the hydro-geo-climatologically diverse extent of Chile. Rather than the traditional strategy of implementing a single, preselected, computational model code to the entire country, or perhaps different variations of a model code to hydrologically different parts of the country, we instead implement three

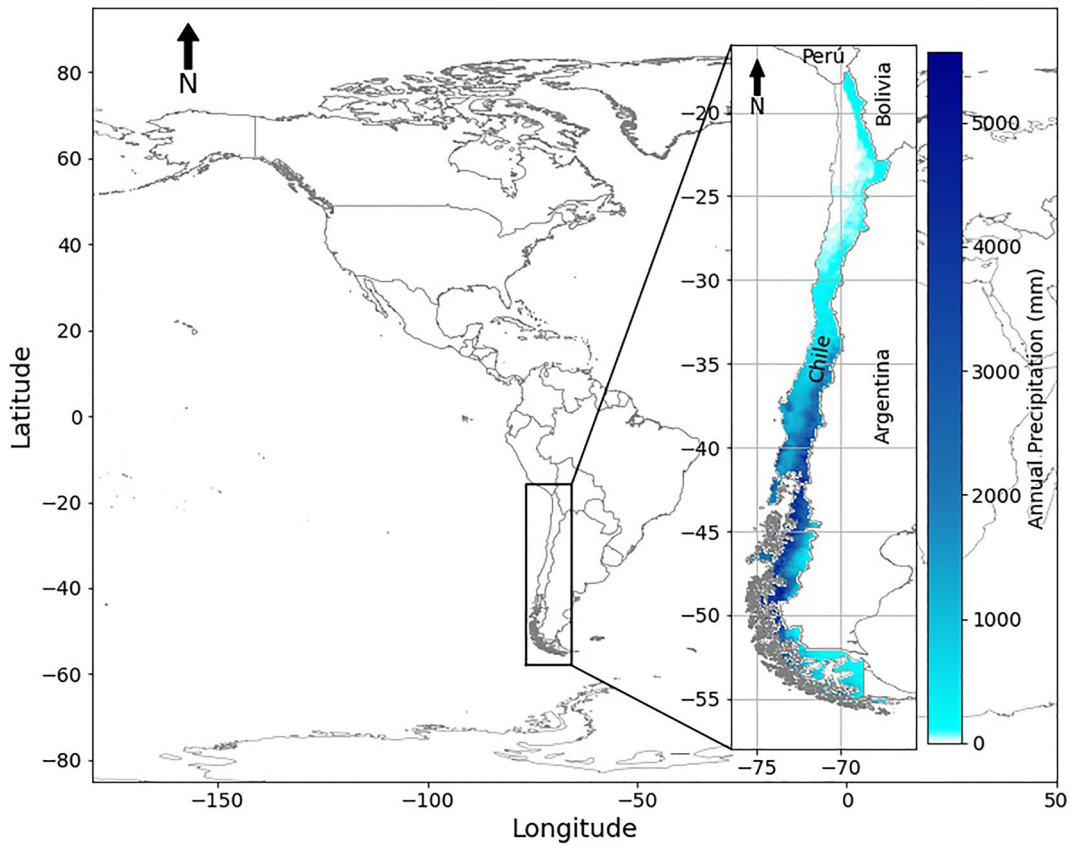

Figure 1. Map showing the geographic location of Chile, and its spatial distribution of mean annual precipitation. In the northern region, the precipitation is almost 0 mm/year and in the southern, it could be higher than 5,000 mm/year.

different representational approaches including one PC-based approach and two ML-based approaches. Our focus is on understanding the strengths and weaknesses associated with each of these three representational approaches, and on exploring the potential richness of inferences that such a multi-representational approach can support.

In the next section, we introduce the problem of catchment-scale hydrological prediction in the context of the climatology of Chile. Section 3 will discuss the study methodology. The study results are presented in Section 4. Finally, we provide a discussion and some thoughts about the implications of this work in Section 5.

To be clear, this study should be considered exploratory, with a view to improving our understanding of how a multi-representational approach can be exploited in the service of enhanced scientific discovery. Further, we make no claim to making any significant fundamental discoveries—what we are suggesting, and seeking to illustrate, is that adoption of a multi-representational approach is advisable in order to maximize the possibility of discovery (as in "*chance favors the prepared mind;*" a phrase is attributed to nineteenth-century bacteriologist Louis Pasteur).

2. The Challenge of Streamflow Prediction Across Hydrologically Diverse Chile

Prediction of streamflow at large scales is challenging, due to the multitude of relevant factors that can vary simultaneously across time and space. In particular, the ability of hydrological models to generalize can be poor in regions where the spatial variability of dynamical forcings and static attributes is large (Malone et al., 2015). This is especially relevant to Chile, which is characterized by tremendous geo-hydro-climatic variability, both along its 4,270 km (2,653 mi) North-South extent and also from East to West (Figure 1). At one extreme, Northern Chile is home to the driest desert in the world, containing regions where no precipitation has been recorded

for more than 25 years. At the other extreme, more than 5,000 mm/year of precipitation has been recorded in parts of the south, where there are also permanent icefields.

Bordered by the Pacific Ocean to the West and Argentina to the East, the country averages just 175 km (109 mi) in width, while the North-South running Andes Mountain range rises to the highest elevation in South America (6,959 m or 22,831 ft). Moreover, a second mountain range, with lower elevations, runs parallel to the coast along almost the entire country. Owing to the high elevations of the mountain ranges, precipitation in the headwater catchments occurs mainly as snow, due to which the corresponding streamflow peak will appear many days or even weeks after the precipitation event. In contrast, where liquid precipitation occurs in catchments with high slopes, the times of concentration can be shorter than 1 day. Other factors, including the variability of forest fraction, degree of human intervention, and valleys created between the two mountain ranges, are also strongly related to the availability of water in the long term.

This immense geographic and hydrologic variability in climatic conditions poses a considerable challenge for any modeling enterprise, and especially for the PC representational approach where model structures must be selected in advance. As such, Chile presents a perfect opportunity to explore the possibility of developing modeling techniques that can deal with large climatic variability, and even exploit it to achieve better model performance.

3. Study Methodology

This section presents and discusses our study methodology, including the data set used (Section 3.1), three representational methods used (Section 3.2), and issues related to the experimental design (Section 3.3).

3.1. Data Set

For this study, we use the information provided by the catchment-scale CAMELS-CL data set (Alvarez-Garreton et al., 2018). This data set includes 11 variables and 105 categorical and numerical attributes for 516 Chilean catchments with daily data. Those catchments are distributed from latitudes 17.8°S to 54.9°S and represent almost the entire spatial extent of Chile (17.5°S–55.5°S). Catchment areas range from 18 to 52,244 km^2 with an average of 2,407 km^2 (median of 780 km^2), and together cover 53% of the entire surface of the country. Variables in the data set include streamflow, precipitation, temperature, potential evapotranspiration (PET), snow water equivalent (SWE), and static catchments attributes. For model development and evaluation, we selected 322 catchments selected to span the country and to have a minimum streamflow record length of 7 years. The literature suggests that 2–3 water years of daily data represents a minimum record length for calibration of conceptual process-resolved models (Gupta & Sorooshian, 1985) while around 8–10 years may be required to ensure some degree of stability with respect to the estimated model (Vrugt et al., 2006). On balance, therefore, 7 years represents a reasonable tradeoff between the availability of the model development and spatial representation of catchments. Note also that the time periods of model development data selected for each catchment are not necessarily identical or even overlapping, they simply represent whatever is available for those catchments. More details on how the data were selected and partitioned appear in De la Fuente (2021).

3.2. Representations Examined

We selected three different representational strategies for investigation, with a view to exploring whether doing so would lead to better understanding of the system. These included one lumped PC-based water balance modeling approach and two ML-based modeling strategies.

The PC-based approach represents the mainstay of how understanding is developed in science. Models based on this representational approach are designed to be structurally and behaviorally isomorphic to the system, and therefore enable theoretical prior knowledge (such as conservation and thermodynamic principles) to be imposed as physical constraints on the allowable input-state-output trajectories of a system (Gharari et al., 2014, 2021). Their strength lies in the ability to constrain model behaviors to be consistent with physical principles, so that meaning can be ascribed to the various components, fluxes, and state variables of the model. This makes it possible, in principle, to transfer understanding between locations and to generalize to classes of systems that share similar representational properties, although the development of reliable functions to achieve such transfers remains an unsolved problem in hydrology (Blöschl et al., 2019). However, this strength can become a weakness

when, by imposing overly strong prior restrictions on model structures, we limit their ability to learn explicitly and directly from data, and to discover things that are inconsistent with the space of hypotheses explicitly covered by those priors (Gharari et al., 2021).

Conversely, ML approaches have recently gained a reputation for being able to help address some of the most challenging tasks in science, particularly where theoretical understanding is lacking or is weak (e.g., Kratzert et al., 2018; Hu et al., 2018; Sudriani et al., 2019; Zhang et al., 2018, among many others). The power of ML arises from its practical ability to extract complex relationships from large data sets, and from its theoretical ability to approximate any input-output (or, where appropriate, input-latent-variable-output) mapping to an arbitrarily high degree of accuracy. Notably, different ML approaches are based in different mathematical perspectives about how to represent the structures underlying a given data set, and/or on how to represent and extract information contained in the data (see below). Accordingly, when different ML approaches are applied to any given data set, each is likely to provide a different "informational" perspective on the underlying nature of the DGP. By understanding how different ML algorithms represent and extract information from data, we can seek to understand the value offered by each perspective and exploit it to obtain a more comprehensive picture of the underlying system.

For the PC-based representation, we chose the GR4J dynamical lumped water balance model (Perrin et al., 2003), due to its relative parsimony and the reports of good performance in other studies (Kunnath-Poovakka & Eldho, 2019; Pagano et al., 2010; Sezen & Partal, 2019), and because the catchment-scale data required for its implementation is available (Table A1). Moreover, we coupled to it the lumped CemaNeige snowmelt module (Valéry et al., 2014) to account for snow process dynamics at high altitudes in Andes Mountain range. For the ML-based representations, we selected the Long Short Term Memory network (LSTM; Hochreiter & Schmidhuber, 1997) and the Random Forest regression-tree method (RF; Breiman, 2001).

While additional representational strategies could also have been included, the three strategies explored here arguably represent sufficiently different approaches to extracting information from data to support the objectives of this study. Further details about these representational approaches are provided in Supporting Information S1; we encourage readers without a background in these approaches to read this material.

3.3. Experimental Design

The main challenge to creating a unified model development methodology is that each of the three representational strategies has different conceptual, mathematical, and coding characteristics, and therefore different structures and processes of implementation, that must be followed to obtain an operational model. It is, therefore, impossible to implement an entirely uniform methodology for model development. Accordingly, we followed the reasonable approach of implementing the commonly followed model development practices for each representational type and comparing the results so obtained. Accordingly, all comparisons are based on the use of the same daily data and performance metrics for model development and evaluation.

3.3.1. Partitioning the Data

A key step in model development is to partition the available data into model development and evaluation subsets, where the former is used for model structure selection and parameter tuning, while the latter is used to assess the generalization performance that can be expected from the developed model. However, no clear guidance exists on how to achieve such a partitioning for data that represent dynamical hydrological systems (Daggupati et al., 2015; Guo et al., 2020; Wu et al., 2013; Zheng et al., 2018). In general, the hydrological literature has traditionally assumed that the entire available data set comes from a stationary underlying DGP, and that any split that preserves the full range of hydrologic variability (dry, medium, and wet) in both sets is satisfactory. Based on this assumption, it is common to use a continuous-time period that makes up ~60%–80% of the available data for model development, while allocating the remaining ~20%–40% for an evaluation of the generalization ability of the model.

In this study, we adopt the strategy of further partitioning the model development subset into "*calibration*" and "*selection*" subsets, where the calibration subset is used for model/network parameter tuning (commonly called "*training*" in the ML literature), and the selection subset is used for model/network structure selection and/or hyperparameter tuning (commonly called "*validation*" in the ML literature). Note that we adopt this naming convention to try and overcome the existing inconsistency in terminology between the ML and hydrological

modeling literature. Accordingly, the available data are partitioned into three subsets, where the first 60% of the data is used for model calibration, the next 24% is used for model/hyperparameter selection, and the final 16% is used for model evaluation (commonly called "*testing*" in the ML literature).

For consistency, all comparisons across the three representations are done using the exact same data subsets. Further, to ensure robust inference, the statistical distributions of all performance evaluation metrics computed on these data subsets were estimated via bootstrapping (Efron & Tibshirani, 1994) and the medians of these distributions were used as the representative values in all comparisons.

3.3.2. Variable Selection

The variables selected from the CAMEL-CL (Alvarez-Garreton et al., 2018) data set include two sources of precipitation (CR2MET and MSWEP, both having long records), three values characterizing temperature (maximum, mean, and minimum), and an estimate of PET obtained via the Hargreaves and Samani (1985) method. Further, the available SWE data does not cover the entire country and was therefore not considered suitable for the current study.

Because the GR4J model has a pre-defined input representation, we used the weighted average of the two sources of precipitation (CR2MET and MSWEP; see Section 3.3.5), temperatures, and PET as input to the GR4J model. This weighted average can be obtained in two ways; (a) by constraining the weights to sum to one (without bias correction), or (b) by allowing the sum of the weights to differ from one (with bias correction). We studied these configurations because it allows different corrections to be applied to the overall precipitation volume, which could be specific for each catchment given the diversity of Chile.

In summary, all three representational approaches are provided with access to the same meteorological forcing information. More details regarding the variables and attributes used for the development of each model type are presented in Tables A1–A3.

3.3.3. Representing the Overall Hydrological Memory

For the RF-based model, which does not explicitly include dynamical state variables, information regarding hydrological memory was included by concatenating past inputs (precipitation, evapotranspiration, and temperature) to the inputs for the current time step, and the number of past input lags was treated as a model hyperparameter. This emulates the idea of a Markov Process, where a state variable can be thought of as a summary property of an infinite number of past inputs to the system. While this strategy enables important information to be made available to the model, it results in a very high cost (in terms of computational and storage resources) because considerable computational memory is required to manage the data set as the number of lags is increased. We found that at 32 days of lagged memory, the computational system became unstable so that any analysis of longer time-lags could not be supported using the available ram memory of the computer (16 GB). This prevented us from explicitly exploring longer memory time scales, such as 270 or 365 days (or longer), and the results presented in the next sections only consider a memory time scale of 16 days (the most stable solution so obtained). To partially address this issue, we augmented the input data to include the month-of-year as a surrogate variable intended to be informative about the longer-term state of the system. The idea was to enable the model to learn a representation of long-term memory as the average behavior associated with different months of the year.

3.3.4. Model Warm-Up

It is recommended, regardless of representational strategy, to use a warm-up period (during which performance metrics are not computed) to minimize errors associated with the initialization of dynamical model states. For lumped water balance modeling it is common to use a full year (365 days) of data for this purpose; for example, Perrin et al. (2003) used a full year to initialize the GR4J model, following the suggestion of Chiew and McMahon (1994). For the LSTM ML approach, Kratzert et al. (2019) used 270 days, after testing 90, 180, 270, and 365 days as different options.

In this study, we adopted the following strategy for warm-up period selection. For the GR4J and LSTM-based models, we followed the strategy of first tuning the LSTM to determine a suitable warm-up period length (as a model hyperparameter) and then using that same period to warm-up the GR4J model; this is possible due to the similarity in definitions of the "warm-up" process in PC-based models and the "sequence information" process in

LSTM-based models. Note that the RF-based model does not need a warm-up period given its lack of Markovian state variables.

3.3.5. Parameters and Hyperparameters to be Tuned

Each model involves different sets of parameters and hyperparameters, depending on its structural form. GR4J contains 4 tunable parameters plus 2 for the CemaNeige snow module, which must be calibrated for each catchment. However, to ensure the best performance of GR4J, three combinations of parameters were evaluated, and the parameter set with the best performance was used when comparing against the ML-based models. The three combinations are (a) GR4J with snow module but *without* implementation of a precipitation bias correction factor (7 parameters), (b) GR4J without snow module but *with* implementation of a precipitation bias correction factor (6 parameters), and (c) GR4J *with* snow module and implementation of precipitation bias correction (8 parameters). Moreover, by employing a Box-Cox transformation (Box & Cox, 1964) on streamflow, all of the configurations include an extra parameter that enables the optimization procedure to correct for skewness and thereby use the best form of the cost function for each catchment. This parameter enables the training approach to determine whether or not a transformation enables better characterization of the time series. A summary of the parameters is presented in Table A4.

For the LSTM-based model, in addition to the large number of system-wide network weights and biases, five hyperparameters must be tuned, namely the sequence length (memory from the past hidden states), number of hidden nodes, batch size, number of epochs, and the Box-Cox transformation parameter.

Finally, in addition to determining the nodal split parameters, the RF-based model requires that 4 hyperparameters be tuned for the entire set of catchments taken together. The first represents hydrological memory (expressed as the number of days previous to the current day for which inputs are simultaneously presented to the model), the second is the Box-Cox transformation parameter, the third is the number of trees, and the fourth is the minimum number of elements that must be retained in the last leaf.

3.3.6. Out-of-Sample Testing

For an additional out-of-sample model evaluation step, we retained all of the CAMELS-CL catchments for which less than 7 years (for training) but more than 1 year of data is available. So, while none of these catchments were included in the model development data set, we can exploit the fact that they have similar climatic variability for model assessment (Figure A1). The resulting 167 catchments facilitate a meaningful out-of-sample operational comparison of the generalization abilities of the LSTM-based and RF-based models. Note that the GR4J model was not tested using this out-of-sample set of catchments since regional generalization of lumped water balance model parameters to ungauged catchments is not within the scope of this paper.

4. Experimental Results

This section consists of two main parts. In Section 4.1 we investigate issues of overall understanding, such as hydrological memory and feature importance, supported by different representational characteristics of each model. In Section 4.2 we investigate how the models performed in terms of the ability to generalize in space and time, looking for causes of similarities and differences in performance.

4.1. Understanding Enabled by the Multi-Representational Approach

Each representational approach responds differently to the fluxes of information through the system, and that response can provide useful insights into the characteristics of the system. Here, we investigate how hydrological memory and the relative importance of attributes provide insights into the underlying nature of the DGP.

4.1.1. Hydrological Memory

For the ML-based representations, an important property is how hydrological memory is characterized, in terms of the number of previous time-steps of input data (meteorological variables) that are determined to provide useful information about the current value of streamflow. Note that this lag-time hyperparameter is not relevant to the GR4J model, which tracks hydrological memory exclusively through its state variables. Figures 2a and 2b

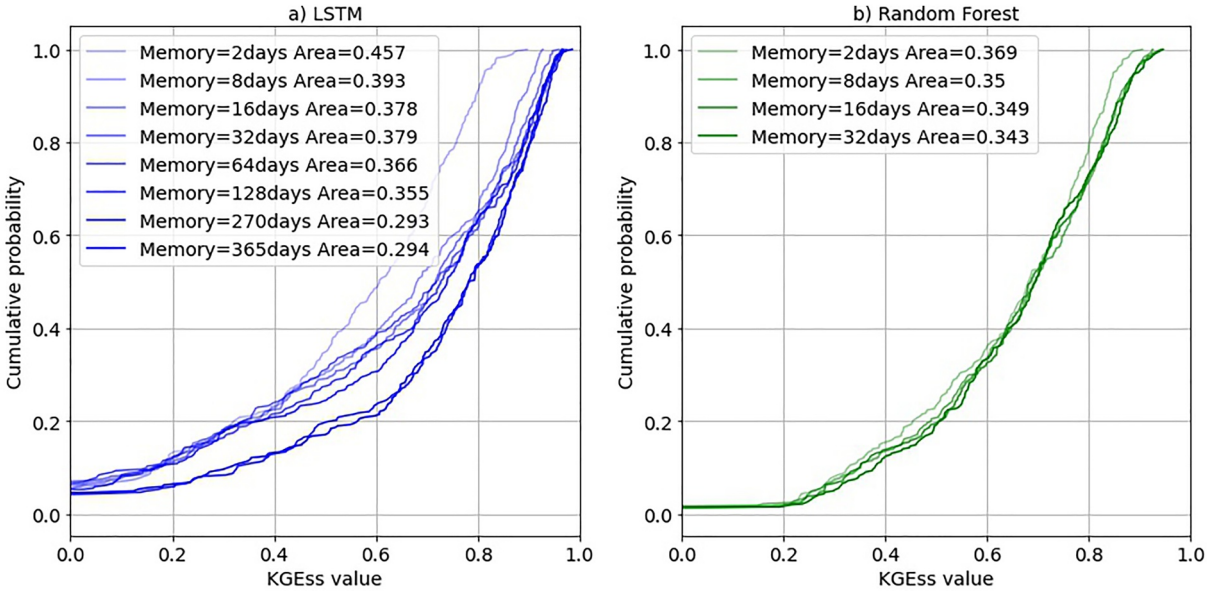

Figure 2. Cumulative density function for the KGE skill score performance for (a) long short-term memory model and (b) RF model (selection period). Lines closer to the right represent better overall performance. The area under each curve is presented as a guideline to check which lag performs better.

show, for LSTM and RF respectively, how the cumulative density functions (CDF) of model performance vary with values of the lag-time hyperparameter.

Consider first the LSTM-based model. For catchments with KGE skill score (KGEss) >∼0.45, the CDFs move progressively to the right (indicating improved performance) as the lag-time is increased from 2 to 32 days, whereas for catchments with KGEss ∼0.45 the results are insensitive to the value of the lag-time hyperparameter (similar area under the curve for 16 and 32 days). Further, while the rate of improvement in performance increase slowly (on average) as the lag-time is increased, we see distinct improvement when going from 128 to 270 days, occurring mainly in catchments with KGEss <∼0.85. A similar, but less strong, result is found for the RF-based model between 2 and 16 days for catchments with KGEss > 0.70 (little change in the area under the curve for different lags).

These results suggest that the ML-based models are detecting the expression of two different kinds of processes giving rise to streamflow generation across the country, one related to (shorter) hydrological memory of around 32 days and the other related to (longer) memory of around 270 days. We will revisit this topic in the next section, where we see that this difference in length of hydrological memory is correlated with climatic attributes. Note that this kind of information about systemic differences between catchments in the study region is somewhat more difficult (and time demanding) to infer from the warm-up period that we could use with the PC-based GR4J representation, because this memory is specific to the starting day in the training period, and is not general as is the memory learned using the ML-based models.

4.1.2. Feature Importance

Another interesting aspect of ML-based approaches is the manner and ease by which the relative informativeness/importance of climatic attributes can be assessed. Whereas this is, in principle, also possible using a PC-based modeling approach, such an inference would have to be done indirectly through an analysis of the spatial patterns of calibrated values of the model parameters, which is arguably a less direct and somewhat more complicated process that is out of the scope of the present research.

Here, we analyze the information provided by feature importance (calculated using Gini importance, Breiman et al., 1984), and by the average values of the thresholds adopted at the first split, which are inherent properties of the RF-based model. If these data-space thresholds occur earlier in construction of the regression tree, they can be interpreted as being more fundamentally or globally important. Figure 3 indicates that the most important attribute is an aridity index (aridity_cr2met, computed using the annual CR2MET precipitation product divided

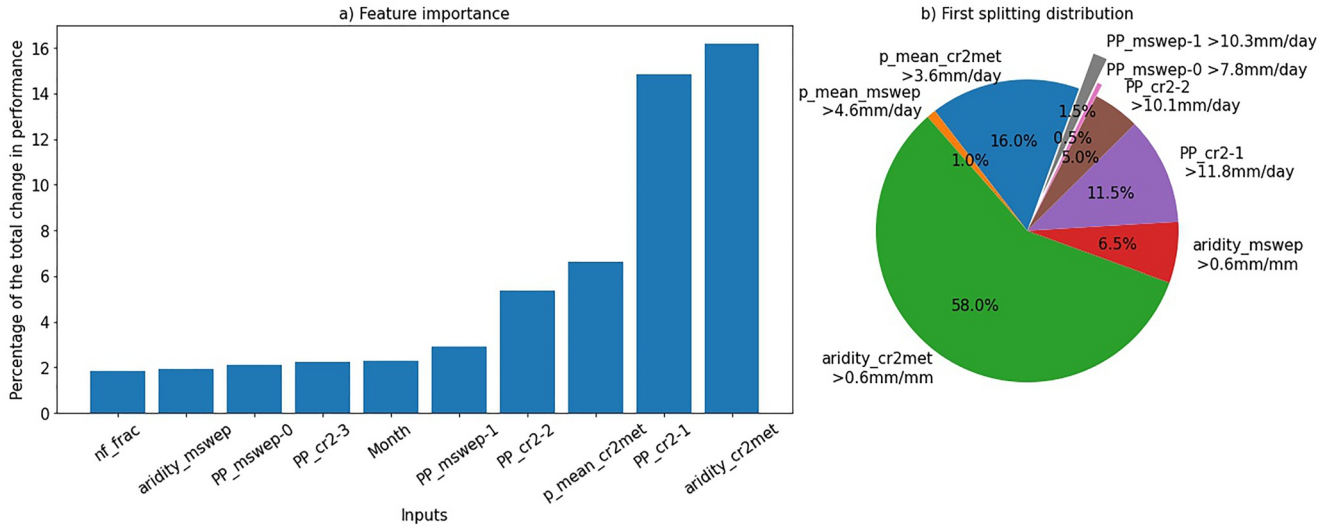

Figure 3. Feature importance and distribution of the first split of the RF model (names in Table A2). (a) It presents the top 10 inputs that generates more performance deterioration when this input is shuffled. (b) Percentage of times and the average threshold that this input was selected as first splitting in the 200 decision trees.

by the annual PET), which strongly suggests that the form of the relationship between the availability of water and the generation of streamflow is different in different parts of the country (e.g., in humid vs. arid regions). While this observation is not novel (Booij et al., 2019; Chen et al., 2019; Meira Neto et al., 2020), it is consistent with the need for flexibility in the architecture and process-parametrization representation of PC models when applied to diverse climatic conditions.

Of course, this does not imply that failure to account for aridity is, per se, a complete and meaningful explanation for the poor performance of any given model type. In general, spatio-temporal changes in aridity index are likely to simply indicate relative changes in the importance of various drivers of streamflow. From Figure 3a we see that the second, fourth, fifth, and seventh most important attributes are daily precipitation values, which indicates that the behavior of the RF model is mainly controlled by aspects related to precipitation, once aridity has been accounted for.

The first split that occurs in most trees of the RF-based model (Figure 3b) occurs at an aridity index threshold of 0.6 mm/mm. This observation suggests that different streamflow generating representations may be required for the model to perform well in regions that are energy-limited (aridity index < 1) as opposed to water-limited (aridity index > 1). When a similar analysis is performed for the daily precipitation threshold (independent of the lag), we find that the nature of the streamflow response is different for values above/below $\sim$10 mm/day, but of course, much more analysis would need to be done to understand the reasons for those specific values.

Finally, we note that of the top 10 most important attributes, the only ones that are not related to aridity and/or precipitation are the month of the year (Month) and forested fraction (nf_frac). The month of year attribute conveys information related to climatic cycling (annual periodicity), whereas the forest fraction could convey (among other things) information about infiltrability and soil water retention capacity of the soil.

The main point of these two (rather simple) examples shown in Figure 3 is that the regression tree representation underlying the RF-based model facilitates a kind of analysis that can provide interesting information that is not easily or directly obtained using either the PC-based or LSTM-based representational approaches. Moreover, while these RF-based results might seem obvious (because they concur with our prior hydrological understanding of catchment behaviors), such checks are important when employing an ML-based representation if we want to proceed further with the process of deepening understanding (such as by investigating the pruning of unnecessary features, and analyzing deeper levels of the decision trees, etc.). In this sense, the RF-based model is a strongly informative tool when our goal is to use modeling in support of scientific understanding and discovery.

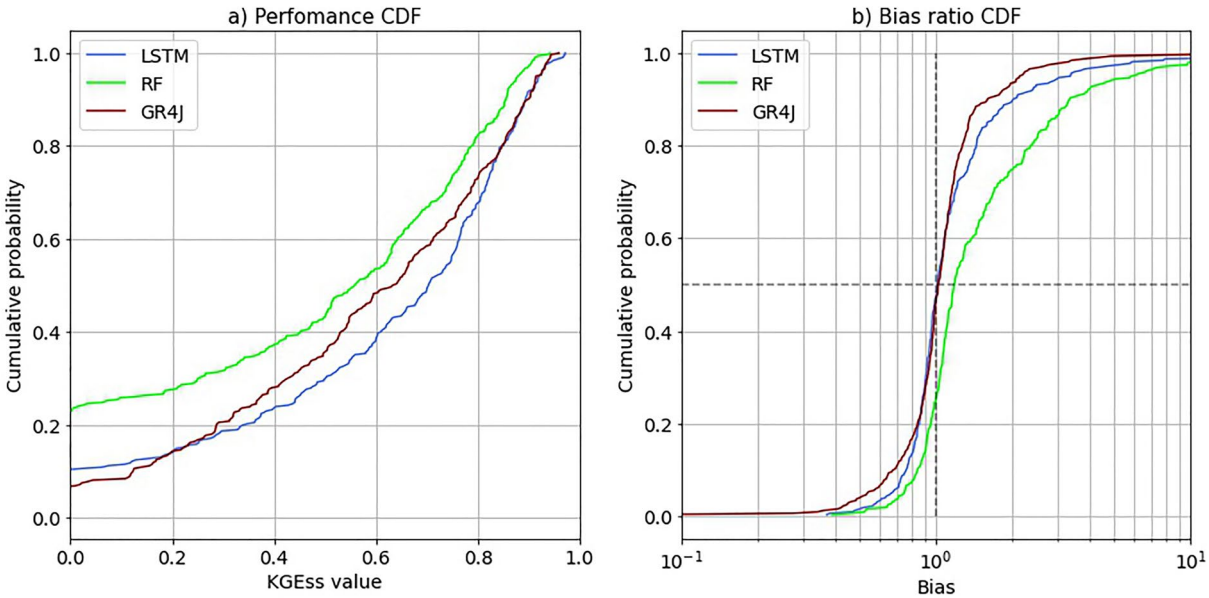

Figure 4. Cumulative density function for the KGE skill score performance and bias ratio for the three models (evaluation period). (a) Lines closer to the right represent better overall performance. (b) Lines closer to the vertical line at 1 (10°) represent better overall performance.

4.2. Analysis of Similarities and Differences in Performance

The relative ability of any properly trained model to perform well (evaluation period) can be considered indicative of how well the underlying DGP has been represented. However, even if all the models tested provide essentially identical values for some aggregate performance metric (such as KGEss or NSE), a deeper analysis may reveal systematic differences in model simulated behaviors, simply because the aggregate metric is not capable of distinguishing between them (Gupta et al., 2008, 2009).

For that reason, in this section, we examine not only the similarities and differences as assessed by an aggregate metric (KGEss), but also more detailed analysis in terms of components of the metric (e.g., bias), the spatial distribution of the metric, and the ability of the model to generalize to unobserved catchments. Through this analysis, we also assess whether each representational approach supports (or not) our prior understanding of the climatic diversity of Chile.

4.2.1. Overall Performance

First, we examine the distributions of overall model performance across the country. Figure 4a shows the CDFs of evaluation period performance (as measured by KGEss) for all locations where KGEss $>$ 0 (where predictions are, on average, better than the no-model prediction that simply uses the observed mean; Knoben et al., 2019). Similar results were obtained using NSE (De la Fuente, 2021). Two interesting points can be noted:

1. The LSTM curve (blue line) is significantly further to the right (~85% of the catchments) over most of the range, indicating statistically better overall performance.
2. The GR4J curve (red line) fails to meet the KGEss $>$ 0 threshold at only ~7% of the catchments, as opposed to ~11% for LSTM and ~22% for RF.

Regarding the first result, the superior performance of the LSTM-based model over most of the range is (arguably) expected given that the LSTM can both (a) explicitly learn about system dynamics and memory through its representation of state variable recurrence, and (b) learn the functional form of the input-state-output mapping due to its structural flexibility. Note that the former ability is not explicitly enabled by the RF architecture (green line), while the latter ability is not possible for the fixed GR4J architecture (red line).

Regarding the second point, given that all three models are trained using the same input-output information, this result suggests that there are climatic conditions under which the GR4J-based model provides useful (lumped water balance) *information that is not directly inferable from the available data* by the LSTM and RF representations

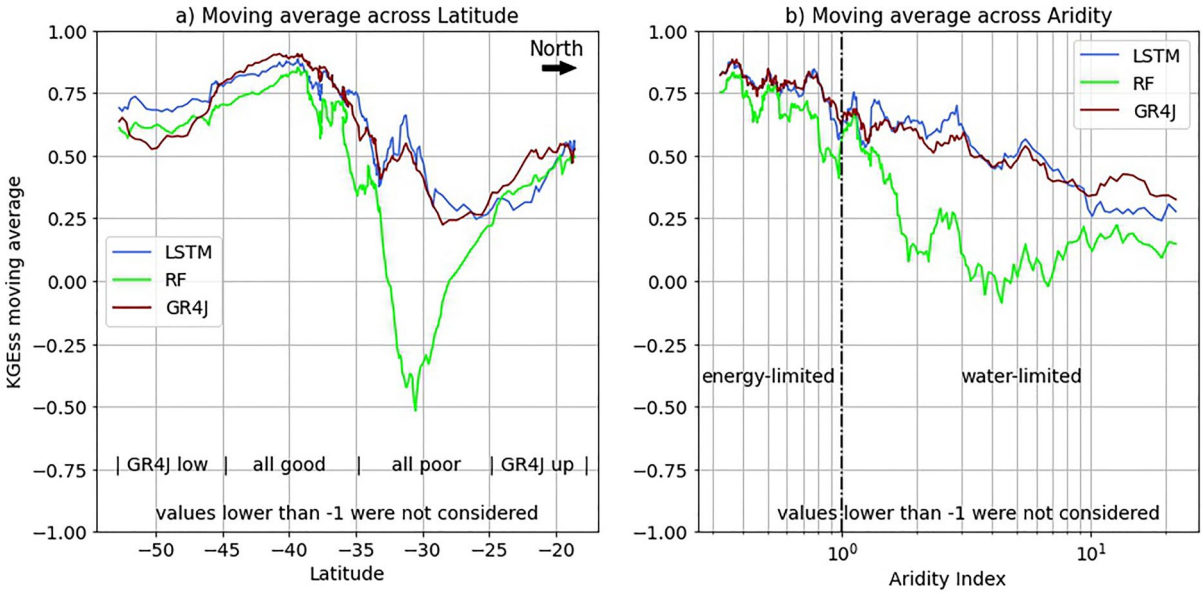

Figure 5. Variation of KGE skill score (KGEss) performance versus aridity and latitude (evaluation period). Figure shows the rolling moving average of 15 catchments sorted by different attributes. The best performance is gotten for KGEss = 1.0 and everything lower than 0.0 means that the prediction is worse than a constant prediction equal to the mean of each catchment.

(~15% of the catchments); this is arguably something that was not known before. Of course, whether this benefit comes from the specific mass-conserving and process-equation nature of the GR4J architecture, or from its ability to compensate for mass-balance errors by importing/exporting groundwater (or some other reason) is not immediately clear, and will require more detailed investigation. Despite the recent study by Hoedt et al. (2021), where a mass-conservative LSTM-based model was found to be able to learn a good latent-variable representation of the dynamics of snow storage, such findings would need to be tested at larger scales over a variety of climatic conditions before more general conclusions can be drawn.

Next, we examine the distributions of the decomposition components of KGEss. While aggregate metrics such as KGEss can provide a good overall idea of model performance, they can often be poor at revealing important differences in characteristic model behaviors, particularly when overall performance is poor (Gupta et al., 2009). Figure 4 provides further discriminatory information by plotting the CDF of model Bias Ratio, where values larger (smaller) than 10^0 (= 1) indicate a tendency to overestimate (underestimate).

This plot reveals that the GR4J and LSTM-based models, which have the explicit ability to simulate system dynamics, tend (on average) to be unbiased, whereas the RF-based model tends to be positively biased. Interestingly, for situations where the models tend to overestimate the mean (Bias Ratio > 1.0), the GR4J model tends to do better (have a lower bias) than the LSTM-based model. However, for situations where the LSTM-based model tends to underestimate the mean (Bias Ratio < 1.0) that situation is reversed. In the case of the RF-based model, its systematic bias does not allow us to consider it in this comparison.

So, while the LSTM-based model is statistically superior in terms of overall KGEss performance for most catchments, the situation is clearly more nuanced—with each representational type providing different characteristic abilities to simulate various attributes of streamflow, even though all the models were trained using the same data. This supports our contention that a multi-representational approach can aid in providing better predictions when developing an ensemble-based model, particularly when faced with significant climatic variability.

4.2.2. Spatial Patterns of Performance

In this section, we investigate how the different representational types perform across the variety of different climatic conditions that characterize Chile. Figures 5a and 5b explore the relationship between model performance and two interesting climatic factors—Latitude, and Aridity. Given the long narrow shape and North-South orientation of the country, these two factors serve as useful surrogates for climatological variability, with the

Northern extent of the country being characterized by very dry conditions and high elevations, the Southern extent being characterized by extreme precipitation and permanent icefields, and the central region is characterized by intermediate degrees of wetness and considerable variability in elevation.

The curves in Figure 5a show smoothed trajectories (using a moving average of 15 catchments) of the variation in KGEss performance with Latitude from South to North (left to right across the *x*-axis). First, we see that, while all three models exhibit relatively good performance in the mid- and south-central (moderately wet) parts of the country [latitude − 45° to − 35°], the performance of GR4J decreases sharply relative to the ML-based RF and LSTM as we move to the southernmost regions [latitude − 55° to − 45°]. This decline in GR4J performance makes sense given that the south is characterized by the existence of glaciers and lakes, which can introduce significant time-lags into the dynamics of the system that cannot easily be reproduced by the GR4J architecture (even when supplemented by the CemaNeige snow module), resulting in flashy and delayed simulated hydrographs compared to observations. In contrast, the flexibility of the ML-based representations enables higher degrees of information extraction from the data.

Meanwhile, all three models exhibit relatively poor performance (KGEss < 0.5) across the north-central parts of the country [latitude − 35° to − 25°]. This region is characterized by steep slopes (very short times of concentration), relatively greater aridity (with only a few precipitation events per year), and snowmelt and groundwater being the relevant processes, which contrasts with the mid/south-central and southern regions. Here, the RF-based model performs particularly poorly, which is largely attributable to the fact that it does not have explicit access to data with greater than 16 days lag time and is, therefore, (unlike in GR4J and LSTM) unable to account for longer time-scales of hydrological memory.

Finally, the northern part of the country [latitude − 25° to − 18°] contains the Atacama Desert, which is the most arid region in the world and has moderate slopes. Here, GR4J exhibits better performance than RF and LSTM. Meanwhile, the relatively poor performance achieved by all models between latitude −35° and −18° suggests that the variables that make up the existing CAMELS-CL data set are not sufficiently informative about the particular input-state-output dynamics of the catchments in this region to enable a robust (*stable under changes*) and accurate model to be developed, and that other variables and attributes should be added to improve model performance.

The curves in Figure 5 show smoothed trajectories for how KGEss performance varies with Aridity Index (computed as the mean of aridity_cr2met and aridity_mswep attribute). Here we see a clear dependence of performance on aridity, with all three models exhibiting better performance (KGEss > 0.5) under wet (i.e., energy limited) conditions but with performance becoming progressively worse as the climatic conditions become increasingly more arid (water-limited). Interestingly, the performance of both the GR4J and the LSTM-based models (that can simulate system dynamics) declines more or less linearly with increasing log-aridity, but RF performance declines somewhat more rapidly and is significantly worse than for GR4J and LSTM when the Aridity Index is between about 1.5 and 8.0. Given that GR4J is designed to represent systems that are primarily driven by precipitation, it is understandable that performance can decline as the direct dependence of streamflow on precipitation becomes less, while the mediating effects of evapotranspiration and long-term groundwater storage become more predominant.

However, while the ML-based models have considerably more flexibility to discover appropriate functional relationships in the data and would therefore normally be expected to serve as indicators of upper bounds on achievable model performance (Nearing et al., 2020), they also show the same declining trend in performance with increasing aridity. This suggests that the information content of the CAMELS-CL data set is biased toward a better representation of the hydrological properties of wet (energy-limited) catchments and is therefore not sufficiently complete to enable model development for arid parts of the country. For example, it is noteworthy that the CAMELS-CL data set does not include information about soil characteristics such as depth to bedrock, hydraulic conductivity, or soil texture, all of which are present in the US version (Addor et al., 2017), and which can be very important in the characterization of the baseflow and streamflow-precipitation elasticity (Addor et al., 2018).

Another interesting observation is that the hydrological memory associated with streamflow generation appears to be different for energy-limited (wet) and water-limited (arid) catchments. Referring to Figure 2a and combining it with the result in Figure 5, we see that catchments with KGEss > 0.45 showed improvement in LSTM performance when provided with ∼32 past days of input data. This suggests the dominance of short-time-scale memory

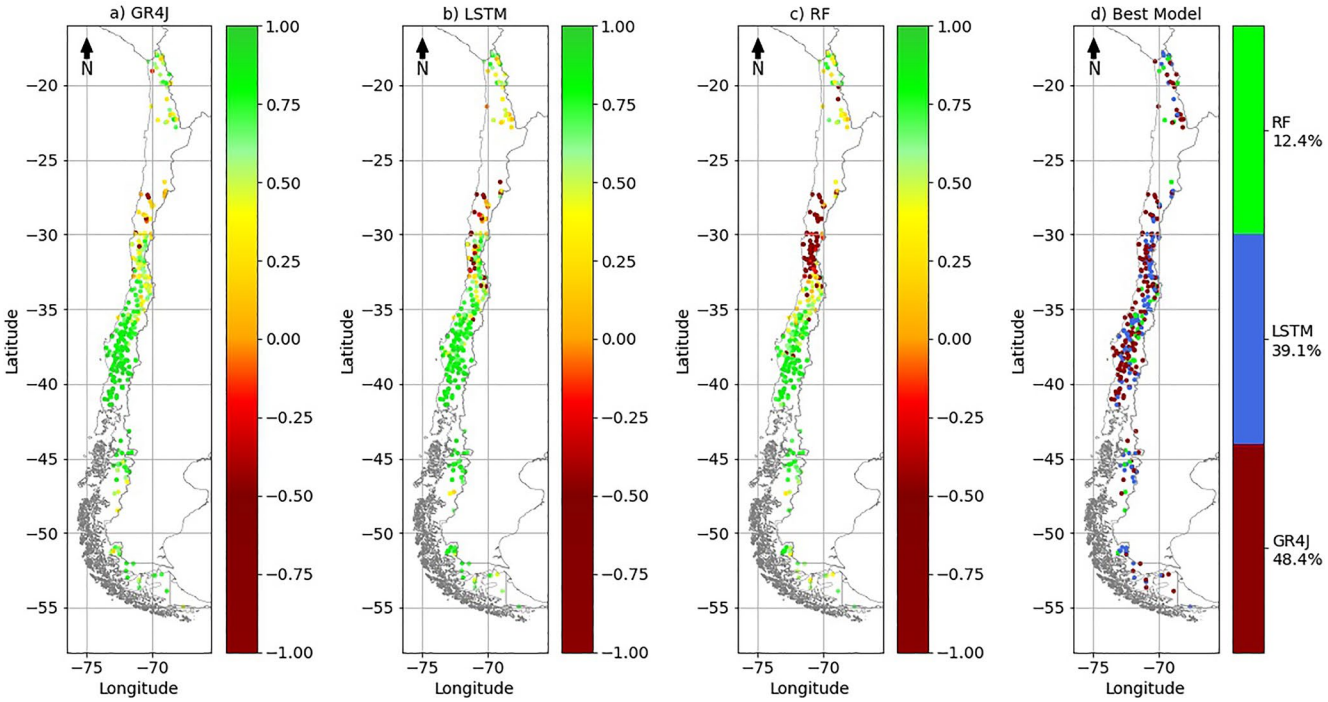

Figure 6. Spatial distributions of KGE skill score performance for each model (evaluation period). (a) GR4J, (b) long short-term memory, (c) RF. Green colors present a good performance and red values very bad performance. Panel (d) presents the best model for each catchment.

processes in energy-limited catchments (Figure 5b). On the other hand, there is a smaller set of catchments with poorer model performance that showed improvement only when provided with 270+ past days of input data, reflecting longer-time-scale memory processes associated with water-limited regions. This indicates that when investigating and modeling the streamflow response of catchments, our representation—whether ML-based or PC-based—must contain structures that make it possible to track memory processes at more than one dominant time-scale, depending on the climatology of the region.

The important point is that the representational type selected for model development should make it possible for information about multiple climatic time scales to be exploited. The GR4J and LSTM-based models contain explicit representations (through dynamic state variables and multiple flow pathways) that to some degree facilitate this, however the assessment of the data-based RF model only included data lagged up to 16 days, which may explain why performance is worse than for the data-based LSTM when the aridity index is in the range of 1.5–8.0.

Figure 6 show evaluation-period KGEss performance for each of the three models at each catchment used for model development (green indicates good performance, yellow-orange indicates poor performance, and red indicates really poor performance). Focusing specifically on the region between latitudes 27°S and 33°S, we see that RF (Figure 6c) performs very poorly throughout this part of the country (see also Figure 5). However, LSTM performs quite well along a narrow strip of this region that borders Argentina. This strip is located at higher elevations where temperatures are low and where snowmelt processes dominate the generation of streamflow. The ability of LSTM to track longer-term memory processes is likely contributing to its good performance here. As we move westward toward the coast, LSTM performance decreases, indicating that the model no longer has access to the information needed to properly simulate the streamflow response (which, in this case, is probably information about connections between groundwater and streamflow). Turning to GR4J, we see that its KGEss performance across the region is just slightly better than 0.0, indicating that the model is mainly reproducing the long-term mean value of streamflow and some of the variability (we checked this visually for some of the catchments). Given that GR4J does not have the explicit ability to represent the complex long-term dynamics of regional groundwater systems, these results make sense.

Figure 6d indicates which model provides the best evaluation-period KGEss performance per catchment (green = RF, blue = LSTM, red = GR4J). Here we simply report the model with the best evaluation-period

Table 1
Summary Statistic in the Evaluation Period (322 Catchments)

Model	Min	25% Percentile	50% Percentile	75% percentile	Max	# Positive	# Best
GR4J	**−5.215**	0.433	0.676	**0.840**	**0.978**	**313**	**156**
RF	−703.128	0.075	0.563	0.762	0.940	249	40
LSTM	−1,170.490	**0.442**	**0.704**	0.826	0.971	289	126

Note. The first five columns describe the KGEss distribution for each model. The sixth column describes the number of cases with positive performance. The last column describes the number of cases where this model had the best performance. Bold highlights the higher value from the three models.

KGEss. No clear pattern emerges, but in general, the blue (LSTM) and red (GR4J) colors dominate, with GR4J generally being the best-performing model across the country. This could be an indication that some extra information is available to GR4J (probably through the regularizing effect of the water balance constraint) which enables it to slightly outperform the other models on average (Figure 5).

Some more nuanced findings emerge from a statistical analysis of KGEss performance by model type, reported in Table 1. While the LSTM has an excellent median KGEss performance of 0.70, it obtains the best KGEss value at only 126 of the 322 catchments (39.1%). Further, where the LSTM fails, it does so badly. In contrast, GR4J performs best at 48.4% of the catchments with a median KGEss performance of 0.68, its distribution has much lower skewness and dispersion, and it achieves positive KGEss values (i.e., KGEss > 0) at a greater number (97%) of the catchments.

4.2.3. Spatial Generalization

The results presented so far indicate that the LSTM-based model has the potential to provide the "*best*" overall (median) performance, while GR4J tends to provide more "*robust*" results in cases where data-based approaches may fail. Meanwhile, the RF-based model is particularly useful for enabling exploration, by providing clues that can lead to hypotheses about what kinds of climatic processes (and hence data sets) should be incorporated into ongoing model development efforts.

However, the previous analysis was for a pseudo-independent data set, consisting of evaluation-period data from the same catchments that were used for model development. As such, the results may not provide a reliable assessment of the quality of model performance that might be expected at (other/new) catchments that are not part of the model development data set. Figures 7a–7c and Table 2 report the results of our out-of-sample analysis. Since GR4J parameter estimates are not available for these catchments (an extra parameter regionalization step would be required, that was not pursued in this study), this assessment was done only for LSTM-based and RF-based models.

Overall, the out-of-sample results indicate that LSTM and RF do not show significantly different (relative to each other) spatial distributions of performance. This tends to conflict with the in-sample evaluation results (Figure 8), even though both the in-sample and out-of-sample catchment locations are distributed similarly with respect to the Aridity Index. When we compare the CDF's of in-sample and out-of-sample performance (Figure 8) for these models, we see that both RF and LSTM exhibit remarkably similar statistical distributions of out-of-sample performance, which suggests that both ML-based approaches have a similar ability to generalize to locations that were not included in the model development data set. There is, however, a larger deterioration in the statistical distribution of model performance from in-sample to out-of-sample for LSTM than for RF.

Finally, Table 2 reports a more detailed statistical analysis of KGEss performance by model type, showing that RF slightly outperforms LSTM on most of the statistical indicators. So, while LSTM achieved in general better temporal generalization (in sample), the results for out-of-sample generalization are less definitive. The tradeoff between temporal and spatial generalization ability may be somehow different for each representational type. While the LSTM-based model is trained using batch data from different catchments, the RF model focuses on finding the best split for all catchments simultaneously, which may make it less sensitive to the new conditions encountered in out-of-sample testing. While this is simply speculative at this point, it would be interesting to further examine this issue using large sample catchment-scale data sets from other parts of the world.

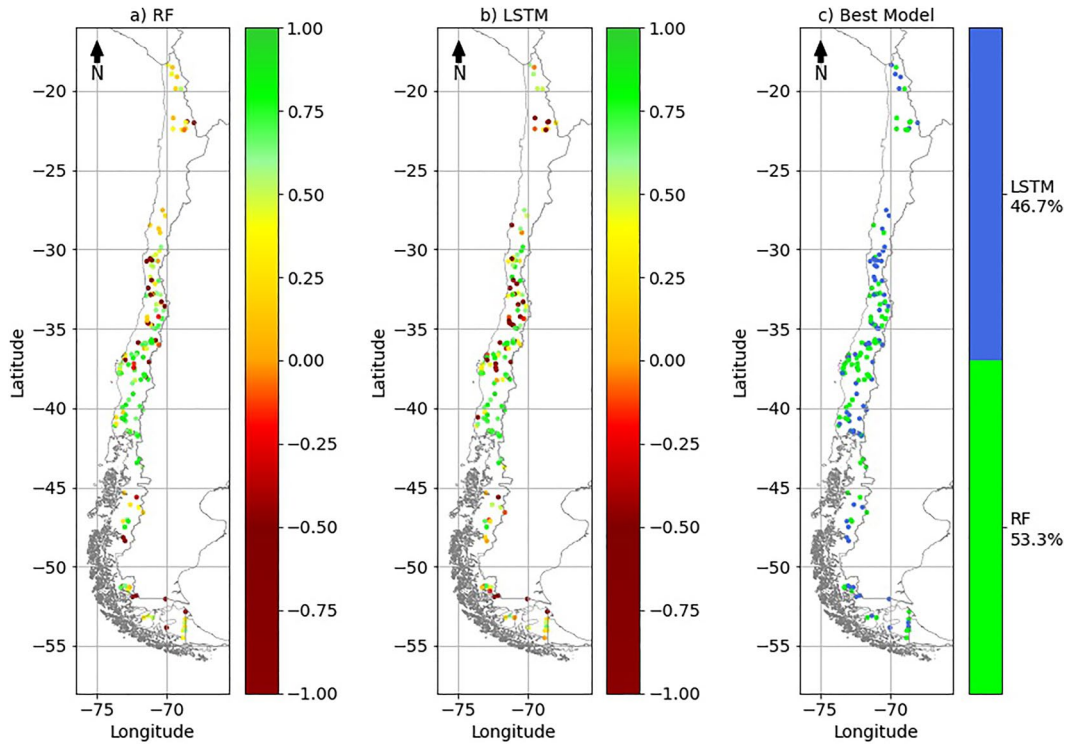

Figure 7. Spatial distributions of KGE skill score performance for each model (out-of-sample). (a) RF and (b) long short-term memory. Green colors present a good performance and red values very bad performance. Panel (c) presents the best model for each catchment.

5. Discussion and Conclusions

An understanding of how hydrological processes vary across large scales is important to the development of strategies for mitigating the effects of floods and droughts (and other natural hazards). Such understanding can be difficult to establish, given the large number of variables, attributes, and relationships that need to be considered. Under such circumstances, the traditional approach of attempting to model the entire diversity of hydro-geo-climatic conditions across a country/region with a single representational approach may not result in a sufficiently accurate characterization of the underlying DGP. Through a case study, we have explored the possibility of using a multi-representational approach to address the challenge of climatic diversity, where the different representations are selected to have different mathematical structures and assumptions, with the goal of maximizing prediction and understanding in support of discovery.

5.1. Opportunities and Challenges of a Multi-Representational Approach

While each representation can support different kinds of investigation, understanding and discovery through the model development and evaluation process, the adoption of a multi-representational approach creates

Table 2
Summary Statistic in Out-of-Sample (167 Catchments)

Model	Min	25% Percentile	50% Percentile	75% Percentile	Max	# Positive	# Better
RF	**−17.501**	**0.118**	**0.45**	0.666	0.897	**130**	**89**
LSTM	−203.124	−0.103	0.429	**0.678**	**0.968**	116	78

Note. The first five columns describe the KGEss distribution for each model. The sixth column describes the number of cases with positive performance. The last column describes the number of cases where this model had the best performance. Bold highlights the higher value from the three models.

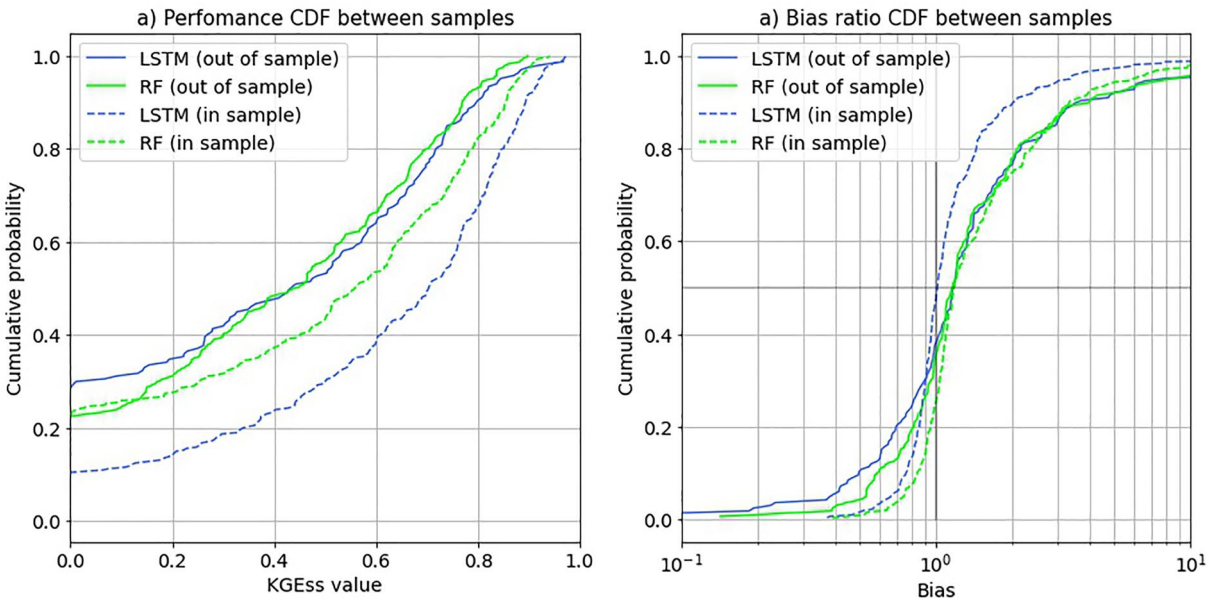

Figure 8. Cumulative density function for the KGE skill score performance and bias ratio for long short-term memory and RF model (evaluation period vs. out-of-sample). (a) Lines closer to the right represent better overall performance. (b) Lines closer to the vertical line at 1 (10°) represent better overall performance.

opportunities that can be exploited and challenges that need to be addressed. As such, the multi-representation approach adopted in this study was valuable in improving our understanding of the DGP in the following ways.

- It facilitated a better understanding of the time-scales of hydrological memory in the system, and specifically that they are associated with degree of aridity. Our analysis suggests that (for Chile) the ability to access input information over the past 32 days (short time scales) is critical to achieving an optimal representation of energy-limited catchments, whereas the memory time-scale required for arid catchments is much longer (~270 days). From our survey of the literature, this understanding is novel.
- It suggests that multi-representational ensembles provide a practical approach to improving streamflow predictions. The model evaluation results indicate that the LSTM-based model provides better overall performance, while the PC-based GR4J model tends to be more robust (providing better performance where KGEss < 0.20). Moreover, the two models exhibit contrasting bias ratio distributions which is desirable to ensemble construction. Meanwhile, the out-of-sample assessment indicated that the RF-based model remains a viable approach as well.
- It clearly revealed the need for more informative data associated with arid regions. Combining the fact of log-linear decline in performance with aridity for both GR4J and the LSTM-based model, with the fact that LSTMs are more capable of maximizing information extraction from data (relative to PC-based approaches), it seems fair to conclude that further performance improvements in arid regions would require catchment-scale data sets that are more informationally complete.
- It facilitated a rapid and simple exploration of features important for prediction, via the regression-tree-based RF approach (complemented by GR4J and LSTM). We find clearly that aridity provides the most important basis for explaining differences between catchment behaviors. Beyond this, various characteristic features and their identified thresholds provide strong explanatory power (e.g., aridity index equal to 0.6 mm/mm, and previous days precipitation equal to 10 mm/day).

Meanwhile, we encountered challenges to be tackled in future research, including.

- Because different representational approaches may exploit the information in data in different ways, and can impose different requirements for inference, it becomes difficult to implement a completely uniform strategy for multi-representational model development.

- For PC-based representations where learning about the spatial variability of hydrological processes is necessarily mediated through an analysis of spatial patterns or parameters, multiple parameter sets can give rise to similar model performance, which can complicate the ability to make meaningful inferences.

Overall, our results illustrate how the process of synthesizing results obtained via a multi-representational approach can lead to a more comprehensive overall picture of the underlying DGP, thereby creating a broader context within which deeper exploration can facilitate further discovery.

5.2. On the Issue of Data Informativeness

In an ensemble approach to prediction, it might make sense to implement a PC-based (e.g., lumped water balance) model as a lower benchmark given its embodiment of valuable regularizing information that may help to prevent model performance from becoming catastrophically poor under conditions where the data is insufficiently informative about the dynamics of streamflow generation. As such, we should require that any ML-based approach under consideration should demonstrate benefits over the benchmark (see e.g., Schaefli and Gupta, 2007). In the case than an ML-based model fails in a comparison against the benchmark, we should consider the possibility that it has been unable to exploit all the information provided by the data set. Further, when every representational approach fails to perform well, this should alert us to the possibility that the data may not be sufficiently informative in quantity and/or quality regarding the processes we seek to model.

Our analysis suggests that this situation is true for the CAMELS-CL data set. Notably, the data set does not include attributes from which it could be possible to infer groundwater-driven baseflow or other related processes, thereby limiting proper characterization of the streamflow response across Chile. However, this situation will require further investigation and exploration of alternative sources of relevant information.

5.3. Conclusions

We propose that a meaningful answer to the question "*Does a single 'correct' catchment-scale hydrological model exist at all?*" expressed by Clark et al. (2011), is that we should instead abandon any concept of a "best" model, and instead consider the value of learning to live with a plurality of representations while developing strategies for extracting important relevant information from the representational ensemble.

Another way of think of this is that it is the ensemble of representations that is actually the "***Model***" per se, since it incorporates a representation of "*what we know that we do not know*" (i.e., our known uncertainties). From this perspective, our task is to populate this ensemble with representations that best support our investigative goals. Returning to the distinction made in the introduction, given that any representational approach can be used to express numerous alternative hypotheses, the aforementioned task is clearly consistent with the idea of a "*multiple hypothesis approach*" (Clark et al., 2011), but one where the hypotheses are selected to be as informationally diverse as possible (containing contrasting structures and assumptions) so that the possibilities of learning and discovery can be maximized. In contrast, an approach where the ensemble consists of hypotheses that may representationally be only marginally different from each other (e.g., that all share the same or similar system architectures while differing only in the forms of the process parameterization equations) may not lend itself to efficient and effective learning (Gharari et al., 2021).

Such a perspective unavoidably affects how we think about the model development process and its role in a scientific investigation. Our view is that conceptual/process/theory-based and ML-data-based approaches to model development must co-exist within such an environment, with neither being the dominant approach, and that a multi-representational strategy is key to promoting model-based scientific discovery. Based on our experience with this exploratory study, we firmly believe that the multi-representational approach will be fundamental to achieving a better understanding of hydrology at the large scale. There, the complexity of the system we seek to understand (and represent) demands access to large and informationally diverse data sets and an analytical strategy that is purposefully diverse. As always, we are keenly interested in dialog and collaboration on this and related issues of how we use models to support prediction and understanding, in support of scientific discovery.

Appendix A: Definition of Metrics

A1. Model Calibration/Training

To calibrate the parameters of the GR4J model to each catchment in the calibration period, we tested both the Root Mean Square Error (RMSE) and the Kling-Gupta Efficiency (KGE, Gupta et al., 2009), as defined below. Overall, we found that KGE provided slightly more robust results (De la Fuente, 2021), and therefore we present here only the results obtained using KGE.

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2}$$

y_i : Measured streamflow

$\hat{y}_i$: Simulated streamflow

n : Total number of data

r : Linear correlation coefficient between y_i and $\hat{y}_i$

α : σ_s/σ_o : relative variability between simulated and observed data.

β : μ_s/μ_o : ratio between simulated and observed data.

For parameter optimization, we used three algorithms from the Spotpy Python library (Houska et al., 2015), namely Maximum Likelihood Estimation, Differential Evolution Adaptive Metropolis (DE-MCz), and Shuffled Complex Evolution. In total, 22 independent optimization runs were done for each catchment, and the parameter set that provided the best performance (of the 22 parameter sets so obtained) on the selection (hyperparameter tuning) data subset was chosen for the specific GR4J configuration used.

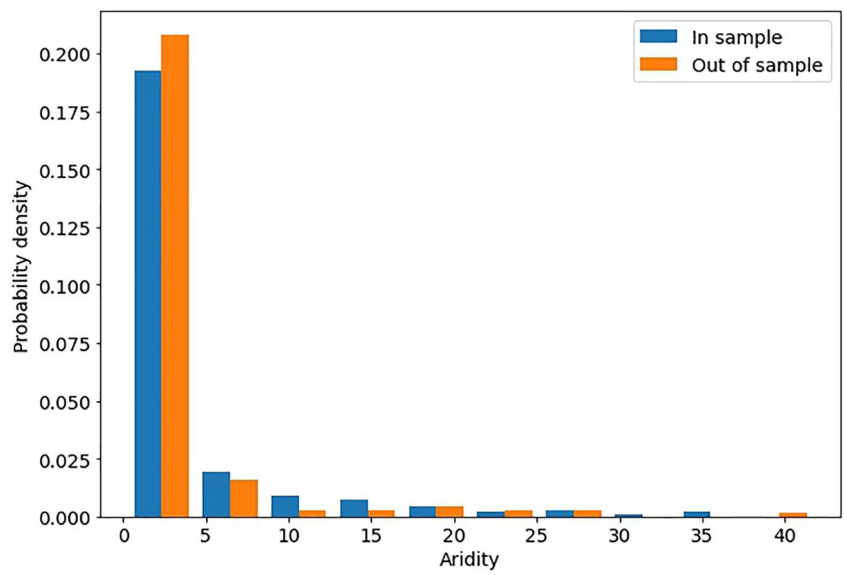

Figure A1. Aridity histogram comparison between both samples used in the performance analysis. In terms of aridity, both samples look very similar.

Table A1
Variables Used in the GR4J Model

Variable	Description
PP_cr2-0	Precipitation in the same day ("0") of the mean streamflow from CR2MET product (mm)
PP_mswep-0	Precipitation in the same day ("0") of the mean streamflow from MSWEP product (mm)
T_{max-0}	Maximum temperature in the same day ("0") of the mean streamflow (°C)
T_{mean-0}	Mean temperature in the same day ("0") of the mean streamflow (°C)
T_{min-0}	Minimum temperature in the same day ("0") of the mean streamflow (°C)
ETP-0	Potential Evapotranspiration in the same day ("0") of the mean streamflow (mm)
Q	Daily mean streamflow (mm)

Note. Static variables were not used given its catchment-by-catchment training.

To develop the RF model, we use the Scikit-learn Python library (Pedregosa et al., 2011). The *RandomForestRegressor* module (version 0.23.1) has two options for performance metrics—Mean Squared Error (MSE) and Mean Absolute Error (MAE). While MAE can be used to reduce the tendency to emphasize larger streamflow values, because we are implementing the Box-Cox transformation on streamflow we chose MSE to be the metric used for RF calibration.

Table A2
Variables Used in the Random Forest Model

No	Variable	Unit	No	Variable	Unit	No	Variable	Unit	No	Variable	Unit	No	Variable	Unit
1	PP_cr2-0	mm	31	PP_mswep-13	mm	61	T_{mean-9}	°C	91	ETP-5	mm	121	gauge_lon	°
2	PP_cr2-1	mm	32	PP_mswep-14	mm	62	$T_{mean-10}$	°C	92	ETP-6	mm	122	grass_frac	%
3	PP_cr2-2	mm	33	PP_mswep-15	mm	63	$T_{mean-11}$	°C	93	ETP-7	mm	123	gw_rights_flow	#
4	PP_cr2-3	mm	34	PP_mswep-16	mm	64	$T_{mean-12}$	°C	94	ETP-8	mm	124	gw_rights_n	#
5	PP_cr2-4	mm	35	T_{max-0}	°C	65	$T_{mean-13}$	°C	95	ETP-9	mm	125	high_prec_dur_cr2met	days
6	PP_cr2-5	mm	36	T_{max-1}	°C	66	$T_{mean-14}$	°C	96	ETP-10	mm	126	high_prec_dur_mswep	days
7	PP_cr2-6	mm	37	T_{max-2}	°C	67	$T_{mean-15}$	°C	97	ETP-11	mm	127	high_prec_freq_cr2met	days/y
8	PP_cr2-7	mm	38	T_{max-3}	°C	68	$T_{mean-16}$	°C	98	ETP-12	mm	128	high_prec_freq_mswep	days/y
9	PP_cr2-8	mm	39	T_{max-4}	°C	69	T_{min-0}	°C	99	ETP-13	mm	129	imp_frac	%
10	PP_cr2-9	mm	40	T_{max-5}	°C	70	T_{min-1}	°C	100	ETP-14	mm	130	lc_barren	%
11	PP_cr2-10	mm	41	T_{max-6}	°C	71	T_{min-2}	°C	101	ETP-15	mm	131	lc_glacier	%
12	PP_cr2-11	mm	42	T_{max-7}	°C	72	T_{min-3}	°C	102	ETP-16	mm	132	low_prec_dur_cr2met	days
13	PP_cr2-12	mm	43	T_{max-8}	°C	73	T_{min-4}	°C	103	Q	mm	133	low_prec_dur_mswep	days
14	PP_cr2-13	mm	44	T_{max-9}	°C	74	T_{min-5}	°C	104	area	km	134	low_prec_freq_cr2met	days/y
15	PP_cr2-14	mm	45	T_{max-10}	°C	75	T_{min-6}	°C	105	aridity_cr2met	-	135	low_prec_freq_mswep	days/y
16	PP_cr2-15	mm	46	T_{max-11}	°C	76	T_{min-7}	°C	106	aridity_mswep	-	136	Month	#
17	PP_cr2-16	mm	47	T_{max-12}	°C	77	T_{min-8}	°C	107	big_dam	#	137	nf_frac	%
18	PP_mswep-0	mm	48	T_{max-13}	°C	78	T_{min-9}	°C	108	carb_rocks_frac	%	138	p_mean_cr2met	mm
19	PP_mswep-1	mm	49	T_{max-14}	°C	79	T_{min-10}	°C	109	crop_frac	%	139	p_mean_mswep	mm
20	PP_mswep-2	mm	50	T_{max-15}	°C	80	T_{min-11}	°C	110	day	#	140	p_mean_spread	mm
21	PP_mswep-3	mm	51	T_{max-16}	°C	81	T_{min-12}	°C	111	elev_gauge	m	141	p_seasonality_cr2met	-
22	PP_mswep-4	mm	52	T_{mean-0}	°C	82	T_{min-13}	°C	112	elev_max	m	142	p_seasonality_mswep	-
23	PP_mswep-5	mm	53	T_{mean-1}	°C	83	T_{min-14}	°C	113	elev_mean	m	143	pet_mean	mm
24	PP_mswep-6	mm	54	T_{mean-2}	°C	84	T_{min-15}	°C	114	elev_med	m	144	shrub_frac	-
25	PP_mswep-7	mm	55	T_{mean-3}	°C	85	T_{min-16}	°C	115	elev_min	m	145	slope_mean	m/km

Table A2
Continued

No	Variable	Unit	No	Variable	Unit	No	Variable	Unit	No	Variable	Unit	No	Variable	Unit
26	PP_mswep-8	mm	56	T_{mean-4}	°C	86	ETP-0	mm	116	forest_frac	%	146	snow_frac	-
27	PP_mswep-9	mm	57	T_{mean-5}	°C	87	ETP-1	mm	117	fp_frac	%	147	sur_rights_flow	#
28	PP_mswep-10	mm	58	T_{mean-6}	°C	88	ETP-2	mm	118	frac_snow_cr2met	%	148	sur_rights_n	#
29	PP_mswep-11	mm	59	T_{mean-7}	°C	89	ETP-3	mm	119	frac_snow_mswep	%	149	wet_frac	-
30	PP_mswep-12	mm	60	T_{mean-8}	°C	90	ETP-4	mm	120	gauge_lat	°			

Note. Dynamic variables are lagged 16 days and all non-categorical attributes from CAMELS-CL data set were used.

Table A3
Variables Used in the LSTM Model

No	Attribute or variable	Unit	No	Attribute or variable	Unit
1	PP_cr2-0	mm	26	high_prec_dur_cr2met	days
2	PP_mswep-0	mm	27	high_prec_dur_mswep	days
3	T_{max-0}	°C	28	high_prec_freq_cr2met	days/y
4	T_{mean-0}	°C	29	high_prec_freq_mswep	days/y
5	T_{min-0}	°C	30	imp_frac	%
6	ETP-0	mm	31	lc_barren	%
7	Q	mm	32	lc_glacier	%
8	Area	km	33	low_prec_dur_cr2met	days
9	aridity_cr2met	–	34	low_prec_dur_mswep	days
10	aridity_mswep	–	35	low_prec_freq_cr2met	days/y
11	big_dam	#	36	low_prec_freq_mswep	days/y
12	carb_rocks_frac	%	37	nf_frac	%
13	crop_frac	%	38	p_mean_cr2met	mm
14	elev_gauge	m	39	p_mean_mswep	mm
15	elev_max	m	40	p_mean_spread	mm
16	elev_mean	m	41	p_seasonality_cr2met	-
17	elev_med	m	42	p_seasonality_mswep	-
18	elev_min	m	43	pet_mean	mm
19	forest_frac	%	44	shrub_frac	-
20	fp_frac	%	45	slope_mean	m/km
21	frac_snow_cr2met	%	46	snow_frac	-
22	frac_snow_mswep	%	47	sur_rights_flow	#
23	grass_frac	%	48	sur_rights_n	#
24	gw_rights_flow	#	49	wet_frac	-
25	gw_rights_n	#			

Note. Dynamic variables are lagged until 270 days but they are not presented in this table. All non-categorical attributes from CAMELS-CL data set were used.

Table A4
Parameters and Search Range Used in the GR4J Optimization

Parameter	Description	Searching range
Alpha1	Amplification factor for CR2MET precipitation product	0–2.5
Alpha2	Amplification factor for MSWEP precipitation product	0–2.5
Theta1	Snowmelt factor	0.1–7
Theta2	Cold content factor	0–1
x1	Storage production capacity	0–5,000
x2	Amplification of water exports	−10 to 10
x3	Storage routing capacity	1–1,500
x4	Time-delay between the initial and maximum values of the hydrograph	0.501–4.5
Lambda	Exponent of Box-Cox transformation	0–2.0

Note. The description of each parameter is presented by its physical meaning. Moreover, the last column shows the range used in the optimization.

$$\text{MSE} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$$

$$\text{MAE} = \frac{\sum_{i=1}^{n} \text{abs}(y_i - \hat{y}_i)}{n}$$

To train the LSTM model, we used the implementation provided by Kratzert et al. (2019) and modified it to conform to the data structures and variables of the CAMELS-CL data set. Whereas the original code enables the choice of either MSE or NSE as the calibration metric, we used only NSE because its normalization of the error enables better comparison across catchments having different amounts of temporal variability.

$$\text{NSE} = 1 - \frac{\text{MSE}}{\sigma_o^2}$$

A2. Model Performance Evaluation

For performance evaluation, we use the KGE skill score (KGEss) (Knoben et al., 2019) computed on this period. KGEss is a rescaled version of the KGE metric such that a value of zero corresponds to the prediction being no better than simply using the mean observed streamflow, in a manner analogous to NSE. While other metrics, including NSE and RMSE, were also used for model evaluation (De la Fuente, 2021), we do not report them here as the conclusions are like those obtained using KGEss. Importantly, we account for sampling variability by computing the estimated posterior distributions of KGEss by bootstrapping 100 times (Efron & Tibshirani, 1994) and using the median value of KGEss in all comparisons.

$$\text{KGE}ss = \frac{\text{KGE} - \text{KGE}_{\text{benchmark}}}{1 - \text{KGE}_{\text{benchmark}}} = \frac{\text{KGE} + \sqrt{2} - 1}{\sqrt{2}} = 1 - \frac{1 - \text{KGE}}{\sqrt{2}}$$

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Data Availability Statement

The CAMELS-CL data set is freely available from https://doi.pangaea.de/10.1594/PANGAEA.894885. The analytical methods and codes are freely available at http://www.hydroshare.org/resource/fc08997100fa4cd6abdd8a4f5731de15.

Acknowledgments
This publication is the product of research done by De la Fuente (2021) to satisfy the requirements for obtaining a Master of Science degree in Hydrology while being funded by the Agencia Nacional de Investigacion y Desarrollo (Chile) through "*Beca de Magíster en el Extranjero, Becas Chile en Áreas Prioritarias, Convocatoria 2018*." Gupta acknowledges partial support from the Australian Research Council (ARC) through the Centre of Excellence for Climate Extremes Grant CE170100023. Condon acknowledges partial support from NSF Early Career Award Grant 1945195.

References

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., & Clark, M. P. (2018). A ranking of hydrological signatures based on their predictability in space. *Water Resources Research*, *54*(11), 8792–8812. https://doi.org/10.1029/2018WR022606

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., et al. (2018). The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies-Chile dataset. *Hydrology and Earth System Sciences*, *22*(11), 5817–5846. https://doi.org/10.5194/hess-22-5817-2018

Blöschl, G., Bierkens, M. F., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., et al. (2019). Twenty-three unsolved problems in hydrology (UPH)–a community perspective. *Hydrological Sciences Journal*, *64*(10), 1141–1158. https://doi.org/10.1080/02626667.2019.1620507

Booij, M. J., Schipper, T. C., & Marhaento, H. (2019). Attributing changes in streamflow to land use and climate change for 472 catchments in Australia and the United States. *Water*, *11*(5), 1059. https://doi.org/10.3390/w11051059

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, *26*(2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Taylor & Francis.

Chen, S. A., Michaelides, K., Grieve, S. W., & Singer, M. B. (2019). Aridity is expressed in river topography globally. *Nature*, *573*(7775), 573–577. https://doi.org/10.1038/s41586-019-1558-8

Chiew, F., & McMahon, T. (1994). Application of the daily rainfall-runoff model MODHYDROLOG to 28 Australian catchments. *Journal of Hydrology*, *153*(1–4), 383–416. https://doi.org/10.1016/0022-1694(94)90200-3

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resources Research*, *47*(9). https://doi.org/10.1029/2010WR009827

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015c). The structure for unifying multiple modeling alternatives (SUMMA), Version 1.0: Technical description. *NCAR Tech*. Note NCAR/TN-5141STR. https://doi.org/10.5065/D6WQ01TD

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015a). A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research*, *51*(4), 2498–2514. https://doi.org/10.1002/2015WR017198

Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015b). A unified approach for process-based hydrologic modeling: 2. Model implementation and case studies. *Water Resources Research*, *51*(4), 2515–2542. https://doi.org/10.1002/2015WR017200

Daggupati, P., Pai, N., Ale, S., Douglas-Mankin, K. R., Zeckoski, R. W., Jeong, J., et al. (2015). A recommended calibration and validation strategy for hydrologic and water quality models. *Transactions of the ASABE*, *58*(6), 1705–1719. https://doi.org/10.13031/trans.58.10712

De la Fuente, L. (2021). *Using big-data to develop catchment-scale hydrological models for Chile (Master dissertation)*. The University of Arizona. Retrieved from https://repository.arizona.edu/handle/10150/656824

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Gharari, S., Gupta, H. V., Clark, M. P., Hrachowitz, M., Fenicia, F., Matgen, P., & Savenije, H. H. (2021). Understanding the information content in the hierarchy of model development decisions: Learning from data. *Water Resources Research*, *57*(6), e2020WR027948. https://doi.org/10.1029/2020WR027948

Gharari, S., Hrachowitz, M., Fenicia, F., Gao, H., & Savenije, H. H. G. (2014). Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration. *Hydrology and Earth System Sciences*, *18*(12), 4839–4859. https://doi.org/10.5194/hess-18-4839-2014

Guo, D., Zheng, F., Gupta, H., & Maier, H. R. (2020). On the robustness of conceptual rainfall-runoff models to calibration and evaluation data set splits selection: A large sample investigation. *Water Resources Research*, *56*(3), e2019WR026752. https://doi.org/10.1029/2019WR026752

Gupta, H. V., Clark, M. P., Vrugt, J. A., Abramowitz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, *48*(8), W08301. https://doi.org/10.1029/2011WR011044

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gupta, H. V., Wagener, T., & Liu, Y. (2008). Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrological Processes: International Journal*, *22*(18), 3802–3813. https://doi.org/10.1002/hyp.6989

Gupta, V. K., & Sorooshian, S. (1985). The relationship between data and the precision of parameter estimates of hydrologic models. *Journal of Hydrology*, *81*(1–2), 57–77. https://doi.org/10.1016/0022-1694(85)90167-2

Hargreaves, G. H., & Samani, Z. A. (1985). Reference crop evapotranspiration from temperature. *Applied Engineering in Agriculture*, *1*(2), 96–99. https://doi.org/10.13031/2013.26773

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hoedt, P. J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G., & Klambauer, G. (2021). MC-LSTM: Mass-Conserving LSTM. arXiv preprint arXiv:2101.05186 https://arxiv.org/abs/2101.05186v3

Houska, T., Kraft, P., Chamorro-Chavez, A., & Breuer, L. (2015). SPOTting model parameters using a ready-made python package. *PLoS One*, *10*(12), e0145180. https://doi.org/10.1371/journal.pone.0145180

Hu, C., Wu, Q., Li, H., Jian, S., Li, N., & Lou, Z. (2018). Deep learning with a long short-term memory networks approach for rainfall-runoff simulation. *Water*, *10*(11), 1543. https://doi.org/10.3390/w10111543

Knoben, W. J., Freer, J. E., & Woods, R. A. (2019). Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005–6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Benchmarking a catchment-aware long short-term memory network (LSTM) for large-scale hydrological modeling (p. 08456). arXiv preprint arXiv:1907. https://doi.org/10.5194/hess-2019-368

Kunnath-Poovakka, A., & Eldho, T. I. (2019). A comparative study of conceptual rainfall-runoff models GR4J, AWBM and Sacramento at catchments in the upper Godavari river basin, India. *Journal of Earth System Science*, *128*(2), 33. https://doi.org/10.1007/s12040-018-1055-8

Malone, R. W., Yagow, G., Baffaut, C., Gitau, M. W., Qi, Z., Amatya, D. M., & Green, T. R. (2015). Parameterization guidelines and considerations for hydrologic models. *Transactions of the ASABE*, *58*(6), 1681–1703. https://doi.org/10.13031/trans.58.10709

Meira Neto, A. A., Roy, T., de Oliveira, P. T. S., & Troch, P. A. (2020). An aridity index-based formulation of streamflow components. *Water Resources Research*, *56*(9), e2020WR027123. https://doi.org/10.1029/2020WR027123

Nearing, G. S., Ruddell, B. L., Bennett, A. R., Prieto, C., & Gupta, H. V. (2020). Does information theory provide a new paradigm for Earth science? Hypothesis testing. *Water Resources Research*, *56*(2), e2019WR024918. https://doi.org/10.1029/2019WR024918

Pagano, T., Hapuarachchi, P., & Wang, Q. J. (2010). *Continuous rainfall-runoff model comparison and short-term daily streamflow forecast skill evaluation*. CSIRO. https://doi.org/10.4225/08/58542c672dd2c

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of hydrology*, *279*(1–4), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Schaefli, B., & Gupta, H. V. (2007). Do Nash values have value? *Hydrological Processes*, *21*(15), 2075–2080. https://doi.org/10.1002/hyp.6825

Sezen, C., & Partal, T. (2019). The utilization of a GR4J model and wavelet-based artificial neural network for rainfall–runoff modelling. *Water Supply*, *19*(5), 1295–1304. https://doi.org/10.2166/ws.2018.189

Sudriani, Y., Ridwansyah, I., & Rustini, H. A. (2019). Long short term memory (LSTM) recurrent neural network (RNN) for discharge level prediction and forecast in Cimandiri River, Indonesia. In *IOP Conference series: Earth and environmental science* (Vol. 299, p. 1012037). IOP Publishing. https://doi.org/10.1088/1755-1315/299/1/012037

Valéry, A., Andréassian, V., & Perrin, C. (2014). 'As simple as possible but not simpler': What is useful in a temperature-based snow-accounting routine? Part 2–Sensitivity analysis of the Cemaneige snow accounting routine on 380 catchments. *Journal of hydrology*, *517*, 1176–1187. https://doi.org/10.1016/j.jhydrol.2014.04.058

Vrugt, J. A., Gupta, H. V., Dekker, S. C., Sorooshian, S., Wagener, T., & Bouten, W. (2006). Application of stochastic parameter optimization to the Sacramento soil moisture accounting model. *Journal of Hydrology*, *325*(1–4), 288–307. https://doi.org/10.1016/j.jhydrol.2005.10.041

Wu, W., May, R. J., Maier, H. R., & Dandy, G. C. (2013). A benchmarking approach for comparing data splitting methods for modeling water resources parameters using artificial neural networks. *Water Resources Research*, *49*(11), 7598–7614. https://doi.org/10.1002/2012WR012713

Zhang, J., Zhu, Y., Zhang, X., Ye, M., & Yang, J. (2018). Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. *Journal of hydrology*, *561*, 918–929. https://doi.org/10.1016/j.jhydrol.2018.04.065

Zheng, F., Maier, H. R., Wu, W., Dandy, G. C., Gupta, H. V., & Zhang, T. (2018). On lack of robustness in hydrological model development due to absence of guidelines for selecting calibration and evaluation data: Demonstration for data-driven models. *Water Resources Research*, *54*(2), 1013–1030. https://doi.org/10.1002/2017WR021470