Towards a Multi-Representational Approach to Prediction, Understanding, and 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 85721, USA

Corresponding authors:

Luis A. De la Fuente (ldelafue@email.arizona.edu), ORCID: 0000-0001-6979-0547

Hoshin V. Gupta (hoshin@email.arizona.edu), ORCID: 0000-0001-9855-2839

Laura E. Condon (lecondon@email.arizona.edu), ORCID: 0000-0003-3639-8076

KEY POINTS

- The representation underlying a model pre-determines what can be learned, which argues for a flexible approach to scientific investigation.
- By employing multiple representational approaches, we improve our chances of properly understanding the underlying Data Generation Process.
- Such an approach helped in understanding how to model precipitation-streamflow response across hydro-geo-climatologically diverse Chile.

KEYWORDS

Representation, Machine Learning, LSTM, Random Forest, GR4J, conceptual model, lumped water balance model, Understanding, Discovery, Hydrological Processes, Catchments, Hydro-geo-climatology

ABSTRACT

A key step in model development is 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 that 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 complementary 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 dataset 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 RF being most supportive of interpretation.

Overall, the complementary nature of the three approaches suggests the value of adopting a multi-representational framework in order to more fully extract information from the data. Our results show that a multi-representational approach better supports the goals of prediction, understanding, and scientific discovery in Hydrology.

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 complementary modeling approaches (one lumped water balance 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 system memory, attributes that correlate with transferability across different regions, and informational deficiencies of the available dataset. Overall, this study suggests the value of adopting a general multi-representational framework to better support prediction, understanding and scientific discovery in the Earth and Environmental Sciences.

1 INTRODUCTION

1.1 The Problem of Selecting an Appropriate Representational System

[1] When developing any dynamical systems model, be it conceptual or data-based, a key step is the selection of an appropriate representational system. This step includes two aspects: (1) the choice of system inputs (driving variables) and boundary conditions relevant to predicting the dynamical evolution of the system states and outputs, and (2) the formal mathematical/algorithmic structure used to construct the input-output (or input-state-output) mappings that are hypothesized to characterize the system (*Gupta et al, 2012; Gharari et al, 2021*).

[2] In hydrology, as in other environmental disciplines, the selection of system inputs and boundary conditions determines the nature and quality of the information that can be brought to bear on the prediction problem – without adequate and informationally relevant data, the task of predicting the system outputs is doomed from the outset. Having done so, the mathematical/algorithmic representational system selected for constructing the input-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* that we can obtain.

[3] For example, a dynamical process-resolved (often called physically-based) catchment-scale hydrological model 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*". If a mass/energy-conserving spatially-lumped bucket-type state-space representation is implemented, then one may be able to obtain insights into aggregate catchment-scale soil moisture storage variations (and their vertical distribution in the soil zone), whereas a spatially-distributed finite-element/difference representation may be used to infer the dynamic evolution of soil-moisture (and other state-variables and fluxes) in three-dimensional space. Such models focus on preserving and tracking "*mass and energy (sometimes also momentum) flows*" through the system.

[4] On the other hand, if a data-based machine-learning type of representation is implemented, then the focus is on preserving and tracking "*information flows*" through the system. In this case, the model may not be as well suited (as a process-resolved representation) to inferring state variables such as "*soil moisture*" or fluxes such as "*percolation, recharge and interflow*" that are constrained (by physics) to obey conservation principles, unless appropriate regularization constraints are also implemented.

[5] In summary, the representational structure (of both the data and the model) selected for the analysis imposes strong constraints on the questions we can ask, the results we can get, and the inferences we can reasonably hope to perform. Consequently, we can expect, a priori, that different representational strategies may provide different perspectives on the factors and processes governing the generation of system behaviors.

1.2 Models as Complementary Perspectives on Reality

[6] For any given application, it can be challenging to determine what the most appropriate model structure might be. 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 modeling approaches brings to mind the classic story of the "*blind people and the elephant*" where each person's interpretation of what constitutes an "*elephant*" is based on their experience being limited to some very specific aspect of the animal, while also being limited by their ability to map that experience onto their previous knowledge (i.e., they are limited by what they can "*recognize*").

[7] Given that any model is a "*relevant simplified representation*" of the world, where the simplifications typically reflect personal biases and preferences, it is quite possible for there to be as many viable "*representations*" to choose from as there are people working on a given problem. In practice, however, only some of these viable representations will tend to perform well (when evaluated against data), thereby considerably decreasing the number of potentially suitable options. Nevertheless, it can remain difficult to identify a single "*best*" model, since many different representations can be found to exhibit similar levels of performance (*Clark et al, 2011*).

[8] Returning to the metaphor of the "*blind people and the elephant*", rather than asking which of the representations is (somehow) "*the best*", one might instead consider whether the multiple complementary perspectives offered by the 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 (in at least some partial sense that contributes to a more complete overall point of view), we can hope to make progress towards uncovering the "*real*" nature of the underlying *Data Generation Process* (DGP).

1.3 The Potential Offered by Lumped Water Balance Modeling

[9] Lumped water balance models, that are structurally and behaviorally isomorphic to the system, are the mainstay of how understanding is developed in science in a very simplified (conceptual) representation. Such representations 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.

[10] The development of such models is structured as a sequence of conditional hypotheses, beginning with the governing conservation laws, and proceeding through the specification of the system architecture, process parameterization equations, and property-to-parameter relationships; see extensive discussion in *Gharari et al (2021)*. Different choices at each stage of development can give rise to different "*compound hypotheses*"; for examples of modular modeling systems in hydrology, see (*Fenicia et al, 2011; Clark et al, 2015; Craig et al, 2020*). This facilitates a multi-hypothesis approach to scientific investigation (*Clark et al, 2011*), in which each model represents a point within a hypothesis space that is strongly constrained by physics, assumptions, and prior knowledge.

[11] The strength of this representational approach is the ability to constrain discovery to model structures that are consistent with physical principles. An important consequence is that "*meaning*" can be ascribed to the various components, fluxes, and state variables of the model, making it possible to transfer understanding between locations, and to generalize to classes of systems that share similar representational properties.

[12] However, this strength can become a weakness when, by imposing strong priors, we limit the ability of the model to learn explicitly and directly from data, and to discover things

that are inconsistent with the space of hypotheses explicitly covered by the priors (*Gharari et al 2021*).

1.4 The Potential Offered by Machine Learning (ML)

[13] Conversely, due to its ability to extract complex relationships from large datasets, Machine Learning (ML) has 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. Recently, applications to hydrology have demonstrated excellent results in different areas. To mention just a few, *Long-Short Term Memory* networks were successfully applied to large-scale streamflow prediction (*Kratzert et al, 2018*), to a long record for one catchment (*Hu et al, 2018*), to the estimation of water table depth for five agricultural areas (*Zhang et al, 2018*), and to 5-day-ahead prediction (*Sudriani et al, 2019*).

[14] Overall, the power of ML-based representations arises from their theoretical and practical ability to approximate any input-state-output mapping to an arbitrarily high degree of accuracy, given sufficient data. Consequently, ML has emerged as a powerful complement/alternative to the hypotheses-driven process-based approach to hydrological modeling. However, as with the lumped water balance approach, each ML algorithm/approach is based on a different mathematical perspective about how to represent the structures underlying a given data set, and/or on how to represent and extract information contained in the data. Accordingly, when different ML algorithms are applied to a given data set, each is (also) likely to provide a different and, in general, complementary 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 particular value offered by each perspective and exploit it to obtain a more comprehensive picture of the underlying system.

1.5 Objectives and Scope of this Paper

[15] The objective of this paper is to explore how a multi-representational approach can help to extract relevant information from a dataset, with a view to improving prediction, understanding, and discovery. Our specific goal is to use such an approach to develop an understanding of catchment hydrology across the hydro-geo-climatologically diverse extent of Chile. Rather than the traditional strategy of implementing a single pre-selected computational model code to the entire country, or perhaps a different model code to hydrologically different parts of the country, we implement three complementary representational approaches (model structures) across the entirety of Chile. These include a a lumped water-balance model (based in a physical understanding of watershed behavior) and two machine learning models (based on information extracted from historical observations). Our focus is on understanding the strengths and weaknesses associated with each representational approach, and on exploring the potential richness of inferences that a multi-representational approach can support.

[16] In the next section, we introduce the problem of catchment-scale hydrological forecasting in the context of the particular hydro-geo-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 to be exploratory, with a view to improving our

130 understanding of how a multi-representational approach can be exploited in the service of
131 enhanced scientific discovery.

132 2 THE CHALLENGE OF STREAMFLOW PREDICTION ACROSS
133 HYDROLOGICALLY DIVERSE CHILE

134 [17] Prediction of streamflow at national scales is challenging, due to the multitude of
135 relevant factors that can vary simultaneously across time and space. In particular, the ability
136 of hydrological models to generalize can be poor in regions where the spatial variability of
137 dynamical forcings and static attributes is large (*Malone et al., 2015*). This is especially
138 relevant to Chile, which is characterized by tremendous geo-hydro-climatic variability, both
139 along its 4,270 km (2,653 mi) North-South extent and also from East to West (***Figure 1***). At
140 one extreme, Northern Chile is home to the driest desert in the world, containing regions
141 where no precipitation has been recorded for more than 25 years. At the other extreme, more
142 than 5000 mm/year of precipitation has been recorded in parts of the south, where there are
143 also permanent icefields.

144 [18] Bordered by the Pacific Ocean to the West and Argentina to the East, the country
145 averages just 175 km (109 mi) in width, while the North-South running Andes Mountain
146 range rises to the highest elevation in South America (6,959 m or 22,831 ft). Moreover, a
147 second mountain range, with lower elevations, runs parallel to the coast along almost the
148 entire country. Owing to the high elevations of the mountain ranges, precipitation in the
149 headwater catchments occurs mainly as snow, due to which the corresponding streamflow
150 peak will appear many days or even weeks after the precipitation event. In contrast, where
151 liquid precipitation occurs in catchments with high slopes, the times of concentration can be
152 shorter than one day. Other factors, including the variability of forest fraction, degree of
153 human intervention, and valleys created between the two mountains range, are also strongly
154 related to the availability of water in the long term.

155 [19] This immense variability in geo-hydro-climatic conditions poses a considerable
156 challenge for any modeling system, and especially for lumped water balance representations
157 where the model structure must be selected in advance. As such, Chile presents a perfect
158 opportunity to explore the possibility of developing modeling techniques that can deal with
159 large geo-hydro-climatic variability, and even exploit it to achieve better model performance.

160 3 STUDY METHODOLOGY

161 [20] This section presents and discusses our study methodology, including the dataset used
162 (section 3.1), three representational methods used (section 3.2), and issues related to the
163 experimental design (section 3.3).

164 3.1 Dataset

165 [21] For the purposes of this study, we will use mainly the information provided by the
166 catchment-scale CAMELS-CL dataset (*Alvarez-Garreton et al, 2018*). This dataset includes
167 11 variables and 105 categorical and numerical attributes for 516 Chilean catchments. For
168 model development and evaluation, we selected 322 catchments selected to span the country
169 and to have a minimum streamflow record length of 7 years. The literature suggests that 2-3
170 water-years of daily data represents a minimum record length for calibration of conceptual
171 process-resolved models (*Gupta and 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

[22] To develop an improved understanding of the nature of catchment-scale hydrology at the national scale across Chile we will use a multi-representational approach. Clearly, this approach must be consistent with the available data (Section 3.1). With this in mind, we chose to investigate three complementary representational strategies – one being a lumped water balance model, and the other two being ML-based modeling strategies. While additional representational strategies could also have been included, these three arguably represent sufficiently different approaches to extracting information from data to support the objectives of this study.

[23] For the lumped water balance model, we chose the GR4J dynamical systems model (*Perrin et al, 2003*), due to its relative parsimony and the reports of good performance in other studies (*Kunnath-Poovakka & Eldho, 2019; Sezen & Partal, 2019; Pagano et al., 2010*), and because the catchment-scale data required for its implementation is available (see ***Appendix Table A-2***). For the ML-based modeling strategies, we selected the LSTM network (*Hochreiter & Schmidhuber, 1997*) and the RF regression-tree algorithm (*Breiman, 2001*). Further details about these modeling strategies are provided below.

3.2.1 The GR4J Lumped Water Balance Representation

[24] The GR4J model is a parsimonious lumped water balance (process-based) representation of daily time-step spatially-lumped catchment-scale hydrology, whose input-state-output behavior can be controlled by adjusting four tunable parameters (***Figure 2a***). GR4J is the outcome of several studies that evaluated a variety of model structures having various levels of complexity, using data from 429 catchments with differing geo-hydro-climatic conditions. While more complex representational structures are available, such as GR5J (*Le Moine, 2008*) and GR6J (*Pushpalatha et al, 2011, CemaNeige; Valéry, 2010*), GR4J provides a relatively simple representative of this class of models, and a search for the "*best*" such model for Chile is not part of the scope of this study.

[25] Importantly, the GR4J representation seems well suited to application across Chile for several reasons, including the fact that it contains a flow path without storage that gives it the ability to simulate the very rapid precipitation-streamflow response that is characteristic of the steep surface slopes that occur across portions of Chile. Further, it has the desirable feature that it includes a parameter that can be tuned to permit the model to either "*import*" or "*export*" water into its main routing storage tank to enable the model to better match the input-output behavior of the system as expressed by the observed time-series data.

[26] As such the GR4J model serves as a kind of lumped water balance benchmark against which the ML-based representations can be compared (no pre-existing benchmark model calibrated across the entire country exists). Following the traditional approach, the parameters of the GR4J model are tuned (calibrated) "*locally*" to be specific to each catchment. This is in contrast to the ML-based approaches (see below) where the parameters

(network weights and/or biases) are tuned "*globally*" to represent all catchments across the study domain (in our case the entire country of Chile). Because GR4J is calibrated locally, it can be considered to represent a performance benchmark that one would want a globally-tuned data-based ML approach to be able to exceed, particularly when requiring the model to generalize well (i.e., to perform well on catchments that are withheld from the model development data set). For more details on the structure of the GR4J model, please see *Perrin et al (2003)*.

3.2.2 The LSTM ML-based Representation

[27] The LSTM model (***Figure 2b***) is a fully connected Recurrent Neural Network (RNN) with the ability to learn from the past. The recurrence feature is akin to that of the tank-like components of physically-based catchment models, but where the nature of the relationships between inputs, state variables, and outputs is learned from the data. The structure of an RNN is based on the linear superposition of non-linear basis functions (known as activation functions), which enables the network to closely match the nature of the true relationships that underlie the data.

[28] The LSTM-based representation is somewhat more complicated than the traditional RNN because additional components (called gates, or gating functions) are used to contextually control how the flow of input, output, and previous state information affects the network response at each time step. In this sense, the LSTM cell-states can be thought of as non-linear extensions of the classic linear reservoir component commonly used for hydrological modeling (see ***Table 1***). As such, the LSTM can be interpreted as a representation of Dynamic Information Storage, where the gating functions act as contextually variable resistances to the flows of different kinds of information. For more detail about the equations used in the LSTM model, please refer to *Kratzert et al. (2018)*.

[29] By drawing an analogy with the linear reservoir (***Table 1***), it is possible to interpret each of the functions and components in the LSTM. The function *g(x,h)* computes a linear combination of inputs and outputs, adds a bias term, and then non-linearly transforms the result onto the range [-1, 1] using the hyperbolic tangent equation. The quantity *c(t)* can be understood as an informational "*state variable*" of the system, where the information carried by *c(t)* is contextually updated based on the values of weights *f(t)* and *i(t)* applied to the past storage information and drivers, respectively. Whereas the linear reservoir computes the output as a constant proportion of the storage *S(t)*, the LSTM defines this proportion dynamically through the gating function *o(t)*. In summary, the gating functions *f(t)*, *i(t),* and *o(t)* act as dynamic amplification factors that can take on values between]0,1[controlled by a sigmoid-shaped activation function. Meanwhile, the variables *g(t)* and *c(t)* represent different informational representations of the input, output, and storage, normalized using the hyperbolic tangent function to take on values between]-1.1[.

[30] Due to the isomorphic relationship of the LSTM to the linear reservoir, it becomes possible for the structure of the GR4J model to be emulated using an LSTM. However, because the LSTM can exploit the information provided by inputs beyond precipitation and potential evapotranspiration, it becomes more difficult to apply physically-based interpretations to the behaviors of its state variables.

3.2.3 The RF-based Representation

[31] The RF is a regression-tree methodology (***Figure 2c***) that adopts the classic strategy of "*divide and conquer*" to construct an approximation of the input-output mapping expressed by the data. The RF algorithm searches for an "*optimal*" partitioning of the input space, that maximizes the similarity within each output cluster that represents a leaf of the decision tree. The similarity measure used is the weighted average dispersion of the outputs within a cluster, where dispersion is measured as the sum of squared deviations from the mean of the cluster members (L2 norm). It is typically assumed that the input-space partition resulting in the smallest average dispersion, weighted by the number of elements, is best. This process is repeated within each partition until a prespecified minimum number of elements remains within a subset, and/or until a predefined number of splits have been conducted.

[32] The RF implements a piecewise-constant approximation of a complex continuous input-output mapping, where for each new split, we look for the minimum difference between the cluster means (predictions) and their corresponding target values (output data). To avoid this deterministic process becoming highly biased by the specific data sample used to construct the decision tree, the RF approach uses a random sample (selected with replacement) from the data set to construct each decision tree and repeats the process multiple times to generate a "forest" (ensemble) of decision trees. The use of randomized ensembles helps to reduce overfitting, while randomness in the input selection helps to improve the accuracy of the classifier and regressor algorithm (*Breiman, 2001*). The final prediction is generated as the average prediction made by each of the trees in the forest. From the perspective of interpretability, it is easy to examine each input-space split associated with the model predictions, making it relatively easy to understand the steps connecting the input to the outputs, without the need to track any intermediate variables or states.

[33] Another issue is related to the nature of the problem. Because streamflow is the result of complex processes within a Dynamical Environmental System, knowledge of the system state can be very important for characterizing how the system will respond to new inputs given past conditions. In other words, the streamflow on a specific day is not just the result of what happens on that day but also depends on what has happened in the past. Moreover, the history of a catchment can be understood in terms of different time scales, such that the current streamflow response is related to both what happened recently (short-term memory) and what has happened in the past months or even years (long-term memory), due to persistence in the behavior of the system. These two kinds of memory are not explicitly represented by the structure of the RF. Accordingly, it is the responsibility of the modeler to ensure that the input data contain variables that provide such memory-related information that can be used by the RF in place of state variables to track both the short-term and long-term dynamics of the system. This feature can be interpreted as both a "*pro*" and a "*con*" of the RF approach; it is pro in the sense that it allows the modeler to exert better control over the model by injecting physical understanding, while it is a con because it imposes higher demands in the form of data preprocessing.

[34] The ML-based RF and lumped water balance GR4J representations are harder to compare (than the LSTM and GR4J representations) because of RF's lack of state variables that mediate between the inputs and outputs, and because of the piecewise-constant nature of the RF representation. However, this does not mean that memory and dynamics are not considered by the RF, because state variables can be thought of as representing the aggregate

effects of an infinite number of past system inputs. Accordingly, provided that the RF is fed with a sufficiently long history of past system inputs, the representation can learn to construct input-space splits that emulate those that would have resulted from the tracking of state variable information.

3.3 Experimental Design

[35] 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 recommended best model development practices for each representational type and compare the results so obtained. However, the overall methodology conforms to a common framework, so as to enable comparative analysis. Accordingly, all comparisons are based on the use of the same data and performance metrics for model development and evaluation.

[36] ***Appendix A1.1*** discusses how we partitioned the data into , that was done using three periods for the purposes of model calibration, selection, and evaluation, consistent with the ML literature. The ***Appendices A1.2*** and ***A1.3*** summarize the variables and attributes used in model development. ***Appendix A1.4*** discusses how the warm-up period used for both GR4J and LSTM was selected. ***Appendix A1.5*** discusses the process of parameter/hyperparameter selection for each model. ***Appendices A1.6*** and ***A1.7*** describe the metrics and algorithms used, and ***Appendix A1.8*** describes how the out-of-sample testing dataset was generated.

4 EXPERIMENTAL RESULTS

[37] In keeping with the objectives of this paper, our analysis pays special attention to how the different representational approaches can be used to make inferences regarding various characteristics of the hydrological processes that underlie the data. Accordingly, this section consists of two parts. In section 4.1 we investigate issues of overall understanding and/or discovery, such as system memory and feature importance, enabled by the multi-representational approach. In section 4.2 we investigate how the different models constructed using the different representational approaches performed in terms of the ability to generalize in space and time.

4.1 Understanding Enabled by the Multi-Representational Approach

[38] 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. This happens because each time that the representation assimilates a new piece of information, it updates its internal structure/parameters, which can be understood as *learning* about *changes in the internal state* of the system. Therefore, the final "learned" version of the representation (the trained model) encapsulates a considerable amount of information that can be subject to analysis. Here, we investigate how data skewness, system memory, and the relative importance of surrogate variables provide insights into the underlying nature of the DGP.

4.1.1 The Box-Cox Transformation Parameter

[39] The Box-Cox power transformation $Y = (y^{\lambda} - 1)/\lambda$ is commonly used in statistical analysis (*Box and Cox, 1964*) to account for skewness in the data (represented here by y). If $\lambda = 1$ the variable Y has essentially the same distributional properties as y (no transformation beyond a shift of origin), while by setting λ to be smaller or larger than 1.0, the skewness of Y can be reduced or increased, respectively, relative to y. In hydrology, strong skewness of the streamflow distribution (corresponding to values of $\lambda \to 0$) is indicative of precipitation being the main driver of system dynamics, while weak streamflow skewness may indicate the dominance of groundwater, snowmelt, or other processes that act as low-pass filters in the generation of streamflow dynamics.

[40] In our implementation of the GR4J model, the λ parameter was allowed to vary with location to allow the model development process to account for the hydro-geo-climatic variability of skewness in the streamflow data. ***Figure 3*** presents the distribution of spatially-varying '*optimal*' λ values obtained for the 322 catchments in the model development dataset. While the optimized λ values vary across the full range tested (0.00 to 2.00), a high concentration (~35%) of values fall in the first bin ($\lambda \sim 0$), consistent with the traditional use of a logarithmic transformation when calibrating daily-time-step models in hydrology (*Hassan & Hassan, 2021*). This result is consistent with the large range of possible streamflow values that can occur across Chile, where streamflow can vary over several orders of magnitude in catchments that respond quickly to intense precipitation events. In contrast, ~10% of the catchments are associated with $\lambda > 1$, where very little variability in the range of streamflow magnitudes can be found (e.g., where precipitation-runoff is weak, or where baseflow tends to be the dominant streamflow generating mechanism).

[41] In the implementations of the LSTM and RF models, single values of lambda were applied across the entire country. ***Figures 4a*** and ***b*** present cumulative density functions (CDF) of KGEss metric performance (*Gupta et al 2009, Knoben et al 2019*; see definitions in ***Appendix A1.7***) computed over the selection period (validation), for different choices of lambda. For each CDF, we report the area under the curve to serve as guidance for selecting the best value of lambda (treated as a hyperparameter), where smaller areas correspond to better overall performance. In contrast with GR4J, where the average value of λ is close to zero, ***Figure 4a*** indicates that better performance of the RF model is obtained with λ close to one, which corresponds to <u>*not*</u> applying a transformation to the streamflow data. This makes sense in retrospect, because when the decision tree splits the data into different clusters it is inherently able to account for skewness, and so the addition of a transformation does not necessarily provide any significant additional value to the model development process (note the relatively weak dependence of performance on λ). Based on this observation, when developing the RF model, we fixed the value to $\lambda = 1$ (corresponding to no transformation) and report only results obtained using this value.

[42] Note that when implementing the LSTM, we obtained better results (***Figure 4b***) by using a '*global*' standardization of the data (subtracting the mean and dividing by the standard deviation of streamflow values computed from the entire Chile-wide dataset), rather than a *'local'* standardization (where the mean and standard deviation were regressed against aridity index). Further, ***Figure 4b*** indicates that the performance of the LSTM model so obtained is not sensitive to the choice of λ, and so we again fixed the hyperparameter $\lambda = 1$

(corresponding to no transformation) for both standardization approaches, and report only results obtained using this value.

[43] These results are interesting because the value of λ encodes information about the DGP when using the GR4J representation but not when using the ML-based representations. One might speculate that this result is a consequence of the fact that the '*local'* GR4J modeling approach permits each catchment to be represented by a different value for λ, while the '*global*' ML-based modeling approaches require the specification of a single country-wide value. However, if this were true we might expect the ML-based '*globally optimal'* λ values to converge to something like the mean or median of the GR4J-based distribution of '*locally optimal*' λ values. Given that this is not the case for both the (quite different) LSTM and RF representational approaches, it is more likely that the ML models are internally able to address problems related to data skewness in some other manner. Nonetheless, this remains an interesting issue for future investigation.

4.1.2 System Memory

[44] For the ML-based representations, an important property is the manner in which the "*system 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 representation, which tracks system memory exclusively through its state variables. ***Figures 5a*** and ***b*** show how the CDFs of the model performance vary with different values of the lag-time hyperparameter for the RF and LSTM models respectively.

[45] Consider first the RF model. For catchments with KGEss better than ~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 less than ~0.45 the results are insensitive to the value of the lag-time hyperparameter. Further, the marginal performance improvement declines, on average, as the lag-time is increased. A similar result is found for the LSTM model, but now we see an additional region of improvement when going from 128 to 270 days, occurring mainly in catchments with KGEss values below ~0.85.

[46] Taken together, 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 a system "*memory*" of around 32 days and the other related to one of around 270 days. We will revisit this topic in the next section, where we see that this difference in length of system memory is correlated with hydro-geo-climatic attributes. Note that this kind of information about systemic differences between different catchments in the study region is somewhat more difficult to infer from the relatively simple GR4J representation used in this study.

4.1.3 Feature Importance

[47] Another interesting aspect of ML-based approaches is the manner and flexibility by which the relative informativeness/importance of different (spatially-varying) hydro-geo-climatic attributes can be assessed. As a consequence, it is (in general) easier to detect which attributes are more/less important when explaining the predictive power of an ML-based model. Whereas this is, in principle, also possible using a conceptual/lumped 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.

[48] In particular, for ML-based models, tools such as the Scikit-learn Python library (*Pedregosa et al., 2011*) facilitate a simple sensitivity analysis that permits a relatively straightforward exploration of the importance of each input variable or system attribute in contributing to the predictions. For conceptual/lumped representations, this exploration is complicated by the fact that the importance/informativeness of a given variable or attribute is mediated by the specific structural assumptions encoded by the system architecture and process parameterization equations chosen for the model. In contrast, the ML-based representational structure is not quite so strongly pre-determined and is therefore, arguably, less likely to bias any inferences of relative feature importance. Of course, this is not *entirely* true since different ML-based approaches also (unavoidably) encode different representational assumptions about how to map system inputs to outputs. However, the relative flexibility of ML-based representations (as well as their focus on the strengths of "*informational*" relationships) should, in principle, enable interesting (and hopefully useful) insights regarding relative feature importance to be inferred. That inference is neither as simple or as direct as in a lumped water balance model, due to the fact that attributes and variables are connected through hundreds (even thousands) of parameters. Nonetheless, tools such as the one mentioned above are increasingly making such analysis possible.

[49] The consequence is that an ML-based assessment of the relative importance of hydro-geo-climatic attributes can provide potentially valuable information regarding what system attributes are likely to be important when constructing a (better) conceptual/lumped model, and regarding how these attributes are likely to vary across space, and must therefore be considered in order to achieve a lumped water balance representation that generalizes well at the large scale.

[50] Here we analyze the information about feature importance that is an inherent property of the RF-based model, where the variables selected for data-space thresholding earlier in the tree (e.g, at the first split) can be interpreted as being more '*fundamentally'* or '*globally'* important to the construction of a decision tree. ***Figure 6a*** indicates that the most important attribute is an "*aridity*" index (aridity_cr2met, computed using the CR2MET precipitation product), 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 versus arid regions). While this observation is not novel (*Neto et al, 2020; Booij et al, 2019, Chen et al, 2019*), it is consistent with the fact that lumped catchment-scale water balance representations, with their fixed architectures and process-parameterization equations, are typically not able to generalize well across different hydro-geo-climatic conditions.

[51] Of course, this does not imply that failure to account for "*aridity*" is, per se, a complete and meaningful explanation for 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 6a*** 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. This is consistent with the interpretation for the Box-Cox transformation parameter λ used with the GR4J model.

[52] In our case, the first split (***Figure 6b***) that occurs in most trees of the RF model occurs at an aridity index threshold of 0.6 mm/day. While, for any given study area, the precise value at which this split occurs will depend on the distribution of wet and arid catchments in the dataset, this observation suggests that different streamflow generating representations may be required for the model to perform well in regions that are "*energy-limited*" as opposed to "*water-limited*". When a similar analysis is performed for the precipitation threshold, we find that the nature of the streamflow response is different for values above/below ~10 mm/day. From this, we could hypothesize that more than ~10mm/day of precipitation is required (on average) to generate surface runoff, but of course, much more analysis would need to be done to test such a hypothesis.

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

[54] The main point of these two rather simple (even trivial) examples shown in ***Figure 6*** is that the RF representation facilitates a kind of analysis that can provide interesting information that is not easily obtained using either the GR4J or LSTM representational approaches. In this sense, the RF approach provides a strong complement to other representational approaches when our goal is to use modeling in support of scientific discovery and understanding.

4.2 Comparative Analysis of Similarities and Differences in Performance

[55] The relative ability of any properly trained model to perform well on independent "*evaluation period*" datasets can be considered indicative of well the corresponding representational approach supports discovery about the underlying DGP. However, even if all of the models tested on the evaluation period provide essentially identical values for some aggregate performance metric (such as KGEss or NSE; see Appendix A.1), deeper analysis may reveal systematic differences in model simulated behaviors that the aggregate metric is not capable of distinguishing between (*Gupta et al 2008, 2009*).

[56] For our rainfall-runoff modeling case study, such behaviors may include things such as the simulated-to-observed long-term water balance and variability ratios, and the timing and shape (measured, for example by cross-correlation strength between the simulated and observed time-series of model output response). By examining how well each representational approach reproduces such behaviors (when trained, as is customary, on an aggregate performance metric), we can hope to obtain insights into the strengths and weaknesses of each, which is the objective of this paper. This section investigates overall and spatial patterns of such differences in model behavior/performance, with a view to understanding the manner and extent to which the models (developed using different representational approaches) are able to generalize well in space and time.

4.2.1 Overall Performance

[57] First, we examine the distributions of overall model performance across the country. ***Figure 7a*** shows the CDFs of evaluation period performance (as measured by KGEss) for all locations where $KGEss > 0$ (where predictions are, on average, better the "no-model" prediction that simply uses the observed mean; *Knoben et al 2019*). Similar results were

obtained using NSE (not shown; for details see *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 model fails to meet the $KGEss > 0$ threshold at only ~5% of the catchments, as opposed to ~11% for LSTM and ~22% for RF.

[58] Regarding the first result, the superior performance of the LSTM 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).

[59] Regarding the second point, given that all three representations are trained using (almost) the same input-output information (GR4J model uses only precipitation and evapotranspiration), this result suggests that there are hydro-geo-climatic conditions under which the GR4J representation provides useful (lumped water balance) information that is not directly inferable from the available data by the LSTM and RF representations. 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. In a recent study by *Hoedt et al. (2021)*, a "*mass-conservative*" LSTM model was found to be able to learn a good state-variable representation of the dynamics of snow storage, but such findings would need to be tested at larger scales over a variety of hydro-geo-climatic conditions before more general conclusions can be drawn.

[60] Next, we examine the distributions of the decomposition components of KGEss (see definitions in Appendix A1.7). While aggregate metrics such as KGEss (and NSE etc.) 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 7b*** provides further discriminatory information by plotting the CDF of model Bias Ratio, where values larger (smaller) than 10^0 $(= 1)$ indicates a tendency to overestimation (underestimation).

[61] This plot reveals that the GR4J and LSTM models, that have the explicit ability to simulate system dynamics, tend (on average) to be unbiased, whereas the RF 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 lower bias) than the two ML-based models, with the RF model being the worst. However, for situations where the models tend to underestimate the mean (Bias Ratio < 1.0) that situation is reversed and the two ML-based models perform better than GR4J, with the RF model being the best. Similar results were found for the Standard Deviation Ratio (results not shown).

[62] So, while the LSTM is statistically superior in terms of overall KGEss performance for the majority of catchments, the situation is clearly more nuanced – with each representational type providing different characteristic abilities to simulate various attributes of streamflow, despite the fact that all the model types were trained using (almost) the same data. This

supports our contention that a multi-representational approach can aid in scientific investigation and discovery, particularly when faced with significant hydro-geo-climatic variability.

[63] Meanwhile, the use of multiple metrics, that target different (ideally complementary) signature properties of the data (*Gupta et al 2008*), can assist in the extraction of different kinds of useful information, enabling inferences about different aspects of the input-(state)-output response of the system.

4.2.2 Spatial Patterns of Performance

[64] The previous statistical analysis is informative about the overall properties and capabilities of the different representational types. However, it is of little value when needing to make statements about actual performance at any catchment. In this section, we investigate how the different representational types perform across the variety of different hydro-geo-climatic conditions that characterize Chile.

[65] ***Figures 8a*** and ***b*** explore the relationship between model performance and two interesting hydro-geo-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 hydro-geo-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 being characterized by intermediate degrees of wetness and considerable variability in elevation.

[66] The curves in ***Figure 8a*** 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°], 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 existing GR4J architecture. In contrast, the flexibility of the ML-based representations enables them to better account for such phenomena.

[67] 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 strong slopes (rapid elevation changes and very short times of concentration) and relatively greater aridity (see next sub-section) than the mid/south-central and southern regions. Here, RF performs particularly poorly, which may be attributable to the fact that it does not have access to data with greater than 16 days lag time and is, therefore, unable to account for longer (seasonal or annual time-scale) system memory, unlike GR4J and LSTM.

[68] Finally, the northern part of the country [latitude -25° to -18°] contains the Atacama Desert, which is the aridest region in the world and has moderate slopes. Here, RF and LSTM both exhibit better performance than GR4J. The inability of the latter to simulate the hydrologic behavior of such extreme conditions is likely due to the fact that GR4J was developed to represent the very different hydro-geo-climatic conditions that characterize France. Meanwhile, the relatively poor performance achieved by both ML-based models

suggests that the variables that make up the existing CAMELS-CL dataset are not sufficiently informative about the particular input-state-output dynamics of the catchments in this region to enable a robust and accurate model to be developed, and that other variables and attributes should be added to improve model performance (more on this later).

[69] The curves in ***Figure 8b*** show smoothed trajectories for how KGEss performance varies with the *Aridity Index* (computed as the mean of *aridity_cr2met* and *aridity_mswep*). 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 hydro-climatic conditions become increasingly more arid (water-limited). Interestingly, the performance of both GR4J and LSTM (that have the ability to 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 to 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.

[70] 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 towards 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 fraction, 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*).

[71] Another interesting observation is that the system memory associated with streamflow generation from precipitation is different for energy-limited (wet) and water-limited (arid) catchments. Referring to ***Figure 5b***, we see that whereas the majority of catchments show improvement of LSTM performance when provided with ~32 past days of input data (reflecting short-time-scale memory processes), there is a smaller set of catchments with poorer model performance that shows improvement only when provided with 270+ past days of input data (reflecting longer-time-scale memory processes). The indication is, therefore, that when investigating and modeling the streamflow response of catchments, our representation – whether ML-based or conceptual/lumped – must contain structures that make it possible to track memory processes at more than one dominant time-scale, depending on the hydro-geo-climatology of the region. For example, one might consider the need to track at least the short-term (weekly/monthly/seasonal), medium-term (annual), and possibly longer-term (climatological) time scales. Of course, to discover and build representations of longer-term (climatological) rainfall-streamflow response one would typically need more than 7 years of data.

[72] The important point, however, is that the representational type selected for model development should (ideally) make it possible for information about multiple hydro-climatic time scales to be exploited. GR4J and LSTM contain explicit representations (through dynamic state variables and multiple flow pathways) that – to some degree – facilitate this, with the LSTM having a much greater degree of flexibility to do so (which may explain its generally better performance in ***Figure 8b***). However, for reasons explained in Appendix A1.4, our implementation of the data-based RF 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 on the range 1.5 to 8.0. Note that this kind of model-enabled analysis and discovery is not easily achieved if only a purely conceptual/lumped approach had been used in this study; by adopting a multi-representational approach that incorporates both conceptual/lumped and a variety of complementary ML-based modeling strategies, the process of analysis and discovery can be greatly enhanced.

[73] Finally, ***Figures 9a-c*** 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 bad performance). Overall, all three models exhibit a tendency to good performance (KGEss >> 0.5) south of latitude ~33°S, and at the very northern tip of the country (north of latitude ~20°S).

[74] Focusing specifically on the region between latitudes 27°S and 33°S, we see that RF (***Figure 9c***) performs very poorly throughout this part of the country (see also ***Figure 8a***). 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 discover and track longer-term memory processes is likely contributing to its good performance here. As we move westward towards 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 barely better than 0.0, indicating that the model is mainly only able to reproduce the long-term mean value of streamflow. Given that GR4J has explicitly neither the ability to represent the dynamics of snow accumulation and melt nor the long-term dynamics of groundwater, this result makes sense.

[75] ***Figure 9d*** indicates, for these same catchments, which model provides the best evaluation-period KGEss performance (red=GR4J, blue=LSTM, green=RF) across the country. Here we simply report the model with the best evaluation-period KGEss, regardless of whether these KGEss values are statistically distinguishable. No clear pattern emerges, but in general, the blue (LSTM) and red (GR4J) colors dominate, with LSTM generally being the best-performing model across the country. This is consistent with the statistical results (CDF plots) shown in ***Figure 7a***.

[76] Some more nuanced findings emerge from a statistical analysis of KGEss performance by model type, reported in ***Table 2***. LSTM is the best performing model at 53% (172 of 322) of the catchments, with an excellent median KGEss performance of 0.70. However, this statistic masks the fact that where LSTM fails, it does so very badly – the worst KGEss value is very poor and, consequently, the dispersion of performance is highly skewed. In contrast, the distribution for GR4J, which performs best at only 30% of the catchments and has a

696 median KGEss performance of 0.56, has much lower skewness and dispersion, and achieves
697 positive KGEss values at a greater number (94%) of the catchments.

698 [77] So, while data-based representations may have a greater potential to learn from the data,
699 and thereby achieve greater predictive performance, the conceptual/lumped representations
700 contain valuable regularizing information that may help to prevent model performance from
701 becoming catastrophically poor under conditions where the data is insufficiently informative
702 about the dynamics of streamflow generation. We can speculate, therefore, that GR4J could
703 help to moderate the dispersion associated with the lower percentiles if an ensemble of these
704 three model types were to be used for operational streamflow prediction across Chile. Of
705 course, to implement such a system for Chile, further work would need to be done to
706 generalize the method for estimating parameter values to enable application at ungauged
707 catchment locations. We do not pursue this possibility further in this paper and leave it for
708 future work.

709 ### *4.2.3 Spatial Generalization*

710 [78] The results presented so far indicate that the data-based LSTM has the potential to
711 provide the "*best*" overall performance, while GR4J tends to provide more "*robust*" results
712 in cases where data-based approaches may fail. Meanwhile, the data-based RF is particularly
713 useful for enabling discovery, by providing clues that can lead to hypotheses about what
714 kinds of hydro-geo-climatic processes (and hence data sets) should be incorporated into
715 ongoing model development efforts.

716 [79] However, the previous analysis was for a "*pseudo-independent*" data set, consisting of
717 evaluation-period data from the same catchments that were used for model development. As
718 such, the results may not provide a reliable assessment of the quality of model performance
719 that might be expected at (other/new) catchments that are not part of the model development
720 dataset. ***Figures 10a-c*** and ***Table 2*** report the results of our "*out-of-sample*" analysis, where
721 model performance was assessed on the 167 CAMELS-CL catchments for which less than 7
722 years but more than 1 year of data were available (these catchments were withheld from the
723 model development dataset). Since GR4J parameter estimates are not available for these
724 catchments (an extra parameter regionalization step would be required, that was not pursued
725 in this study), this assessment was done only for LSTM and RF.

726 [80] Overall, the out-of-sample results indicate that LSTM and RF do not show significantly
727 different (relative to each other) spatial distributions of performance. This tends to conflict
728 with the in-sample evaluation results (***Figure 9***), despite the fact that both the in-sample and
729 out-of sample catchment locations are distributed similarly with respect to the *Aridity Index.*
730 When we compare the CDF's of in-sample and out-of-sample performance (***Figure 11a***) for
731 these models, we see that both RF and LSTM exhibit remarkably similar statistical
732 distributions of out-of-sample performance, which suggests that both of these ML-based
733 approaches have a similar ability to generalize to locations that were not included in the
734 model development dataset. There is, however, a larger deterioration in the statistical
735 distribution of model performance from in-sample to out-of-sample for LSTM than for RF.

736 [81] Meanwhile, the CDF of streamflow prediction bias (***Figure 11b***) shows that RF retains
737 the same tendencies in- and out-of-sample tendencies to overestimate the long-term mean
738 streamflow (compare with ***Figure 7b***). This is encouraging, as it suggests the possibility of
739 being able to learn and correct for any long-term predictive bias at a given location.

740 [82] Finally, ***Table 3*** reports a more detailed statistical analysis of KGEss performance by
741 model type, showing that RF slightly outperforms LSTM on most of the statistical indicators.
742 So, while LSTM clearly achieved (in general) better temporal (in sample) generalization, the
743 results for out-of-sample generalization are less definitive. It is possible that the tradeoff
744 between temporal- and spatial-generalization ability is somehow different for each
745 representational type. Further, this may be partially related to the differences in model
746 development strategies – while the LSTM was sequentially fed with the information from
747 different catchments (model parameters are updated using the data from each catchment in
748 turn), the RF model development focuses on finding the best split for all catchments
749 simultaneously, which may make it less sensitive to the new conditions encountered in out-
750 of-sample testing. While this is simply speculative at this point, it would be interesting to
751 further examine this issue using large-sample catchment-scale data sets from other parts of
752 the world.

753 # 5 DISCUSSION

754 [83] An understanding of how hydrological processes vary at large (e.g. national) scales is
755 important to the development of strategies for mitigating the effects of floods and droughts
756 (and other natural hazards). Such understanding can be difficult to establish, given the large
757 number of variables, attributes, and relationships that need to be considered. Under such
758 circumstances, the traditional approach of attempting to model the entire diversity of hydro-
759 geo-climatic conditions across an entire country/region with a single representational
760 approach may not result in a sufficiently accurate characterization of the underlying *Data*
761 *Generation Process (DGP)*. Through a case study, we have explored the possibility of using
762 a multi-representational approach to address the challenge of large-scale model development,
763 where the different representations are selected to have complementary strengths and with
764 the goal of maximizing learning and discovery.

765 ## 5.1 Challenges and Opportunities of a Multi-Representational Approach

766 [84] While each representation can support different kinds of discovery through the model
767 development and evaluation process, adoption of a multi-representational approach brings
768 forth both opportunities and challenges to be addressed.

769 1) It becomes difficult to implement a completely "*uniform*" strategy for model
770 development since each representational approach may exploit the information in
771 data differently and can have different requirements for inference.

772 2) For conceptual/process-resolved representations, discovery/learning about the spatial
773 variability of hydrological processes is mediated through an analysis of spatial
774 patterns or parameters.

775 [85] However, multiple parameter sets can give rise to similar model performance, thereby
776 complicating our ability to make meaningful inferences. In contrast, for the ML-based
777 models, the need for transformations and/or standardizations of the data was found to be
778 unnecessary, and even to bring about declines in overall performance telling us how different
779 representations are dealing with the data.

3) ML-based approaches facilitate an exploration of varying memory time scales.

[86] Our analysis suggested that, for energy-limited catchments in Chile, the ability to access input information over the past 32 days was critical to achieving an optimal representation, whereas for arid catchments the memory time-scale required was much longer (~270 days). In this regard, 32 days is likely associated with rapid time-scale precipitation-driven processes such as surface runoff and lateral flow, and while 270 days is likely associated with slower time-scale groundwater driven processes such as baseflow. While further investigation is needed to test these findings, such findings illustrate the power of ML-based approaches to support learning and discovery.

4) The RF architecture enables an exploration of feature importance, potentially enabling a higher degree of interpretability and discovery than the GR4J and LSTM representations.

[87] In our case, aridity was seen to provide the highest-level segregation of catchments, which makes sense given that the nature of the hydrological processes underlying the generation of streamflow depends, unavoidably, on the availability of water. Beyond this, various characteristic features associated with precipitation in the period just prior to the streamflow event of interest were seen to provide strong explanatory power. By exploring the structure of the decision tree model, it is possible to gain insight into the main relationships or drivers governing the behavior of the system under investigation.

[88] Overall, by synthesizing the results obtained using a multi-representational approach, we can obtain a more comprehensive overall picture of the underlying DGP, which in turn creates a better context for a more in-depth investigation of the capabilities and performance of each specific modeling approach.

5.2 On the Issue of Data Informativeness

[89] In terms of overall performance during the independent evaluation period (temporal generalization), the LSTM model provided better overall (statistical) performance than GR4J and RF. On the other hand, the GR4J model tended to be more robust, providing the best performance for locations where KGEss was lower than 0.15. It might make sense, therefore, to implement a lumped water balance model as a "*lower benchmark*" in any multi-representational ensemble of component models, and to generally require that any ML-based approach under consideration for inclusion in such an ensemble should demonstrate some benefits over the benchmark. Further, when an ML-based model fails to perform well when compared with the lumped water balance benchmark, this should alert us to the possibility that the data may not be sufficiently informative regarding the processes we seek to model.

[90] In this regard, note also that RF performed slightly better than LSTM when tested out-of-sample (spatial generalization). The reasons for this are not yet clear, but it is possible that the LSTM model development strategy employed tends to overfit the temporal/sequence patterns in the data. Regardless, this result also lends support to the idea that a multi-representational ensemble has the potential to be superior to one that is less representationally diverse.

[91] One common finding for all three models was their poor performance in one particular region of Chile. Given that the most important shared commonality of the three models is their access to the same dataset, coupled with the fact that ML-based approaches are highly flexible, this result strongly indicates that the dataset is not sufficiently informative to enable a suitable characterization of the streamflow response of this region, and that the main driver of local streamflow is not precipitation. Unfortunately, the Chilean CAMELS dataset does not include attributes from which it could be possible to infer groundwater-driven baseflow, or other related processes, and so proper characterization of the streamflow response of this region will require further investigation and exploration of alternative sources of relevant information.

[92] Overall, these observations point to the issue of whether the available data is informative enough for a sufficiently robust characterization of the underlying DGP to be achieved. While a multi-representation approach cannot (by itself) solve that issue, it can certainly help us to recognize the existence of the problem so that we can seek additional relevant information that may help us in the process of learning and discovery.

5.3 Relationship of the Multi-Representational Approach to Hypothesis Testing

[93] Given the tendency for each of the three representation types to provide better performance under different hydro-geo-climatic conditions, and the fact that each one facilitates different (complementary) kinds of information extraction and degree of interpretability, it seems clear that the three models can collectively be treated as a valuable tool for gaining insights regarding the underlying *DGP*. From this point of view, a meaningful answer to the question "*Does a single "correct" catchment-scale hydrological model exist at all?*" expressed by *Clark et al (2011)* may be that:

> "*It seems sensible to 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*".

This is, of course, because any model is unavoidably a "*simplified*" (and hopefully informationally relevant) representation of reality.

[94] Another way of think of this is that it is the "*ensemble of representations*" (and not each of the individual components thereof) that is actually "*the model*" per se, since it helps to meet the goal of incorporating within the "***Model***" (writ large) 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. This 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 (potentially) *informationally complementary* as possible, so that learning/discovery can be maximized. In contrast, an approach where the ensemble consists of hypotheses that may only be marginally different from each other (e.g, that all share the same system architecture while differing only in the forms of the process parameterization equations) may not lend itself to efficient and effective learning (*Gharari et al., 2021*).

[95] 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 a key to promoting model-based scientific discovery. While this perspective is likely to promote (as is currently happening) interest in hybrid approaches that integrate theory-based and data-based strengths, it is not clear that such a push towards reductionism through integration will necessarily obviate the need for a continued multi-representational approach in order for models to be tools that enable scientific discovery.

6 CONCLUSIONS

[96] In conclusion, while the metaphor of the "*blind people and the elephant*" is highly suggestive, it is not completely accurate. In the metaphor, each person constructs a different representation based on potentially different prior knowledge and clearly different sensory information (data). In our case, all of the representational approaches have access to the same sources of information (dataset) but differ in their abilities to fully exploit that information due to (prior) representational restrictions.

[97] So, while one might debate how to improve the metaphorical story to match the current situation, more important is the fact that an optimal strategy for scientific discovery would seem to be one that combines multiple complementary model structural representations (modeling strategies) with multiple complementary mechanisms for extracting information from data (inferential strategies). In this regard, it is perhaps worth noting that the strategy of "*multi-headed attention*" that has recently become the topic of intense inquiry in fields such as text prediction, translation, and speech recognition (*Vaswani et al, 2017; Devlin et al, 2018; Luo et al, 2021*), is explicitly based on the notion that multiple attentional perspectives bring considerable value to such tasks.

[98] This paper seeks to explicitly promote adoption of a multi-representational approach to learning, understanding, and discovery in the hydrological sciences. We believe that the multi-representational approach is fundamental to understanding hydrology at a large scale, where 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 dialogue and collaboration on this and related issues of how we use models to support prediction, understanding, and scientific discovery.

ACKNOWLEDGMENTS

[99] 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 Chilean Government scholarship "*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. The authors declare no conflicts of interest.

900 # CODE AND DATA AVAILABILITY

901 [100] The CAMELS-CL dataset is freely available from
902 https://doi.pangaea.de/10.1594/PANGAEA.894885. The analytical methods are presented
903 as a Jupiter notebook freely available at
904 http://www.hydroshare.org/resource/fc08997100fa4cd6abdd8a4f5731de15.

905 # APPENDIX

906 ## A1. Model Development Strategies

907 ### *A1.1 Partitioning the Data*

908 [101] A key step in model development is to partition the available data into '*model*
909 *development'* and '*evaluation'* subsets, where the former is used for model structure selection
910 and parameter tuning, while the latter is used to assess the generalization performance that
911 can be expected from the developed model. However, no clear guidance exists for how to
912 achieve such a partitioning for data that represent dynamical hydrological systems (*Wu et al,*
913 *2013, Daggupati et al, 2015, Zheng et al 2018, Guo et al 2020*). In general, the hydrological
914 literature has traditionally assumed that the entire available dataset comes from a stationary
915 underlying data generating process, and that any split that preserves the full range of
916 hydrologic variability (dry, medium, and wet) in both sets is satisfactory. Based on this
917 assumption, it is common to use a continuous-time period that makes up ~60-80% of the
918 available data for model development, while allocating the remaining ~20-40% for an
919 evaluation of the generalization ability of the model.

920 [102] In this study, we adopt the strategy of further partitioning the '*model development'*
921 subset into '*calibration'* and '*selection'* subsets, where the calibration subset is used for
922 model/network parameter tuning (commonly called '*training'* in the ML literature), and the
923 selection subset is used for model/network structure selection and/or hyperparameter tuning
924 (commonly called '*validation'* in the ML literature). Note that we adopt this naming
925 convention to try and overcome the inconsistency in terminology between the ML and
926 hydrological modeling literature. Accordingly, the available data are partitioned into three
927 subsets, where the first 60% of the data is used for model calibration, the next 24% is used
928 for model selection, and the final 16% is used for model evaluation (commonly called
929 '*testing'* in the ML literature)

930 ### *A1.2 Variable Selection*

931 [103] The variables selected from the CAMEL-CL dataset include two sources of
932 precipitation (CR2MET and MSWEP, both having long records), three values characterizing
933 temperature (Maximum, mean, and minimum), and potential evapotranspiration (PET)
934 estimated via the *Hargreaves and Samani (1985)* method. The PET value derived from
935 MODIS was not used because its time step is higher than daily (8 days). Further, the snow
936 water equivalent (SWE) data does not cover the entire country and was therefore not
937 considered suitable for the current study.

[104] Because the GR4J model has a pre-defined input representation, it is unable to use any other sources of data and so we used the weighted average of the two sources of precipitation as input to the GR4J model. In contrast, the ML-based models are able to use the information provided by all of the available variables and attributes, but in different ways. While the RF model used lagged input variables as surrogates for system memory (lag memory), the LSTM model used internal state variables to characterize system memory (sequential memory). More details regarding the variables and attributes used for the development of each model type are presented in Tables A-1, A-2, and A-3.

A1.3 Representing System Memory

[105] For the RF representation, which does not explicitly include dynamical state variables, system memory was included by concatenating past inputs (precipitation, evapotranspiration, and temperature) to the inputs for the current time step. This follows 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. For the RF, the number of past input lags was treated as a model hyperparameter. 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 huge system memory is required to manage the dataset as the number of lags is increased. We found that at 32 days of lagged memory, the computation became unstable. This prevented us from readily exploring longer memory time scales, such as 270 or 365 days (or longer), and the results presented only consider a memory time-scale of 16 days.

A1.4 Model Warm-Up

[106] 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 machine-learning approach, *Kratzert et al. (2019)* used 270 days, after testing 90, 180, 270, and 365 days as different options.

[107] In this study, we adopted the following strategy for warm-up period selection. For the GR4J and LSTM representations, because preliminary testing suggested that the LSTM requires the longest warm-up period to ensure stable results, 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.

A1.5 Parameters and Hyperparameters to be Tuned

[108] Each representational strategy involves different sets of parameters and hyperparameters, depending on its structural form. Whereas the original GR4J model contains 4 tunable parameters that must be calibrated for each catchment, our implementation includes an additional 3 parameters, two of which are used to facilitate driving the model by a weighted average of the two available precipitation products (CR2MET and MSWEP), and the third being the Box-Cox transformation parameter.

978 [109] For the LSTM model, in addition to a large number of system-wide network weights
979 and biases, 6 hyperparameters must be tuned, namely the sequence length (memory from the
980 past hidden states), number of hidden nodes, batch size, number of epochs, standardization
981 parameters, and the Box-Cox transformation parameter. To standardize the data (centering
982 by subtracting the mean, and rescaling by dividing by the standard deviation), we investigated
983 two options – global and local standardization. In global standardization, for each variable,
984 we use the mean and standard deviation computed from the entire dataset, whereas in local
985 standardization (which is applied only to the precipitation and streamflow variables) we
986 assume that the local means and standard deviations vary as functions of the aridity index.

987 [110] Finally, in addition to determining the nodal split "parameters", the RF model requires
988 the tuning of 4 hyperparameters for the entire set of catchments taken together, the first
989 representing system "memory" (expressed as the number of days previous to the current day
990 for which inputs are simultaneously presented to the model), the second being the Box-Cox
991 transformation parameter, the third being the number of trees, and the fourth being the
992 minimum number of elements that must be retained in the last leaf. To attempt to circumvent
993 the problem of being unable to input lagged daily inputs beyond 32 days to the RF model to
994 account for "memory" in the system, we augment the input data to include surrogate variables
995 intended to be informative about the state of the system. Specifically, we included the
996 "month-of-year" as an attribute, to enable the model to learn a representation of long-term
997 memory as the average behavior associated with different months of the year. Meanwhile, as
998 mentioned above, the short-term memory was treated as a hyperparameter.

999 ***A1.6 Model Calibration/Training***

1000 [111] To calibrate the parameters of the GR4J model to each catchment in the calibration
1001 period, we tested both the *Root Mean Square Error* (RMSE) and the *Kling-Gupta Efficiency*
1002 (KGE, *Gupta et al., 2009*), as defined below. Overall, we found that KGE provided slightly
1003 more robust results (*De la Fuente, 2021*), and therefore we present here only the results
1004 obtained using KGE in this paper.

1005 $$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

1006 $$KGE = 1 - \sqrt{(r-1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}$$

1007 y_i: $Measured\ streamflow$
1008 $\hat{y}_i$: $Simulated\ streamflow$
1009 n: $Total\ number\ of\ data$
1010 r: $Linear\ correlation\ coefficient\ between\ y_i\ and\ \hat{y}_i$
1011 α: σ_s/σ_o: $relative\ variability\ between\ simulated\ and\ observed\ data.$
1012 β: μ_s/μ_o: $ratio\ between\ simulated\ and\ observed\ data.$

1013 [112] For parameter optimization, we used three algorithms from the *Spotpy Python* library
1014 (*Houska et al., 2015*), namely *Maximum Likelihood Estimation* (MLE), *Differential*
1015 *Evolution Adaptive Metropolis* (DE-MCz), and *Shuffled Complex Evolution* (SCE-UA). In

1016 total, 22 independent optimization runs were done for each catchment, and the parameter set
1017 that provided the best performance (out of the 22 parameter sets so obtained) on the '*selection*
1018 (*hyperparameter tuning)*' data subset was chosen.

1019 [113] To develop the RF model, we use the *Scikit-learn Python library* (*Pedregosa et al.,*
1020 *2011*). The *RandomForestRegressor* module (version 0.23.1) has two options for
1021 performance metrics – *Mean Squared Error* (MSE) and *Mean Absolute Error* (MAE). While
1022 MAE can be used to reduce the tendency to emphasize larger streamflow values, because we
1023 are implementing the Box-Cox transformation on streamflow we chose MSE to be the metric
1024 used for RF calibration.

1025 $$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

1026 $$MAE = \frac{\sum_{i=1}^{n} abs(y_i - \hat{y}_i)}{n}$$

1027 [114] To train the LSTM model, we used the implementation provided by *Kratzert et al*
1028 *(2019)* and modified it to conform to the data structures and variables of the CAMELS-CL
1029 dataset. Whereas the original code enables the choice of either MSE or NSE as the calibration
1030 metric, we used only NSE because its normalization of the error enables better comparison
1031 across catchments having different amounts of temporal variability.

1032 $$NSE = 1 - \frac{MSE}{\sigma_o^2}$$

1033 ***A1.7 Model Performance Evaluation***

1034 [115] For performance evaluation, we use the KGE skill score (KGEss) (*Knoben et al., 2019*)
1035 computed on the evaluation-period data. KGEss is a rescaled version of the KGE metric such
1036 that a value of zero corresponds to the prediction being no better than simply using the mean
1037 observed streamflow, in a manner analogous to NSE. While other metrics, including NSE
1038 and RMSE, were also used for model evaluation (*De la Fuente, 2021*), we do not report them
1039 here as the conclusions are similar to those obtained using KGEss. Importantly, we account
1040 for sampling variability by computing the estimated posterior distributions of KGEss by
1041 bootstrapping 100 times (*Efron and Tibshirani, 1994*) and using the median value of KGEss
1042 in all comparisons.

1043 $$KGEss = \frac{KGE - KGE_{benchmark}}{1 - KGE_{benchmark}} = \frac{KGE + \sqrt{2} - 1}{\sqrt{2}} = 1 - \frac{1 - \mathrm{KGE}}{\sqrt{2}}$$

1044 ***A1.8 Out-of-Sample Testing***

1045 [116] For an additional '*out-of-sample*' model evaluation step, we retained all of the
1046 CAMELS-CL catchments for which less than 7 years but more than 1 year of data is
1047 available; being less than 7 years of record length, none of these catchments are included in
1048 the model development data set. The resulting 167 catchments facilitate a meaningful out-

of-sample operational comparison of the generalization abilities of the LSTM and RF ML-based representations. 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.

[117] Note that, because the additional '*out-of-sample*' model evaluation data set is completely independent of the data set used for model development, while being similarly representative of the geo-hydro-climatic variability across the country (***Figure A-1***), the model performed using those data can be considered to be similar to the idea of "*Proxy-basin differential split-sample testing*" (*Klemeš, 1986)*.

REFERENCES

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

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

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., ... & Ayala, A. (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

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

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 (Methodological), 26*(2), 211-243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

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., ... & Rasmussen, R. M. (2015). A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research, 51*(4), 2498-2514. https://doi.org/10.1002/2015WR017198

Craig, J. R., Brown, G., Chlumsky, R., Jenkinson, R. W., Jost, G., Lee, K., ... & Tolson, B. A. (2020). Flexible watershed simulation with the Raven hydrological modelling framework. *Environmental Modelling & Software, 129*, 104728. https://doi.org/10.1016/j.envsoft.2020.104728

Daggupati, P., Pai, N., Ale, S., Douglas-Mankin, K. R., Zeckoski, R. W., Jeong, J., ... & Youssef, M. A. (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

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint https://arxiv.org/abs/1810.04805

De la Fuente, L. (2021*). Using Big-Data to Develop Catchment-Scale Hydrological Models for Chile* (Master dissertation, The University of Arizona). https://repository.arizona.edu/handle/10150/656824

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

Fenicia, F., Kavetski, D., & Savenije, H. H. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research, 47*(11). https://doi.org/10.1029/2010WR010174

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*, https://doi.org/10.1029/2020WR027948

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, 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

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

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

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

Hassan, M., & Hassan, I. (2021). Improving Artificial Neural Network Based Streamflow Forecasting Models through Data Preprocessing. *KSCE Journal of Civil Engineering*, 1-13. https://doi.org/10.1007/s12205-021-1859-y

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). 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

Klemeš, V. (1986). Operational testing of hydrological simulation models. *Hydrological sciences journal, 31*(1), 13-24. https://doi.org/10.1080/02626668609491024

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. *arXiv preprint arXiv:1907.08456*. 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

Le Moine, N. (2008). *Le bassin versant de surface vu par le souterrain: une voie d'amélioration des performances et du réalisme des modèles pluie-débit?* (Doctoral dissertation, Doctorat Géosciences et Ressources Naturelles, Université Pierre et Marie Curie Paris VI).

Luo, H., Zhang, S., Lei, M., & Xie, L. (2021, January). Simplified self-attention for transformer-based end-to-end speech recognition. In *2021 IEEE Spoken Language*

Technology Workshop (SLT) (pp. 75-81). IEEE. https://doi.org/10.1109/SLT48900.2021.9383581

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

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). 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*; 2010. 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. *The 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

Pushpalatha, R., Perrin, C., Le Moine, N., Mathevet, T., & Andréassian, V. (2011). A downward structural sensitivity analysis of hydrological models to improve low-flow simulation. *Journal of hydrology*, 411(1-2), 66-76. https://doi.org/10.1016/j.jhydrol.2011.09.034

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, July). 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, No. 1, p. 012037). IOP Publishing. https://doi.org/10.1088/1755-315/299/1/012037

Valéry, A. (2010). *Modélisation précipitations débit sous influence nivale: Elaboration d'un module neige et évaluation sur 380 bassins versants* (Doctoral dissertation, Doctorat Hydrobiologie, Institut des Sciences et Industries du Vivant et de l'Environnement AgroParisTech).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008). https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

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

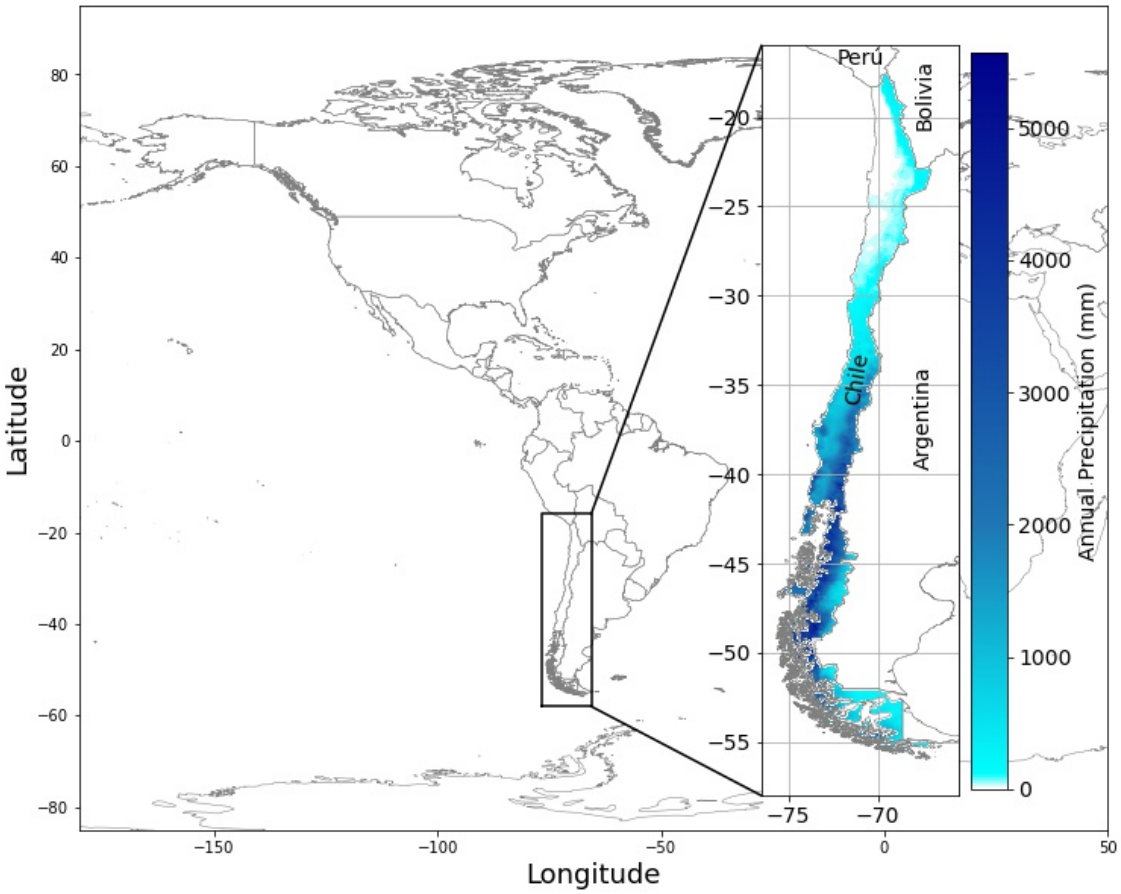

Figure 1. Map showing the geographic location of Chile, and its spatial distribution of annual precipitation.

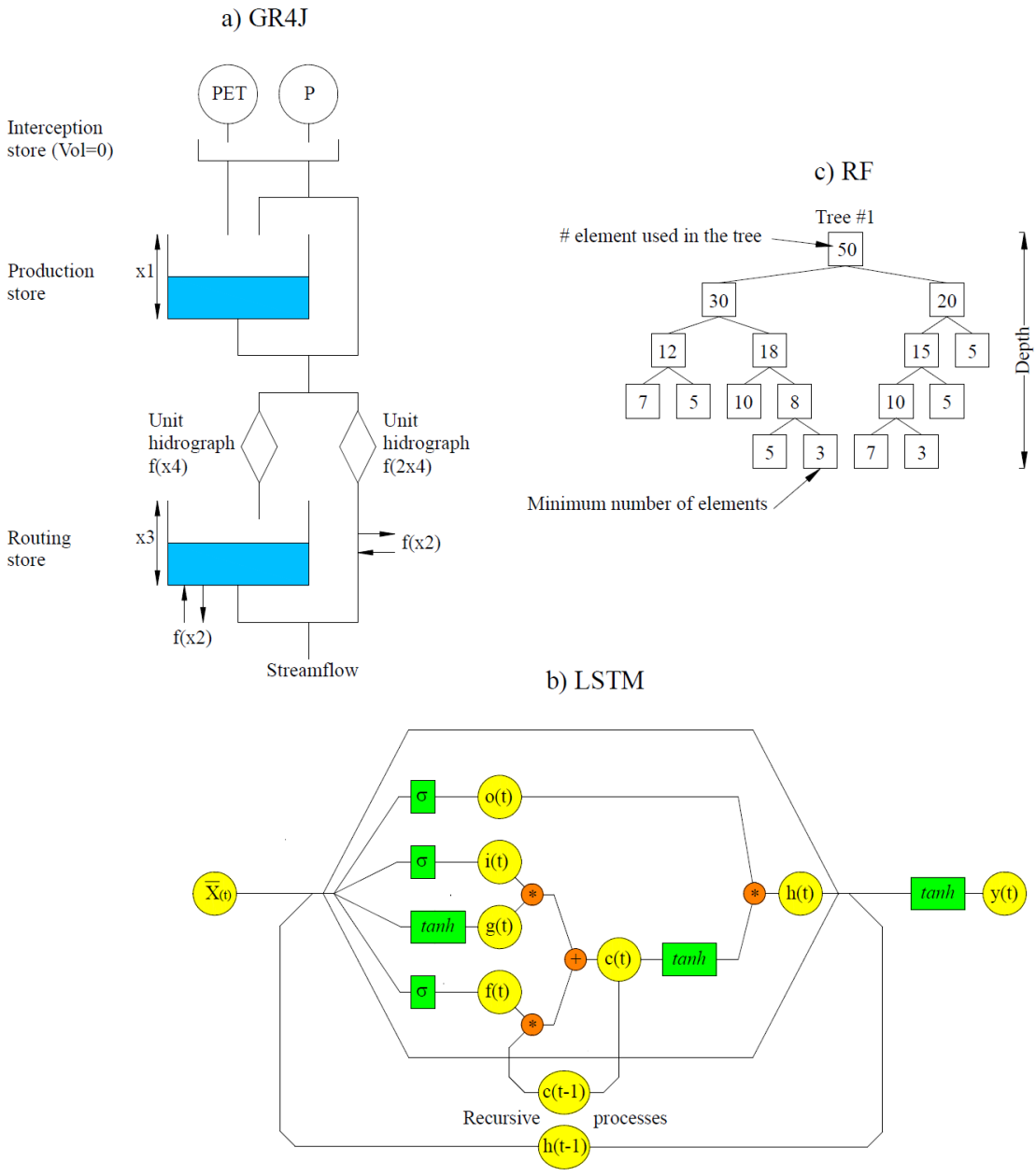

Figure 2. Representational structures of the three different models used; a) The GR4J lumped-water balance model; b) The *Long-Short Term Memory* (LSTM) machine learning model; and c) The *Random Forest* (RF) machine learning model.

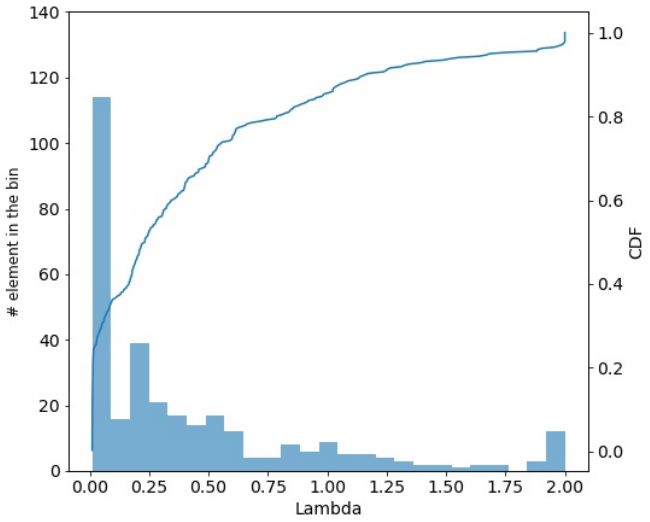

Figure 3. Frequency distribution of the λ hyperparameter for the 322 catchments when using the GR4J model.

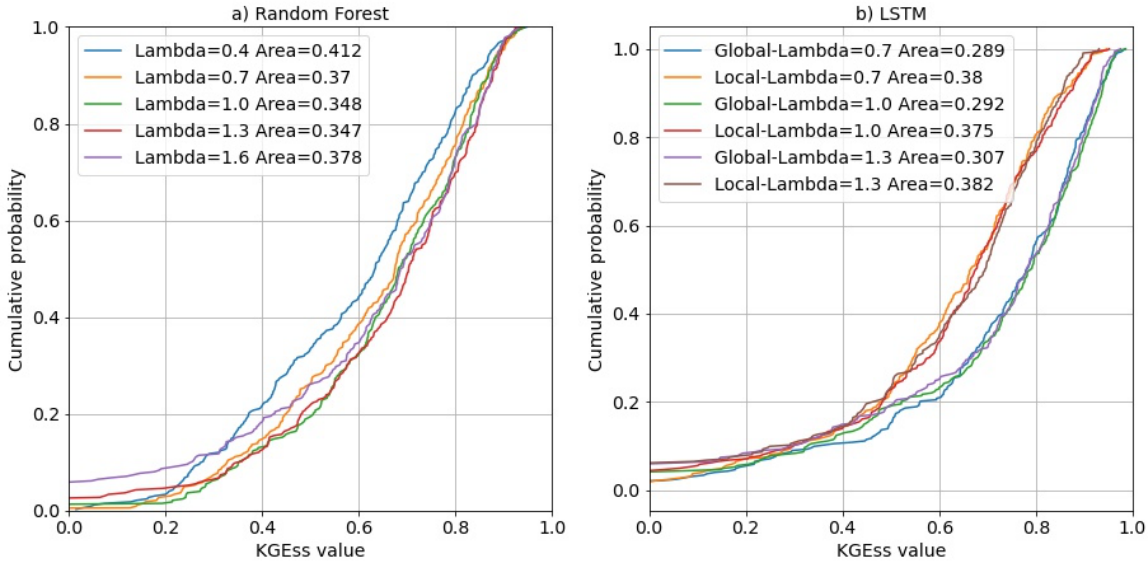

Figure 4. Selection period performance CDF's obtained using different values of the λ hyperparameter; the Left subplot is for RF and the right subplot is for LSTM.

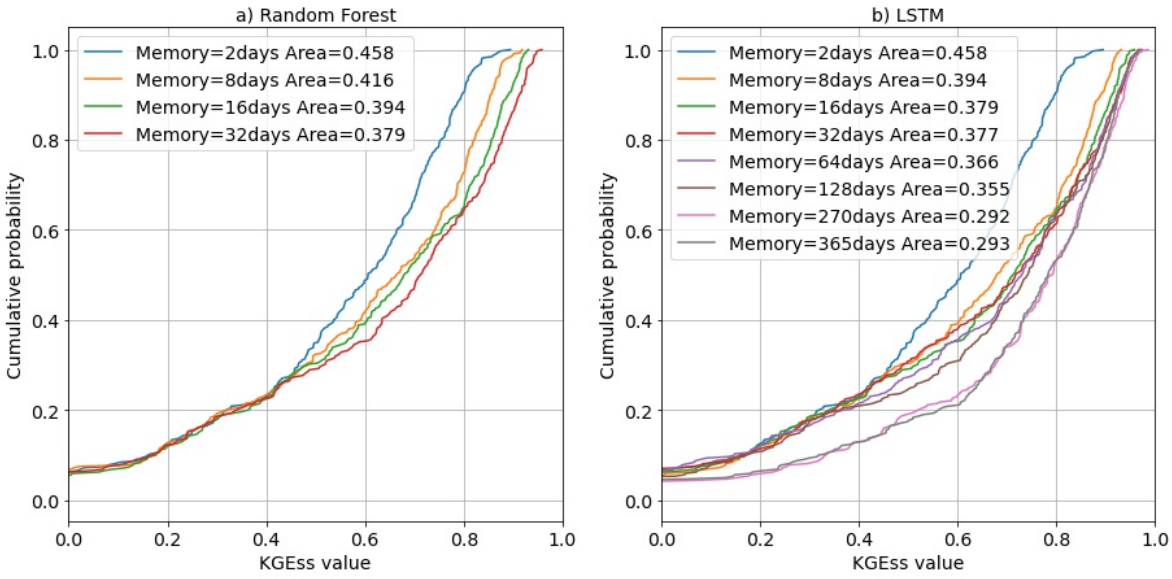

Figure 5. Selection period performance CDF's for the RF and LSTM models, showing dependence on memory lag.

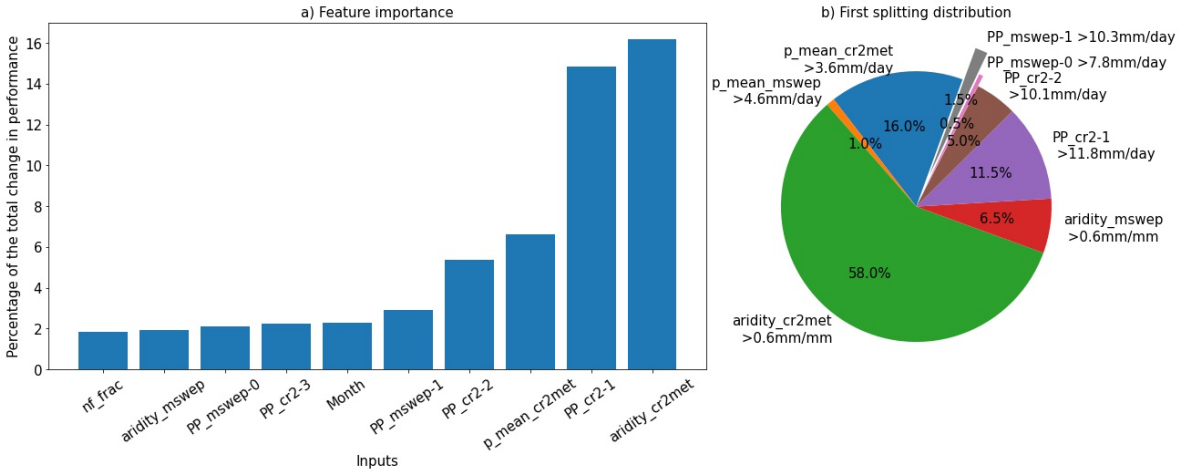

Figure 6. Feature importance and distribution of the first split of the RF model (For a description of the name of each attribute or variable, see Table A-3).

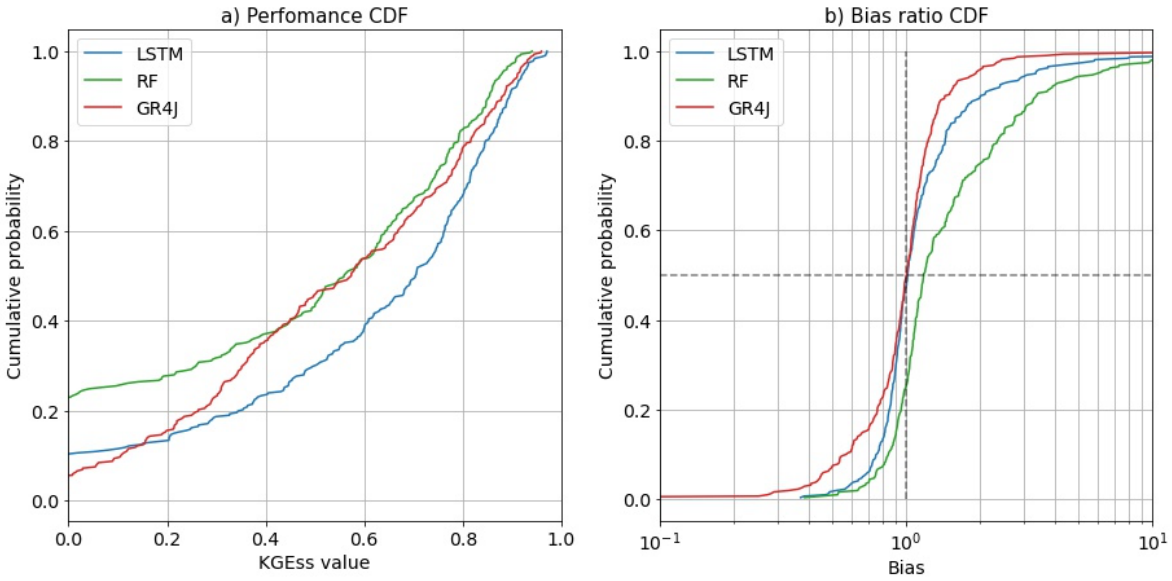

Figure 7. Evaluation period performance CDFs for the three models. The left subplot shows KGEss and the right subplot shows Bias Ratio.

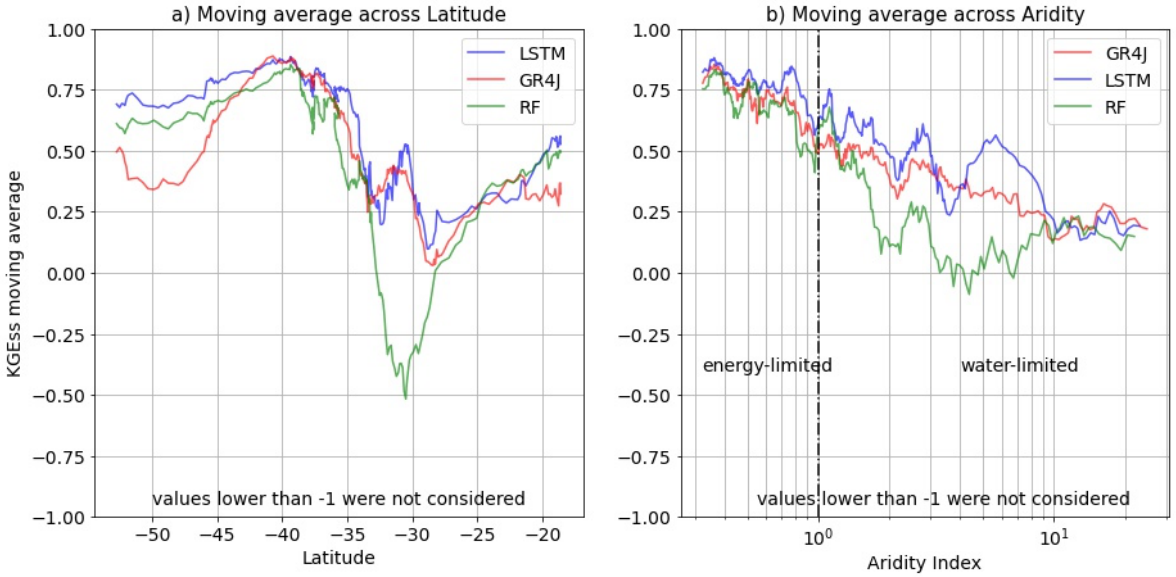

Figure 8. Variation of evaluation period KGEss performance with aridity and latitude.

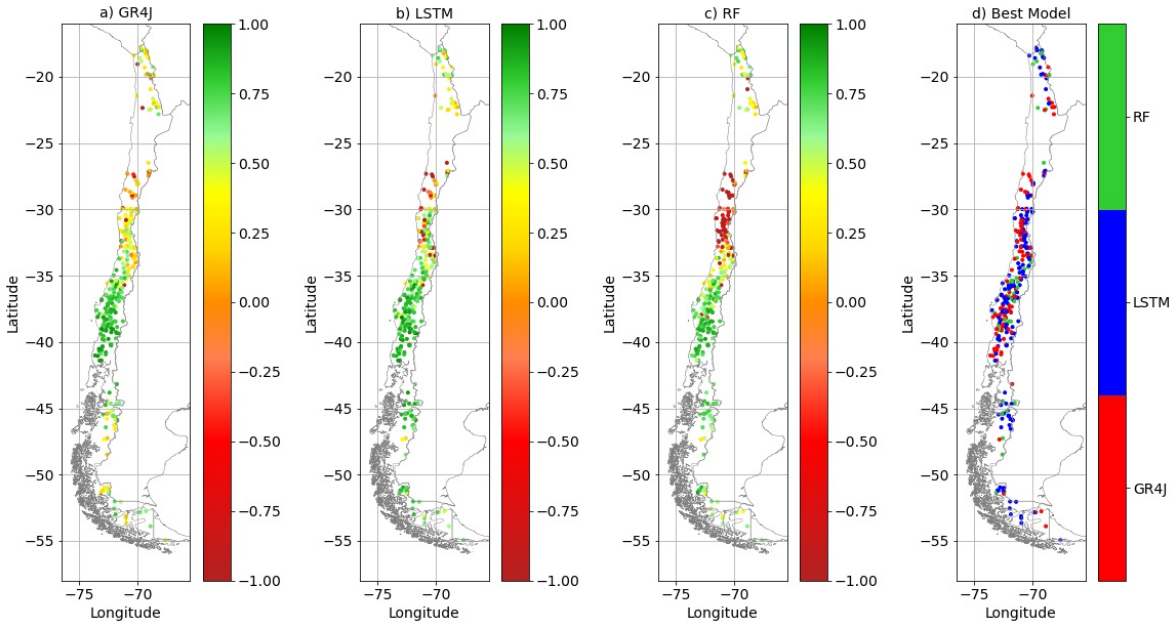

Figure 9. Spatial distributions of evaluation period model performance for a) GR4J, b) LSTM, c) RF, and d) the "best" performing model.

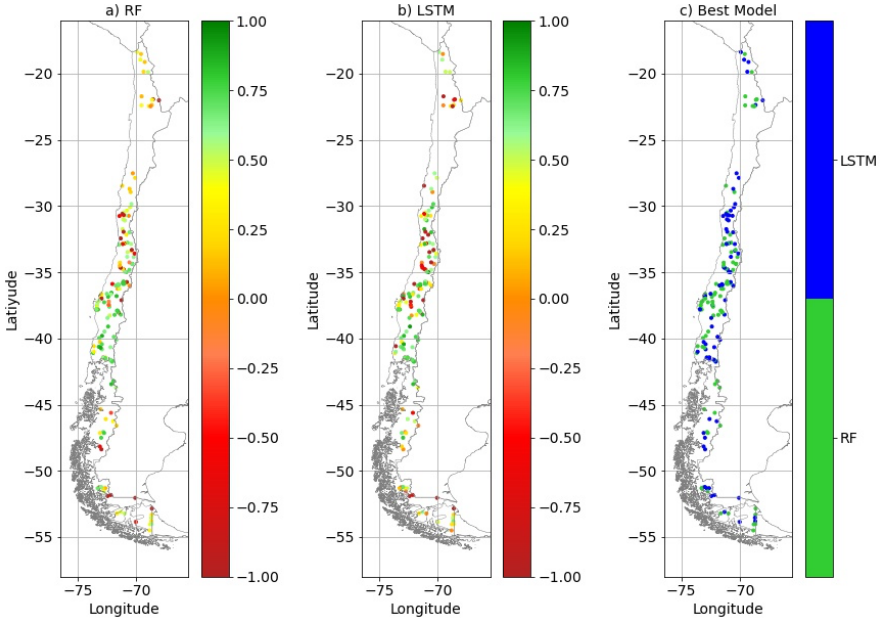

Figure 10. Spatial distributions "out-of-sample" performance for a) RF and b) LSTM. The right subplot (c) indicates the "best" performing model.

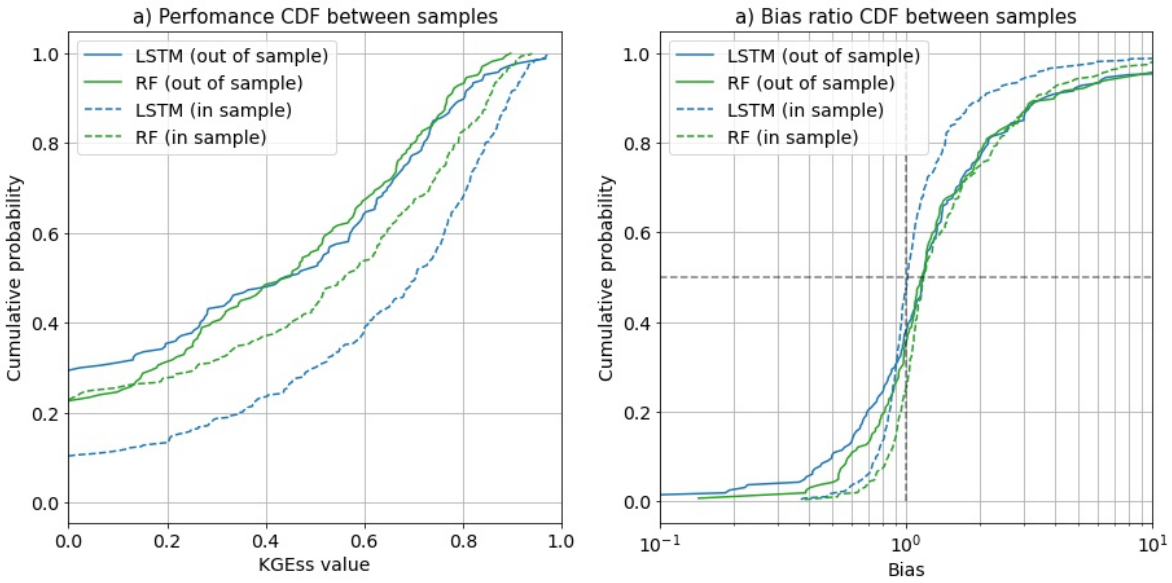

Figure 11. Performance CDFs for temporal (in-sample) and spatial (out-of-sample) generalization.

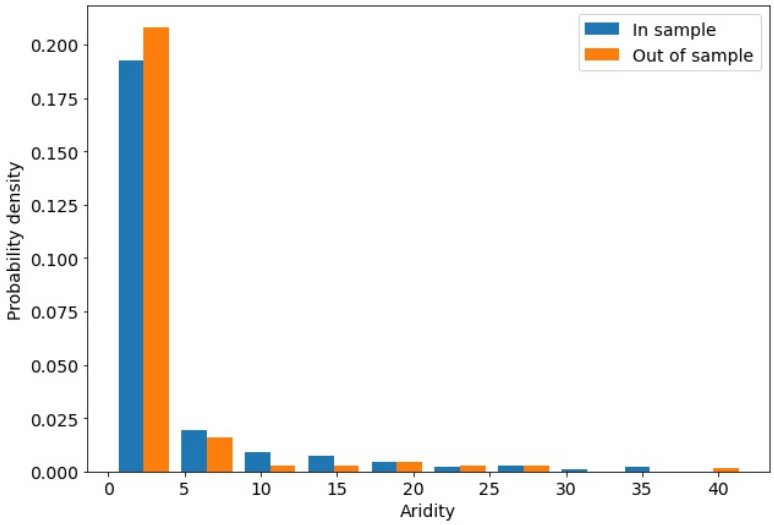

Figure A-1: Comparison of the histogram for both samples used in the performance analysis.

Table 1. Comparison between Linear Reservoir and LSTM

Linear Reservoir	**LSTM**
$\frac{dS}{dt} = I - O$	$\frac{dc}{dt} = g(x, h)$
$S: water\ storage$	$c: information\ storage$
$I: input$	$x: input$
$O: output$	$h: output$
$S(t) = \alpha \cdot S(t-1) + \beta \cdot (I - O)$	$c(t) = \alpha \cdot c(t-1) + \beta \cdot g(x, h)$
$\alpha = 1$	$\alpha = f(t)$
$\beta = 1$	$\beta = i(t)$
$O = k \cdot S(t)$	$h = k \cdot \bar{\bar{c}}(t)$
$k =]0,1[$	$k = o(t)$
	$\bar{\bar{c}}(t) = c(t)\ normalized\ between\ ]-1,1[$

Table 2. Evaluation period performance statistics.

Model	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**	**# Positive**	**# Best**
GR4J	**0.417**	**0.960**	**-11.621**	0.311	0.561	0.789	0.960	**303**	98
RF	-2.411	39.310	-703.128	0.075	0.563	0.762	0.940	249	52
LSTM	-3.282	65.269	-1170.490	**0.442**	**0.704**	**0.826**	**0.971**	289	**172**

Table 3. Out-of-sample performance statistics.

Model	**Mean**	**Std**	**Min**	**25%**	**50%**	**75%**	**Max**	**# Positive**	**# Better**
RF	**-0.356**	**2.605**	**-17.501**	**0.118**	**0.45**	0.666	0.897	**130**	**89**
LSTM	-2.474	17.341	-203.124	-0.103	0.429	**0.678**	**0.968**	116	78

Table A-4 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	0-2.0
x1	Storage production capacity	0-5000
x2	Amplification of water exports	-10 to 10
x3	Storage routing capacity	0-1500
x4	Time-delay between the initial and maximum values of the hydrograph	0.5-4.5
Lambda	Exponent of Box-Cox transformation	0-2.0

Table A-5 Variables used in the GR4J model.

Variable	Description
PP_cr2-0	Precipitation in the same day (“0”) of the mean streamflow from CR2MET product
PP_mswep-0	Precipitation in the same day (“0”) of the mean streamflow from MSWEP product
ETP-0	Potential Evapotranspiration in the same day (“0”) of the mean streamflow
Q	Mean streamflow

Table A-6 Variables used in the Random Forest model.

n°	Attribute or variable	n°	Attribute or variable	n°	Attribute or variable	n°	Attribute or variable	n°	Attribute or variable
1	area	31	fp_frac	61	PP_cr2-3	91	PP_mswep-16	121	Tmean-6
2	aridity_cr2met	32	frac_snow_cr2met	62	PP_cr2-4	92	Q	122	Tmean-7
3	aridity_mswep	33	frac_snow_mswep	63	PP_cr2-5	93	shrub_frac	123	Tmean-8
4	big_dam	34	gauge_lat	64	PP_cr2-6	94	slope_mean	124	Tmean-9
5	carb_rocks_frac	35	gauge_lon	65	PP_cr2-7	95	snow_frac	125	Tmean-10
6	crop_frac	36	grass_frac	66	PP_cr2-8	96	sur_rights_flow	126	Tmean-11
7	Day	37	gw_rights_flow	67	PP_cr2-9	97	sur_rights_n	127	Tmean-12
8	elev_gauge	38	gw_rights_n	68	PP_cr2-10	98	Tmax-0	128	Tmean-13
9	elev_max	39	high_prec_dur_cr2met	69	PP_cr2-11	99	Tmax-1	129	Tmean-14
10	elev_mean	40	high_prec_dur_mswep	70	PP_cr2-12	100	Tmax-2	130	Tmean-15
11	elev_med	41	high_prec_freq_cr2met	71	PP_cr2-13	101	Tmax-3	131	Tmean-16
12	elev_min	42	high_prec_freq_mswep	72	PP_cr2-14	102	Tmax-4	132	Tmin-0
13	ETP-0	43	imp_frac	73	PP_cr2-15	103	Tmax-5	133	Tmin-1
14	ETP-1	44	lc_barren	74	PP_cr2-16	104	Tmax-6	134	Tmin-2
15	ETP-2	45	lc_glacier	75	PP_mswep-0	105	Tmax-7	135	Tmin-3
16	ETP-3	46	low_prec_dur_cr2met	76	PP_mswep-1	106	Tmax-8	136	Tmin-4
17	ETP-4	47	low_prec_dur_mswep	77	PP_mswep-2	107	Tmax-9	137	Tmin-5
18	ETP-5	48	low_prec_freq_cr2met	78	PP_mswep-3	108	Tmax-10	138	Tmin-6
19	ETP-6	49	low_prec_freq_mswep	79	PP_mswep-4	109	Tmax-11	139	Tmin-7
20	ETP-7	50	Month	80	PP_mswep-5	110	Tmax-12	140	Tmin-8
21	ETP-8	51	nf_frac	81	PP_mswep-6	111	Tmax-13	141	Tmin-9
22	ETP-9	52	p_mean_cr2met	82	PP_mswep-7	112	Tmax-14	142	Tmin-10
23	ETP-10	53	p_mean_mswep	83	PP_mswep-8	113	Tmax-15	143	Tmin-11
24	ETP-11	54	p_mean_spread	84	PP_mswep-9	114	Tmax-16	144	Tmin-12
25	ETP-12	55	p_seasonality_cr2met	85	PP_mswep-10	115	Tmean-0	145	Tmin-13
26	ETP-13	56	p_seasonality_mswep	86	PP_mswep-11	116	Tmean-1	146	Tmin-14
27	ETP-14	57	pet_mean	87	PP_mswep-12	117	Tmean-2	147	Tmin-15
28	ETP-15	58	PP_cr2-0	88	PP_mswep-13	118	Tmean-3	148	Tmin-16
29	ETP-16	59	PP_cr2-1	89	PP_mswep-14	119	Tmean-4	149	wet_frac
30	forest_frac	60	PP_cr2-2	90	PP_mswep-15	120	Tmean-5		

Table A-4 Variables used in the LSTM model.

n°	Attribute or variable	n°	Attribute or variable
1	area	31	p_mean_cr2met
2	aridity_cr2met	32	p_mean_mswep
3	aridity_mswep	33	p_mean_spread
4	big_dam	34	p_seasonality_cr2met
5	carb_rocks_frac	35	p_seasonality_mswep
6	crop_frac	36	pet_mean
7	elev_gauge	37	shrub_frac
8	elev_max	38	slope_mean
9	elev_mean	39	snow_frac
10	elev_med	40	sur_rights_flow
11	elev_min	41	sur_rights_n
12	forest_frac	42	wet_frac
13	fp_frac	43	PP_cr2-0
14	frac_snow_cr2met	44	PP_mswep-0
15	frac_snow_mswep	45	Tmin-0
16	grass_frac	46	Tmean-0
17	gw_rights_flow	47	Tmax-0
18	gw_rights_n	48	ETP-0
19	high_prec_dur_cr2met		
20	high_prec_dur_mswep		
21	high_prec_freq_cr2met		
22	high_prec_freq_mswep		
23	imp_frac		
24	lc_barren		
25	lc_glacier		
26	low_prec_dur_cr2met		
27	low_prec_dur_mswep		
28	low_prec_freq_cr2met		
29	low_prec_freq_mswep		
30	nf_frac		