Data Assimilation Informed model Structure Improvement (DAISI) for robust prediction under climate change: Application to 201 catchments in southeastern Australia

Julien Lerat[1], Francis Hock Soon Chiew[1], David Ewen Robertson[1], Vazken Andréassian[2], and Hongxing Zheng[1]

[1]Commonwealth Scientific and Industrial Research Organisation (CSIRO)
[2]Irstea (formerly Cemagref)

November 14, 2023

Abstract

This paper presents a method to analyze and improve the set of equations constituting a rainfall-runoff model structure based on a combination of a data assimilation algorithm and polynomial updates to the state equations. The method, which we have called "Data Assimilation Informed model Structure Improvement" (DAISI) is generic, modular, and demonstrated with an application to the GR2M model and 201 catchments in South-East Australia. Our results show that the updated model generated with DAISI generally performed better for all metrics considered included KGE, NSE on log transform flow and flow duration curve bias. In addition, the modelled elasticity of runoff to rainfall is higher in the updated model, which suggests that the structural changes could have a significant impact on climate change simulations. Finally, the DAISI diagnostic identified a reduced number of update configurations in the GR2M structure with distinct regional patterns in three sub-regions of the modelling domain (Western Victoria, central region, and Northern New South Wales). These configurations correspond to specific polynomials of the state variables that could be used to improve equations in a revised model. Several potential improvements of DAISI are proposed including the use of additional observed variables such as actual evapotranspiration to better constrain the model internal fluxes.

Hosted file

`977967_0_art_file_11536405_s3hdyk.docx` available at https://authorea.com/users/613569/articles/685421-data-assimilation-informed-model-structure-improvement-daisi-for-robust-prediction-under-climate-change-application-to-201-catchments-in-southeastern-australia

Hosted file

`977967_0_supp_11536411_s3d8d6.docx` available at https://authorea.com/users/613569/articles/685421-data-assimilation-informed-model-structure-improvement-daisi-for-robust-prediction-under-climate-change-application-to-201-catchments-in-southeastern-australia

Data Assimilation Informed model Structure Improvement (DAISI) for robust prediction under climate change: Application to 201 catchments in southeastern Australia

Julien Lerat(1), Francis Chiew(1), David Robertson(2), Vazken Andréassian(3), Hongxing Zheng(1)

(1) CSIRO Environment, Canberra, ACT, Australia
(2) CSIRO Environment, Clayton, VIC, Australia
(3) INRAE, Antony, France

Abstract

This paper presents a method to analyze and improve the set of equations constituting a rainfall-runoff model structure based on a combination of a data assimilation algorithm and polynomial updates to the state equations. The method, which we have called "Data Assimilation Informed model Structure Improvement" (DAISI) is generic, modular, and demonstrated with an application to the GR2M model and 201 catchments in South-East Australia. Our results show that the updated model generated with DAISI generally performed better for all metrics considered included KGE, NSE on log transform flow and flow duration curve bias. In addition, the elasticity of modelled runoff to rainfall is higher in the updated model, which suggests that the structural changes could have a significant impact on climate change simulations. Finally, the DAISI diagnostic identified a reduced number of update configurations in the GR2M structure with distinct regional patterns in three sub-regions of the modelling domain (Western Victoria, central region, and Northern New South Wales). These configurations correspond to specific polynomials of the state variables that could be used to improve equations in a revised model. Several potential improvements of DAISI are proposed including the use of additional observed variables such as actual evapotranspiration to better constrain the model internal fluxes.

Key words [6 max]: Model structure, Model diagnostic, data assimilation, Ensemble Smoother, Climate change scenario

Key points

1. DAISI method diagnoses hydrological model structures by combining data assimilation with a polynomial update of state equations.
2. The method was applied to the GR2M rainfall-runoff model with significantly improved streamflow simulations in 201 Australian catchments.

3. The method identified updates to state equations with marked regional characteristics that could guide future improvement of GR2M.

Plain language summary

This paper presents a data-driven method to improve rainfall-runoff models used to generate future water resources scenario in climate change studies. The method, which we have called “Data Assimilation Informed model Structure Improvement” (DAISI) is generic, modular, and demonstrated with an application to monthly streamflow simulations over a large dataset of catchments in South-East Australia. Our results show that DAISI improves model performance for a wide range of metrics and increases the sensitivity of the model to climate inputs, which is critical in climate change scenarios. Finally, the improvements identified by DAISI take a simple mathematical form with distinct regional patterns in three sub-regions of the study domain (Western Victoria, central region, and Northern New South Wales). Several improvements of DAISI are discussed including the inclusion of additional observed variables such as evapotranspiration to better constrain model simulations.

Contents

Notations

N	Number of rainfall-runoff model state equations.
P	Number of observed variables.
R	Number of data assimilation ensembles.
B	Number of catchments in the study area.
T	Number of time steps.
V	Number of components in state vector.
O	Number of output variables in the state vector.
V_n	Number of variables affecting the n^{th} state variable.
$\tilde{x}_t$	State vector at time t.
$\tilde{u}_t$	Input vector at time t.
$\tilde{m}_t$	Model output vector at time t.
$\tilde{d}_t$	Observed data vector at time t.
f	Model dynamic equation.
L_n	Number of coefficients in the n^{th} update equation.
X^f	Forecast state matrix of dimension $T(\sum_n V_n) \times R$.
X^a	Analysis state matrix of dimension $T(\sum_n V_n) \times R$.
$\tilde{x}_t[r]$	r^{th} ensemble of the state vector at time step t in the assimilated ensemble.
$\tilde{y}_t$	Normalized state vector at time t.
$\delta_{n,t}$	Update term for the n^{th} normalized state variable at time step t.
$\Delta_{n,t}$	Assimilated update term for the n^{th} state variable at time step t.
KGE	KGE performance metric
F_B	Flow duration curve bias performance metric

ϵ_P Elasticity of modelled streamflow to rainfall evaluation metric

C_n Matrix of dimension $2B \times L_n$ containing the update coefficients for the n^{th} state variable, all catchments in the dataset and two calibration periods.

$s_{n,k}$ k^{th} singular value of C_n

1. Introduction

The pressure on water resources is reaching unprecedented levels in many catchments around the world due to increasing anthropogenic presence and higher variability induced by climate change. In this tense context, catchment scale rainfall-runoff models are one of the main quantitative tools used by water managers to translate future climate predictions into water volumes and assess water sharing scenarios. Estimation of future streamflows like in the study by Chiew, Vaze et al. (2008) is generally done by selecting a few rainfall-runoff models to generate streamflow projections based on future climate inputs. Unfortunately, the performance of these models degrades significantly when predicting values beyond the range of hydro-climate conditions seen during their calibration (Coron, Andréassian et al. 2012). Of particular worry is the tendency for rainfall-runoff models to over-estimate streamflow in dry years which are expected to become more common in the future in many regions, for example in South Eastern Australia (Chiew, Young et al. 2011). This paper presents a method to analyze and improve the equations constituting a rainfall-runoff model structure in the context of climate change scenario modelling demonstrated with an application to the GR2M model (Mouelhi, Michel et al. 2006) and a large dataset of catchments in South-East Australia.

Most rainfall-runoff models are empirical and hence require their parameters to be calibrated based on observed data. Once input and output data of acceptable quality are obtained, improving model calibration is the first step to obtain defensible simulations of future streamflow. Calibration algorithms are the topic of a considerable literature including the development of stochastic (see the review by Arsenault, Poulin et al. 2014), probabilistic (Kuczera and Parent 1998, Beven and Freer 2001, Vrugt and Ter Braak 2011) or multi-objective (see the review by Efstratiadis and Koutsoyiannis 2010) algorithms. These advances have allowed for highly parameterized models to be routinely calibrated within operational systems. However, there are limits to what a better calibration strategy can achieve to simulate streamflow in a changing climate. Coron, Andréassian et al. (2014) showed that models are often incapable of simulating significant changes in rainfall-runoff relationships regardless of how they are calibrated. Zheng, Chiew et al. (2022) go further by saying that “calibration can only marginally (if at all) improve the quantification of uncertainty in future runoff projection due to hydrological nonstationarity”. These studies suggest that improvement in model structures is critical to obtain more robust streamflow projections. Pursuing this idea, Fowler, Knoben et al. (2020) identified that most model structures are not able to simulate multi-year processes that are driving changes in rainfall-runoff relationship during drought periods.

Unfortunately, formulating an efficient rainfall-runoff model structure is not straightforward because of the difficulty to describe physical processes at the catchment scale (Beven 2001) leading to a certain level of subjectivity in the process. To overcome these limitations, one can assemble a large collection of published models and compare their performance as was done by Perrin, Michel et al. (2001) and more recently by Knoben, Freer et al. (2020). These studies have laid the foundations for the development of robust model structures such as the GR4J model (Perrin, Michel et al. 2003). However, they also concluded that no single structure outperforms the others systematically and that the difference in performance between structures is not well explained by catchment descriptors. As a result, it is difficult to define a clear path leading to model structure improvement from these approaches. To overcome these limitations, flexible software frameworks such as FUSE (Clark, Slater et al. 2008) or SUPERFLEX (Fenicia, Kavetski et al. 2011) have been proposed to create arbitrary model structures from selected components and hence allow the comparison of a much larger set of candidate structures. These tools remain complex to implement and few authors have applied them beyond pure research applications in a single catchment. A notable exception is the study by Van Esse, Perrin et al. (2013) who applied a large number of model structures to 237 catchments in France. Van Esse, Perrin et al. (2013) concluded on the difficulty to relate model structures with catchment characteristics. The study named several modelling components that proved generally beneficial (e.g., parallel routing stores, bypass flows), but did not offer a simple diagnostic to improve a particular model such as the ones used in climate change studies.

As an alternative to the previous approaches, data itself can guide the identification of model structures. Machine learning method such as deep learning offers powerful tools to generate purely data-driven model structures (Nearing, Kratzert et al. 2021). However, as suggested by Wi and Steinschneider (2022), pure machine learning models may lack the capacity to extrapolate far beyond historical conditions such as required in climate change studies. In addition, machine learning models remain complex compared to empirical lumped rainfall-runoff models which does not facilitate their use in an operational context. Consequently, this paper focuses on classical modelling approaches based on empirical equations derived from physical system knowledge.

In this context, Lamb and Beven (1997) and subsequently Kirchner (2009) used data analysis to infer the form of model equations. Their approach remains limited to specific hydrological processes (recession for Lamb and Beven 1997) or catchment characteristics (dominant base flow contribution for Kirchner 2009). This concept was expanded further by Gharari, Gupta et al. (2021) in theoretical experiments who explored the uncertainty in model structure via randomly generated piecewise linear functions. Overall, these attempts of data-driven model structure identification are promising but lack practical and large-scale applications to improve climate projections in the short term. In contrast, the field of data assimilation has produced firmly established algorithms such as the Ensemble Kalman Filter (Evensen 2009) to efficiently blend model simulations with observed data over large spatial domains. These algorithms have been used in hydrology for several decades as reviewed by Ghorbanidehno, Kokkinaki et al. (2020). For example, Pathiraja, Marshall et al. (2016) used data assimilation to estimate time-varying parameters in synthetic case studies, which is a powerful approach to remediate model structure deficiencies. However, according to Beck (1985), large time variations of parameters could also be interpreted as a structural deficiency requiring remediation. To our knowledge, only Bulygina and Gupta (2009) have demonstrated the use of a data assimilation algorithm for the identification of a complete set of rainfall-runoff model equations and applied their model beyond synthetic experiments to observed data. The approach of Bulygina and Gupta (2009) relies on an iterative algorithm where the particle filter (Doucet, Godsill et al. 2000) is used within each iteration to generate probabilistic model equations sampled from a mixture of multivariate normal distributions. This approach is elegant because it allows combining a prior estimate of the model structure with observed data in a fully Bayesian inference scheme. However, it is significantly more complex than classical data assimilation algorithms because it requires a repeated application of the particle filter (hence ensuring that the filter does not degenerate as warned by Moradkhani, DeChant et al. 2012) and a customized sampling scheme described by Bulygina and Gupta (2011). The method is promising but was applied to a single catchment in the United States with results qualified as "preliminary" by Bulygina and Gupta (2011). Consequently, it does not seem applicable to a large catchment dataset in an operational context.

The previous review of the literature shows important research gaps related to the improvement of rainfall-runoff model structures:

- **Lack of methods to improve model structure beyond trial and error of pre-defined structures**: the most advanced methods currently available to identify model structures are based on trial and error of pre-defined model structures either collected from published literature or built from selected components. These approaches often lead to model equifinality where multiple structures are seen as equally applicable, which provides limited guidance for the improvement of a specific model.
- **Estimation of variable structure remains theoretical**: methods have been developed to infer variable model structures directly from observed data, but they remain essentially theoretical with limited or no application to real catchments. Data assimilation offers promising avenues in this field with well-established algorithms to blend models and data.
- **Limited research on model structure improvement in a climate change context**: Most of the research done to improve model simulations in a climate change context has focused on model parameterization and calibration. Methods to perform model structure diagnostic in this context are emerging but lack quantitative approaches to accelerate progress.

Consequently, the three objectives of this paper are:

- **Define a new approach** where data assimilation is used to identify structural improvements of an existing rainfall-runoff model structure. The method should be robust and computationally tractable enough to be applicable to a large dataset of catchments offering insight into regional trends in model structure updates. The method focuses exclusively on model structure improvement and not on model parameterization by assuming that the model has been calibrated prior to the structural diagnostic. The rationale behind this choice is to avoid duplicating existing research on model calibration and leave open the choice of the calibration process.
- **Demonstrate that the improvement can benefit model simulations** using a wide range of metrics and provide a **detailed diagnostic** on the structural improvement to guide future model development.
- **Present an example of the overall process using the GR2M monthly rainfall-runoff model**, an existing data assimilation scheme (Ensemble Smoother) and a large data set of catchments in a region experiencing a pronounced climate change signal (South-East Australia).

The proposed method is presented in Section 2 including its objective and principles. Section 3 describes the empirical case study with the description of the GR2M model, the evaluation process, and the catchment dataset. Application of the method is presented in details for one example catchment in section 4.1 and then generalized to 201 catchments in section 4.2 and 4.3. The strength and weaknesses of the method, the knowledge gained on the GR2M structure and potential future development of the method are discussed in section 5. Section 6 concludes the paper.

2. Theory

2.1. Objective and Principles of Data Assimilation Informed model Structure Improvement (DAISI)

The main goal of DAISI is to provide a rapid diagnostic of an existing rainfall-runoff model structure by analyzing time series of state variables generated by a data assimilation algorithm. DAISI relies on Bayesian inference but aims at providing simple diagnostics that can be used outside of a probabilistic framework. The method can be applied to a single catchment or to a large dataset of catchments to obtain a more robust diagnostic on the model structure.

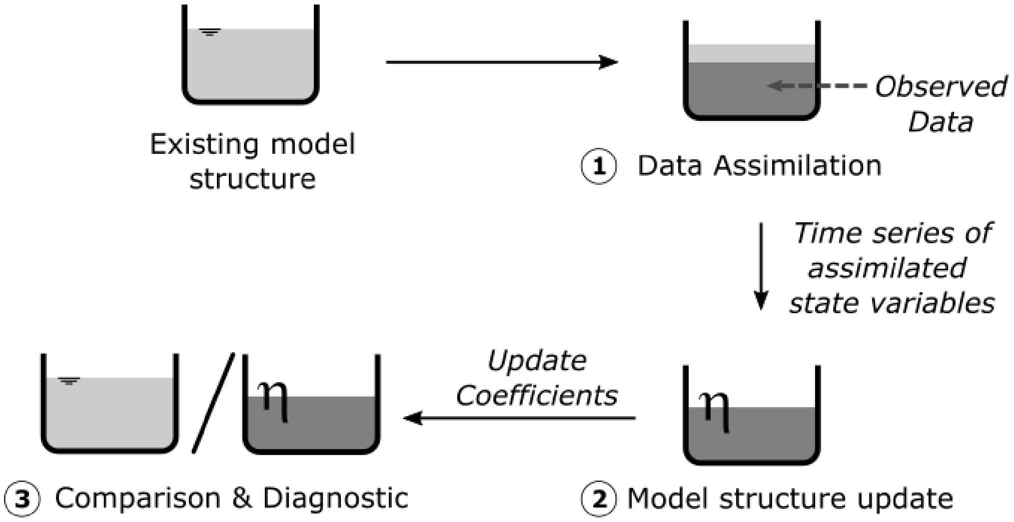

Figure 1: The three steps of the DAISI method

DAISI is based on three steps outlined in Figure 1. The method starts by assimilating observations during a calibration period (e.g., observed streamflow data) resulting in an ensemble of state variables. In a second step, the assimilated states are used to update the model structure. Finally, the new model structure is run over an independent validation period similarly to a classical rainfall-runoff model (i.e., without the use of assimilation or structure update) and compared with the original structure in terms of model behavior and performance.

2.2. Step 1: Data Assimilation

A rainfall-runoff model is a numerical solution to a set of ordinary differential equations describing the water storages and fluxes at a catchment scale. When integrated over a time step, these equations take the following form referred to as “state equations” in the rest of the paper:

$$\tilde{x}_{t+1} = f\left(\tilde{u}_t, \tilde{x}_t, \tilde{\theta}\right) \qquad \text{Eq. 1}$$

Where t is the time step, $\tilde{u}_t$ is the input vector, $\tilde{x}_t$ the state vector of length V, f is a vector value function characterizing its dynamic, and $\tilde{\theta}$ is a parameter vector assumed to be obtained from a prior calibration exercise. The state vector includes all variables that affect the dynamic of the model such as internal stores, fluxes and model outputs denoted as $\tilde{m}_t$ (e.g., streamflow). Observed data corresponding to $\tilde{m}_t$ are denoted $\tilde{d}_t$. The concatenation of all $\tilde{x}_t$ vectors for $t = 1, \dots, T$ is denoted $\tilde{x}$. Similar notations are used for vectors $\tilde{u}$ and $\tilde{d}$.

The first step of DAISI is a smoothing data assimilation algorithm that aims at estimating the probability distribution of states $\tilde{x}$ given f, inputs $\tilde{u}$ and observed data $\tilde{d}$ over the whole calibration period ($t = 1\!:\!T$). Among the wide range of methods described in the literature (Van Delft, El Serafy et al. 2009, Moradkhani, DeChant et al. 2012), the linear ensemble smoother (ES) introduced by van Leeuwen and Evensen (1996) is one of the simplest algorithms where model errors are assumed to be linearly related to observations and the prior distribution of errors is assumed Gaussian. Despite its limitations in handling non-linear dynamics, the high computational efficiency of ES, especially its non-sequential nature, is appealing in a diagnostic tool such as DAISI. Note that the use of smoothing algorithms remains limited in hydrology, mostly due to their high computing requirements (Li, Ryu et al. 2014). As this is a well-known algorithm, the presentation of ES is deferred to Appendix A. Our implementation of ES relies on a single tuning factor α_e which relates the covariance of the perturbations applied to the state and input variables to their covariance in the original model simulation (see Appendix A, Eq. 24). It is fixed to a value of 0.1 (i.e., covariance perturbation equal to α_e^2=1% of the original covariance) for all instances of ES. The impact of this factor on the performance of DAISI was found to be small as shown in Supplementary Material S3. Consequently, the fixed value of 0.1 was adopted throughout this paper.

The outcome of ES is a set of R ensemble vectors $\{\tilde{x}[r], r = 1, \dots, R\}$ denoted as "analyzed states" or X^a (see Appendix A) containing samples from the posterior distribution $p(\tilde{x}|\tilde{u}, \tilde{d}, f, \tilde{\theta})$. It is highlighted that the choice of ES does not prevent the use of more sophisticated smoothing algorithms in DAISI. This point is discussed in Section 5.

Data assimilation schemes are notoriously complex to configure with parameters that are difficult to relate to actual observations. Consequently, it is important to verify that the assimilated ensemble is statistically consistent with observed data (reliable). This consistency was measured with the normalized RMSE ratio (Moradkhani, Sorooshian et al. 2005, Fortin, Abaza et al. 2014, Thiboult and Anctil 2015):

$$N_R[k] = \frac{\sqrt{\frac{1}{T}\sum_t \left(\frac{1}{R}\sum_r m_t[k,r] - d_t[k]\right)^2}}{\frac{1}{R}\left\{\sum_r \sqrt{\frac{1}{T}\sum_t (m_t[k,r] - d_t[k])^2}\right\}} \sqrt{\frac{2R}{R+1}} \qquad \text{Eq. 2}$$

Where $m_t[k,r]$ is the r^{th} ensemble of the k^{th} model output corresponding to observation $d_t[k]$. A value of N_R close to one indicates statistical reliability while N_R substantially smaller or greater than 1 suggests a too wide or narrow ensemble, respectively.

2.3. Step 2: Model Structure Update

In the second step of DAISI, the analyzed states X^a are used to estimate updates in the state equation (Eq. 1). A preliminary transformation of the state equations, referred to as "normalization" in the remainder of the paper, is undertaken if DAISI is applied to a large number of catchments. This normalization aims at removing site specific parameters from the state equations to allow the comparison of structural updates between catchments (see examples in Section 3.1). It is acknowledged that such normalization is not always possible, which is an important limitation of DAISI. Let us assume that the normalization of the state and input vectors is given by

$$\tilde{y}_t = \psi_x(\tilde{x}_t, \tilde{\theta}), \qquad \tilde{v}_t = \psi_u(\tilde{u}_t, \tilde{\theta}) \qquad \text{Eq. 3}$$

Where ψ_x and ψ_u are normalization functions for states and inputs, respectively. Using this normalization, the state equation becomes:

$$\tilde{y}_{t+1} = f^*(\tilde{y}_t) \qquad \text{Eq. 4}$$

Where f^* is the normalized function with no dependency on rainfall-runoff model parameters $\tilde{\theta}$ as opposed to f. Note that if DAISI is applied to a single catchment, the normalization process is not required and $\tilde{y}_t$ and f^* can be replaced by $\tilde{x}_t$ and f, respectively.

The fundamental concept in DAISI is to alter the state variables with an update term as follows:

$$\hat{y}_{n,t+1} = y_{n,t+1} + \delta_{n,t} \qquad \text{Eq. 5}$$

Where $y_{n,t}$ is the n^{th} component of $\tilde{y}_t$, $\hat{y}_{n,t}$ is the updated state and $\delta_{n,t}$ is the update term. Let us assume that a subset of V_n variables, noted $\{y_{i,t}\}_{i=1,..,V_n}$, affects $y_{n,t}$ among the full set of V state variables. The update term is computed as a quadratic form of $y_{i,t}$ written as:

$$\delta_{n,t} = \eta_n[0] + \sum_{i=1}^{V_n} \eta_n[i]\ y_{i,t} + \sum_{i=1}^{V_n} \eta_n[V_n + i]\ y_{i,t}^2 + \sum_{1 \le i < j \le V_n} \eta_n[k(i,j)]\ y_{i,t}\ y_{j,t} \qquad \text{Eq. 6}$$

Where is $\eta_n[i]$ the i^{th} update coefficient and $k(i,j) = 2V_n + i + (j-1)(j-2)/2$ is the index for cross-product terms. The coefficient vector $\tilde{\eta}_n$ is of length L_n where

$$L_n = 1 + 2V_n + V_n(V_n - 1)/2 \qquad \text{Eq. 7}$$

Eq. 5 is the fundamental equations of DAISI and is referred to as the “update equation” in the rest of the paper. The form of the update term in Eq. 6 was chosen because it is a non-linear function of the state variables but a linear function of the coefficients which greatly facilitates their estimation as discussed below. Eq. 5 provides a simple way to explore alternative model structures continuously (as opposed to pre-defined or discrete structures) by varying the coefficients $\tilde{\eta}_n$. Note that Eq. 6 has a similar form to the second order Taylor series expansion of the n^{th} component of f^* at the origin. Consequently, the update coefficients $\tilde{\eta}_n$ can be interpreted as modifications of its partial derivatives up to order 2 close to the origin. Despite these attractive properties, Eq. 6 does not impose any physical constraint on the update term, which can lead to non-physical values of the updated state $\hat{y}_{n,t+1}$. This is an important limitation of the method and the price to pay for the flexibility offered by Eq. 6. When running the updated model structure, the lack of physical constraints requires that checks on their bounds be placed on the updated state variables to ensure their physical realism. An example of such checks is provided in Appendix C.

If the normalized states are known, for example via data assimilation as presented in the previous section, it is possible to compute what is referred to as the "assimilated update", denoted $\Delta_{n,t}[r]$, for each ensemble member r:

$$\Delta_{n,t}[r] = y_{n,t+1}[r] - f_n^*(\tilde{y}_t[r]) \quad \text{Eq. 8}$$

Where $\tilde{y}_t[r]$ is the assimilated normalized state vector from the r^{th}ensemble member and f_n^* is the n^{th} component of f^*. The assimilated update $\Delta_{n,t}[r]$ can be subsequently combined with Eq. 5 in a regression equation:

$$\Delta_{n,t}[r] = \delta_{n,t}[r] + \epsilon_{n,t}[r] \quad \text{Eq. 9}$$

Where $\epsilon_{n,t}[r]$ is a residual assumed to follow a normal distribution with mean 0 and standard deviation σ_n. Eq. 9 is an ordinary multivariate regression that can be solved easily by Bayesian inference if non-informative priors are assumed (Gelman, Carlin et al. 2013). Consequently, Eq. 9 provides a way to estimate update coefficients $\tilde{\eta}_n$ for each ensemble member.

Generalizing the approach described above, DAISI aims at estimating the distribution of $\tilde{\eta}_n$ for each state equation given the model structure f, model parameters $\tilde{\theta}$, and input ($\tilde{u}$) and observed data ($\tilde{d}$) over a calibration period $t = 1{:}T$. This probability is noted $P(\tilde{\eta}_n|\tilde{u}, \tilde{d}, f, \tilde{\theta})$. Using the posterior distribution of state variables $\tilde{x}$ estimated by data assimilation presented in the Section 2.2, the distribution of $\tilde{\eta}_n$ can be obtained by introducing $\tilde{x}$ and integrating as follows:

$$P(\tilde{\eta}_n|\tilde{u}, \tilde{d}, f, \tilde{\theta}) = \int_{\tilde{x}} P(\tilde{\eta}_n|\tilde{x}, \tilde{u}, \tilde{d}, f, \tilde{\theta})\, P(\tilde{x}|\tilde{u}, \tilde{d}, f, \tilde{\theta})\, d\tilde{x} \quad \text{Eq. 10}$$

Introducing the assimilated ensemble, Eq. 10 can be approximated as

$$P(\tilde{\eta}_n|\tilde{u}, \tilde{d}, f, \tilde{\theta}) \approx \frac{1}{R}\sum_r P(\tilde{\eta}_n|\tilde{x}[r], \tilde{u}[r], \tilde{d}[r], f, \tilde{\theta}) \quad \text{Eq. 11}$$

$$\approx \frac{1}{R}\sum_{r} P\left(\tilde{\eta}_n \middle| \tilde{y}[r], \tilde{u}[r], \tilde{d}[r], f^*, \tilde{\theta}\right) \quad \text{Eq. 12}$$

This paper aims at producing a deterministic run of the modified model which requires a single estimate of $\tilde{\eta}_n$. The choice made here is to compute this estimate as the expected value of $P\left(\tilde{\eta}_n \middle| \tilde{u}, \tilde{d}, f, \tilde{\theta}\right)$, denoted $\tilde{\eta}_n^a$ and computed as follows:

$$\tilde{\eta}_n^a = \int_{\tilde{\eta}_n} \tilde{\eta}_n\, P\left(\tilde{\eta}_n \middle| \tilde{u}, \tilde{d}, f, \tilde{\theta}\right) d\tilde{\eta}_n \approx \frac{1}{R}\sum_{r} \int_{\tilde{\eta}_n} \tilde{\eta}_n\, P\left(\tilde{\eta}_n \middle| \tilde{y}[r], \tilde{u}[r], \tilde{d}[r], f^*, \tilde{\theta}\right) d\tilde{\eta}_n \quad \text{Eq. 13}$$

If a noninformative prior on $\tilde{\eta}_n$ is assumed, the integral on the right-hand side of Eq. 13 is the posterior mean of the coefficients in a multivariate regression which is equal to the ordinary least square solution (Box and Tiao 2011):

$$\int_{\tilde{\eta}_n} \tilde{\eta}_n\, P\left(\tilde{\eta}_n \middle| \tilde{y}[r], \tilde{u}[r], \tilde{d}[r], f^*, \tilde{\theta}\right) d\tilde{\eta}_n = \left(Y[r]^T Y[r]\right)^{-1} \tilde{y}[r]^T \Delta_{t,n}^{(r)} \quad \text{Eq. 14}$$

Where $Y[r]$ is the predictor matrix associated with assimilated ensemble r in which the columns are the L_n predictor variables in the right-hand side of Eq. 6.

In summary, the second step of DAISI aims at modifying the N state equations, and hence the model structure, using a multivariate polynomial regression parameterized by coefficients $\tilde{\eta}_n$. Expected values of these coefficients, denoted $\tilde{\eta}_n^a$, can be estimated to obtain a single set of update coefficients.

2.4. Step 3: Model Diagnostics

Once the expected coefficients $\tilde{\eta}_n^a$ are obtained from Step 2, the model can be run using the updated state equation (Eq. 5), leading to modified simulated variables. It is highlighted that data assimilation and coefficient fitting are not used at this stage of DAISI and that the updated model runs exactly like a classical rainfall-runoff model. The last step of DAISI compares the original and updated model by answering three questions: (1) Is the updated model a robust alternative to the original model? (2) What is driving the updates? (3) Are there dominant functional forms of the update?

To answer the first question, the simulations produced by both structures are compared over a validation period using evaluation metrics. Four metrics were selected starting from the KGE performance metric (Gupta, Kling et al. 2009) which summarize model performance by aggregating measures of bias in the mean, bias in variance and correlation into a single metric. KGE alone is not sufficient to assess model performance, especially on low flows (Pushpalatha, Perrin et al. 2012). To assess low-flow performance, we used the Nash-Stucliffe efficiency computed on log-transform flow with an offset of 1 mm/month to handle zero values. Furthermore, following the recommendations of

Refsgaard, Madsen et al. (2014) in the evaluation of climate change scenario, we included the flow duration curve bias index (Lerat, Thyer et al. 2020):

$$F_B(\tilde{q}^o, \tilde{q}^s) = 1 - \frac{1}{100}\sum_{k=1}^{100}\left|1 - \frac{Pct(\tilde{q}^s, k)}{Pct(\tilde{q}^o, k)}\right| \quad \text{Eq. 15}$$

Where $Pct(\tilde{q}, k)$ is the k^{th} percentile of streamflow time series $\tilde{q}$, and $\tilde{q}^o$ and $\tilde{q}^s$ are the observed and simulated streamflow series, respectively. F_B is equal to 1 for a perfect simulation. The fourth metric is the relative elasticity of modelled streamflow to rainfall computed as:

$$\epsilon_P = \frac{E[\tilde{p}]}{E[\tilde{q}^s(\tilde{p})]}\frac{E[\tilde{q}^s(\tilde{p}^+)] - E[\tilde{q}^s(\tilde{p}^-)]}{E[\tilde{p}^+] - E[\tilde{p}^-]} \quad \text{Eq. 16}$$

Where $\tilde{p}^+$ and $\tilde{p}^-$ are two rainfall scenarios in which historical rainfall series $\tilde{p}$ are scaled up and down by +10% and -10%, respectively. $\tilde{q}^s(\tilde{p})$ is the streamflow simulation obtained when forcing the model with rainfall scenario $\tilde{p}$ and $E[\tilde{p}]$ is the mean value of $\tilde{p}$. The choice of 10% as a scaling factor was guided by the range of rainfall variability expected in South-East Australia (Charles, Chiew et al. 2020). ϵ_P is distinct from the three previous metrics because it does not compare the model with an observed reference. Comparing ϵ_P between the original and updated model quantifies the impact of the DAISI structural update on climate change scenarios. This last test is important because better performance, as measured by the three previous metrics, does not guarantee that the updated model will yield significantly different climate change scenario when forced with different climatological inputs (e.g., reduced rainfall scenario).

Additional metrics including absolute bias, NSE, NSE on reciprocal flow (Pushpalatha, Perrin et al. 2012), the recent PMR robustness (Royer-Gaspard, Andreássian et al. 2021) and split KGE (Fowler, Peel et al. 2018) metrics are included in the Supplementary Material S2. These metrics lead to similar conclusions than the three described in the previous paragraphs.

The second element of the DAISI diagnostic explores the trends in the update term $\delta_{n,t}$. To visualize how state variables affect the update term, a scatterplot is generated by plotting $\delta_{n,t}$ on the vertical axis versus the percentile rank of one of the state variables on the horizontal axis. The percentile rank is used to allow the plotting of data from multiple sites in a single plot and hence analyze regional trends. The choice of the state variable on the horizontal axis is subjective and depends on the model and state equation. All variables were trialed and the one leading to the easiest plot to interpret was retained. When multiple sites are plotted simultaneously, the update terms are binned based on the state variable. The median, 25% and 75% quantiles of the update term are computed for each bin and added to the plot to ease interpretation.

The third element of the DAISI diagnostic aims to find dominant patterns in the functional form of the updates. Let us assume that DAISI was applied to B sites and 2 calibration periods. A matrix C_n of size $2B \times L_n$ (L_n is the number of update coefficients in Eq. 6) is constructed for each state variable n by concatenating as rows all the update coefficient vectors $\tilde{\eta}_n^a$ for each site and calibration period. The influence of outliers in this matrix is tempered by clipping the values between -1 and 1. These bounds are subjective and may vary depending on the model. Dominant patterns are identified in C_n through a reduced singular value decomposition (Lawson and Hanson 1974):

$$C_n = A_n\, S_n\, B_n^T \qquad \text{Eq. 17}$$

Where A_n and B_n are orthogonal matrices of size $2B \times L_n$ and $L_n \times L_n$, respectively, and S_n is a diagonal matrix of size $L_n \times L_n$ containing the singular values $s_{n,1} \dots s_{n,L_n}$ along its diagonal in decreasing order by convention. The columns of B_n are referred to as singular vectors. Eq. 17 provides important insights into the functional form of the update. First, the components of the singular vectors are directly related to the predictor variables in the update equation (see Eq. 6). Consequently, each singular vector corresponds to a set of coefficients and hence to a specific update polynomial. Second, assume that the weights $\omega_{n,k}$ are defined from the singular values as:

$$\omega_{n,k} = \frac{s_{n,i}^2}{\sum_{i=1}^{L_n} s_{n,i}^2} \qquad \text{Eq. 18}$$

$\omega_{n,k}$ varies between $1/L_n$ and 1 and represents the total distance between the rows of C_n and their projection on the k^{th} singular vector as per the inner-product (Hastie, Tibshirani et al. 2009). For example, a value of $\omega_{n,1}$ close to 1 indicates that the rows of C_n are nearly colinear with the first singular vector, suggesting that the update polynomial has a form similar to the first singular vector for all sites and periods. Finally, the product $s_{n,k} \times A_n[:,k]$, referred to as principal component k, contains the projection of each row of C_n on the k^{th} singular vector. These projections can be used to find groups of sites where the update coefficients are colinear to the k^{th} singular vector, and hence where the update is close to the corresponding polynomial. The uncertainty in the decomposition was assessed by replicating Eq. 17 whilst bootstrapping the rows of C_n to obtain confidence intervals on the singular vectors.

3. Empirical Case Study Methods

3.1. Rainfall-runoff Model

The DAISI method is applied to the GR2M monthly rainfall-runoff model (Makhlouf and Michel 1994, Mouelhi, Michel et al. 2006) presented in detail in Appendix B. The model runs a sequence of two stores. The first one referred to as the "production" store (S) receives rainfall (P) and potential evapotranspiration (E). It generates effective rainfall P_e which is then transferred to the "routing" store of fixed capacity $\theta_r = 60$mm which in turn produces streamflow (Q). The model has two calibrated parameters: the capacity of the production store θ_1 (mm) and the inter-basin exchange coefficient θ_2 (-) which controls the amount of water gained or lost from the surface water catchment (Mouelhi, Michel et al. 2006). The GR2M model has four state variables listed in Table 1 with more details provided in Appendix B. In this table, variables corresponding to the end of the time step are marked with a "+".

Table 1: GR2M state variables

State variable	**Variables affecting the state variable**	**Normalization functions**		**Number of update coefficients**
Production store (S^+)	S, P, E	$y_s = \frac{S}{\theta_1}$ $y_p = \frac{P}{\theta_1}$ $y_e = \frac{E}{\theta_1}$	$y_{s^+} = \frac{S^+}{\theta_1}$	10
Effective rainfall (P_e)	S, P, E	$y_s = \frac{S}{\theta_1}$ $y_p = \frac{P}{\theta_1}$ $y_e = \frac{E}{\theta_1}$	$y_{p_e} = \frac{P_e}{\theta_1}$	10
Routing store (R^+)	R, P_e	$y_r = \theta_2 \frac{R}{\theta_r}$ $y_{p_e^*} = \theta_2 \frac{P_e}{\theta_r}$	$y_{r^+} = \frac{R^+}{\theta_r}$	6
Streamflow (Q)	R, P_e	$y_r = \theta_2 \frac{R}{\theta_r}$ $y_{p_e^*} = \theta_2 \frac{P_e}{\theta_r}$	$y_q = \frac{Q}{\theta_r}$	6

Note that the notation is modified in the rest of the paper to improve readability by referring to specific GR2M state variables using the names indicated in Table 1 as a lower-case subscript instead of the state variable number n used in the previous sections (for example $y_{s^+,t}$ instead of $y_{1,t}$). To further simplify notations, reference to time step is also dropped when possible.

This model was chosen because it has been applied to a wide range of catchments across the world (Huard and Mailhot 2008, Ditthakit, Pinthong et al. 2021). It also has a smooth and parsimonious structure which simplifies the application of DAISI. Finally, GR2M shares its production store with the daily GR4J model which has been applied even more widely than GR2M, especially in Australia (Lerat, Thyer et al. 2020, Hapuarachchi, Bari et al. 2022). The monthly time step further simplifies the process by reducing time lags between the model variables and corresponding observations that can penalize certain assimilation schemes significantly (Li, Ryu et al. 2014).

In this paper, the GR2M model is calibrated by maximizing the Kling-Gupta Efficiency (KGE, Gupta, Kling et al. 2009). The calibration algorithm is a two-step approach where 10,000 random parameter sets are first generated followed by a Nelder-Mead gradient descent (Nelder and Mead 1965) applied to the best parameter set. The overall algorithm is detailed in Lerat, Thyer et al. (2020). This configuration is referred to as “GR2M-kge” in the rest of this paper.

To assess the influence of the objective function on GR2M performance and compare it with the DAISI performance, a benchmark configuration is obtained by calibrating GR2M using the sum of squared Box-Cox transformed flows with an exponent of 0.2. McInerney, Thyer et al. (2017) found that this objective function is a satisfactory compromise between a wide range of performance metrics. The function is computed as follows:

$$BC02(\tilde{q}^o, \tilde{q}^s) = \sum_t \left[BC(\tilde{q}_t^s, 0.2) - BC(\tilde{q}_t^o, 0.2)\right]^2 \qquad \text{Eq. 19}$$

Where $BC(\tilde{q}_t, \lambda)$ is the Box-Cox transformation of $\tilde{q}_t$ with exponent λ. In the rest of this paper, the calibration of GR2M using the BC02 objective function is referred to as “GR2M-bc02”. Several other benchmarks are presented in Supplementary Material S2.

As part of DAISI Step 1, the linear Ensemble Smoother data assimilation algorithm described in Appendix A was applied to GR2M using the parameters obtained from the GR2M-kge calibration and ensemble of $R = 500$ members following the algorithm described in Appendix A. For the DAISI Step 2 (model update), the expected update coefficients introduced in Eq. 13 are computed for the four state variables described in Table 1.

It is highlighted that the update terms applied to the effective rainfall (δ_{p_e}) and the simulated streamflow (δ_q) are flux corrections. In other words, these corrections alter the way GR2M computes effective rainfall and streamflow from its inputs and internal variables. The interpretation of the update terms corresponding to the production (δ_{s^+}) and routing store (δ_{r^+}) is more subtle. Via rearrangement of the model equations, Appendix C concludes that the opposite of the sum $\delta_{p_e} + \delta_{s^+}$ is the update term for the actual evapotranspiration, hence a correction on the evapotranspiration flux. Similarly, the opposite of the sum $\delta_q + \delta_{r^+}$ is the update term for the inter-basin exchange flux, hence a correction on this flux.

3.2. Model Evaluation

The GR2M model calibration and application of the DAISI method were implemented within a split-sample cross-validation scheme where the GR2M model parameters, analyzed ensembles and update coefficients are obtained using half of the total data period. These model parameters and update coefficients are then applied to the second half of the period without any use of data assimilation. Both sub-periods are subsequently exchanged. As detailed in the following section, the study region used in this paper experienced a prolonged dry period during the second half of the period known as the “Millenium drought” (Chiew, Potter et al. 2014), which leads to significantly different hydro-climate conditions between the two sub-periods.

Overall, three modelling scenarios are run for each catchment and each validation period: (1) a GR2M simulation using parameters calibrated over the independent corresponding calibration period with the KGE objective function (GR2M-kge), (2) GR2M calibrated with BC02 objective function (GR2M-bc02), and (3) the DAISI updated model structure using the GR2M-kge parameters and update coefficients fitted for the same calibration period.

3.3. Catchment Dataset

The DAISI approach was tested on a set of 201 catchments located in South-Eastern Australia as shown in Figure 2. The hydro-climatic catchment characteristics are provided in Table 2. The data was extracted from the datasets collated by Lerat, Thyer et al. (2020) including rainfall and potential-evapotranspiration data obtained from the Bureau of Meteorology Australian Water Outlook website (Frost, Ramchurn et al. 2016) and streamflow data obtained from the Bureau of Meteorology Water Data Online website (Bureau of Meteorology 2019). The data was collected over the period from 1980 to 2018, split into the two sub-periods 1980-1999 (Period 1) and 1999-2018 (Period 2).

In addition, two sub-groups of stations including stations located in Western Victoria (WVIC) and Northern New South Wales (NNSW) are located in Figure 2 to support the presentation of results in Section 4.

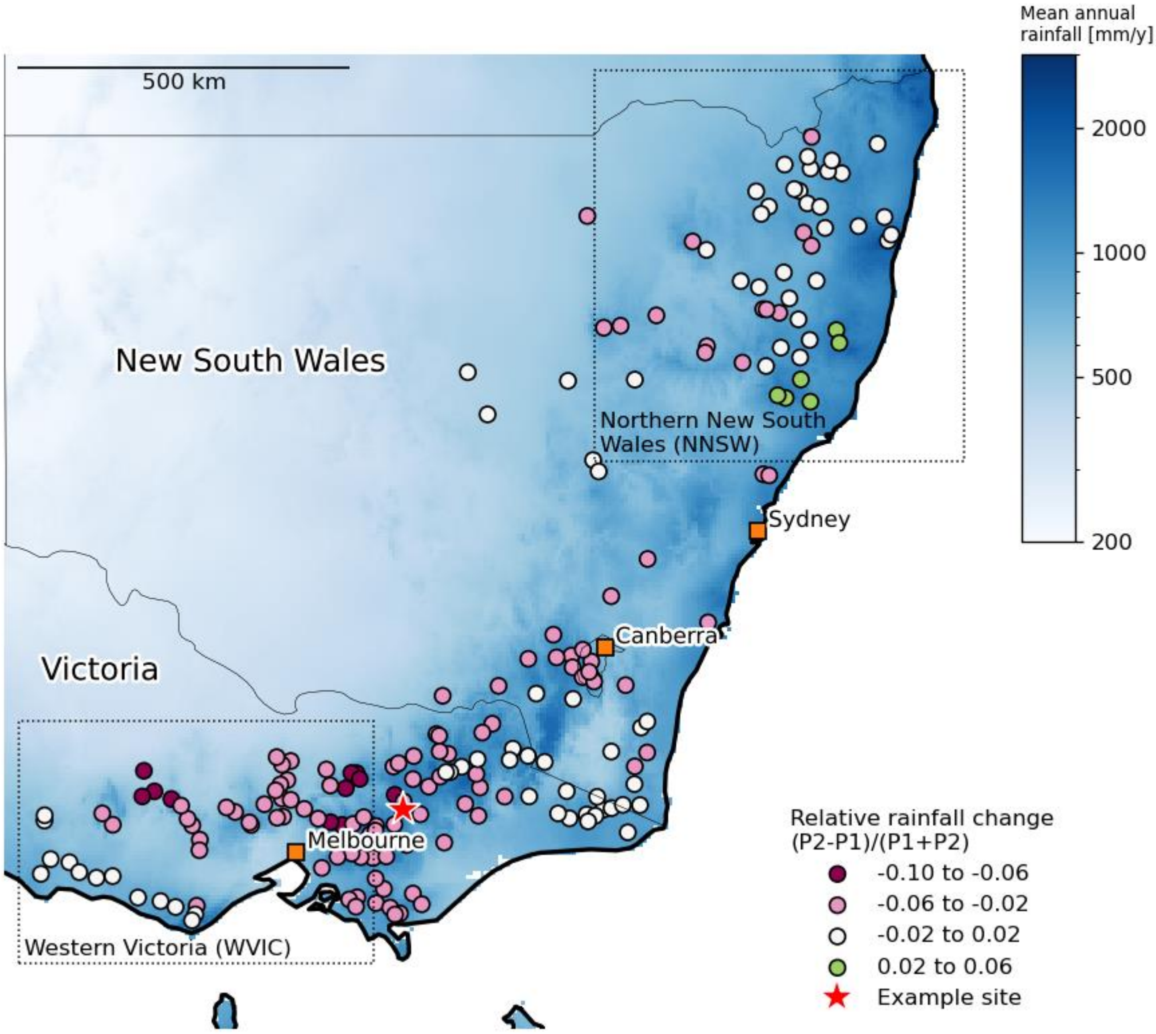

Figure 2: Site locations and relative change in rainfall between the two calibration periods

Table 2 highlights the predominance of semi-arid conditions in this dataset with median runoff coefficients of 0.17 and 0.12 for periods 1 and 2, respectively. The table also reveals that Period 1 was much wetter with median runoff of 160 mm/y against 113 mm/y for Period 2. Figure 2 shows that the relative reduction of rainfall between periods 1 and 2 reaches -10% for certain catchments located in the state of Victoria.

Among the 201 study catchments, the Jamieson River at Gerrang Bridge (station id 405218) was selected as an example to illustrate the DAISI method. This catchment represents 9% of the total catchment area of lake Eildon, one of the largest reservoirs in Australia with a maximum storage capacity of 3,334 Mm^3. Lake Eildon is a key piece of infrastructure which supports irrigation and environmental flows along the Goulburn and Murray Rivers.

Table 2: Hydro-climatic characteristics of the 201 case study catchments

Variable	Period	Min	Q25	Median	Q75	Max	*Gerrang (405218)*
Catchment area (km2)	-	54	180	388	766	34179	364
Mean annual rainfall (mm/y)	1980-1999	360	765	914	1116	1733	1190
	1999-2018	341	724	884	1043	1795	1093
Mean annual PET (mm/y)	1980-1999	1077	1222	1291	1405	1953	1221
	1999-2018	1079	1232	1292	1394	1982	1211
Mean annual streamflow (mm/y)	1980-1999	8	84	160	288	899	577
	1999-2018	3	53	113	213	745	486
Aridity index rain/PET (-)	1980-1999	0.18	0.56	0.71	0.86	1.58	0.98
	1999-2018	0.17	0.54	0.69	0.81	1.48	0.9
Runoff coeff. streamflow/rain (-)	1980-1999	0.02	0.11	0.17	0.25	0.61	0.48
	1999-2018	0.01	0.07	0.12	0.2	0.53	0.44

4. Results

4.1. Example of DAISI Workflow Applied to the Jamieson River at Gerrang Bridge

This section follows the three steps of the DAISI workflows applied to the example catchment. The parameters and diagnostic metrics for this catchment are provided in Table 3.

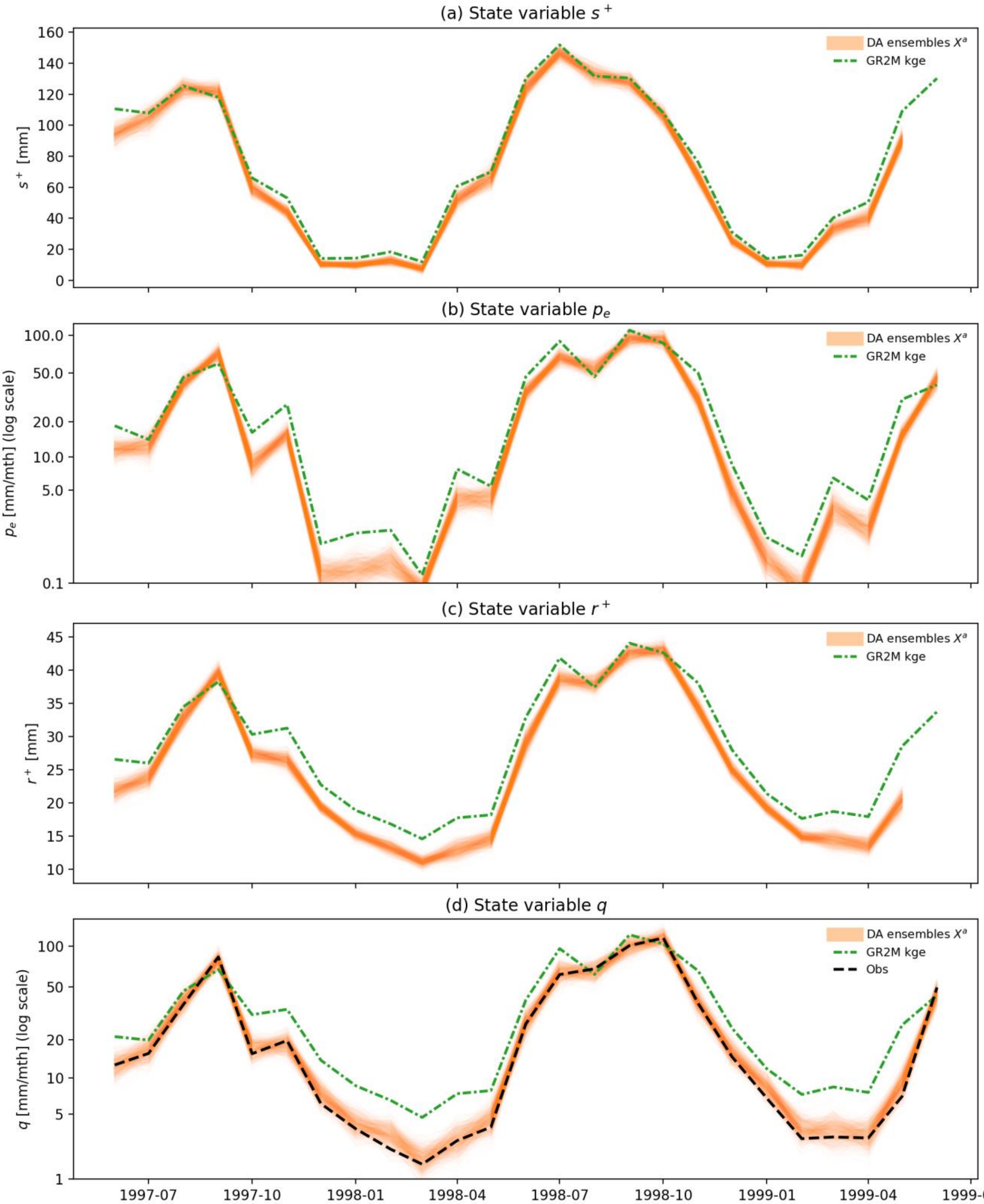

Figure 3: DAISI Step 1 – GR2M-kge variables (green dashed line) and assimilated ensembles (orange lines) for the Jamieson River at Gerrang bridge catchment. The plot covers the last two years of the calibration period. Plots (b) and (d) use a log scale for the vertical axis. The black dotted line in plot (d) is the observed streamflow.

4.1.1. Step 1: Data Assimilation

Figure 3 illustrates Step 1 of DAISI by showing as orange lines the ensemble time series resulting from the Ensemble Smoother data assimilation algorithm applied to the GR2M model calibrated using the KGE objective function (GR2M-kge) over the first period (1980-1999). Each plot in this figure corresponds to the four states listed in Table 1. In addition, the figure shows the original GR2M-kge simulations in green along with the observed streamflow data as black dotted lines in Figure 3.d.

The comparison between streamflow observations and GR2M-kge simulations in Figure 3.d highlights the systematic overestimation of low to mid flows by GR2M-kge. This overestimation is particularly pronounced between the second half of 1997 and the first half of 1998 for which GR2M-kge simulation stays above 5mm/mth whereas observations are close to cease-to-flow conditions with values as low as 1 mm/mth. As can be seen in Figure 3.d, assimilation corrects the low-flow bias of GR2M-kge by bringing the ensemble closer to streamflow observations. During high flow periods, the assimilation does not affect the simulation significantly as can be seen during the period from July to October 1998.

Assimilation impacts the GR2M routing store shown in Figure 3.c in a similar way than streamflow by decreasing the store level by 5 to 10 mm during the low flow periods compared to the original model. Like the two previous variables, the assimilated effective rainfall (P_e, see Figure 3.b) is reduced during low-flow periods but remains largely unaffected during high flow periods. The assimilated production store level (S^+) shown in Figure 3.a remains close to its value in GR2M-kge throughout the simulation. Overall, the effect of data assimilation decreases for state variables located further apart from streamflow within the model structure. This is expected as their correlation with observed streamflow estimated via the Kalman gain matrix is likely to decrease (see Appendix A).

The RMSE ratio metric N_R reported in Table 3 measures the statistical reliability of assimilated streasmflow ensembles and reaches 0.73 and 0.78 for the two calibration periods. These values are below one, which denotes an ensemble that is slightly too wide. Similar results are obtained across the whole catchment data set; hence the discussion of this point is deferred to Section 4.2.

Table 3: Model parameters and metrics for the Jamieson River at Gerrang Bridge. Numbers highlighted in bold correspond to metrics computed over a validation period.

Variable	Evaluation period	Calibration over Period 1 (1980-1999)			Calibration over Period 2 (1999-2018)		
		GR2M KGE	*GR2M BC02*	*DAISI*	*GR2M KGE*	*GR2M BC02*	*DAISI*
$\boldsymbol{\theta_1}$ (mm)	-	238	319	238	212	249	212
$\boldsymbol{\theta_2}$ (-)	-	1.12	1.01	1.12	1.12	1.01	1.12
$\mathbf{N_R}$ (-)	P1	-	-	0.73	-	-	-
$\mathbf{N_R}$ (-)	P2	-	-	-	-	-	0.78
KGE (-)	P1	0.86	0.70	0.94	**0.86**	**0.74**	**0.92**
KGE (-)	P2	**0.82**	**0.65**	**0.87**	0.83	0.70	0.88
NSElog (-)	P1	0.84	0.92	0.92	**0.83**	**0.91**	**0.91**
NSElog (-)	P2	**0.84**	**0.90**	**0.92**	0.84	0.91	0.92
$\boldsymbol{F_B}$ (-)	P1	0.61	0.82	0.91	**0.60**	**0.83**	**0.86**
$\boldsymbol{F_B}$ (-)	P2	**0.63**	**0.84**	**0.93**	0.63	0.85	0.88
$\boldsymbol{\epsilon_P}$ (-)	P1	1.66	1.82	1.84	**1.64**	**1.76**	**1.70**
$\boldsymbol{\epsilon_P}$ (-)	P2	**1.75**	**1.93**	**1.98**	1.73	1.87	1.84

4.1.2. Step 2: Model Structure Update

In Step 2 of DAISI, the update equation (Eq. 5) is fitted for each assimilated ensemble to obtain the update coefficients $\tilde{\eta}_n$ for each state variable. The process is illustrated in Figure 4 where the fitting is undertaken using data from the first calibration period. Figure 4.a and Figure 4.c show time series of the update terms corresponding to the routing store state (y_{r^+}) and the first two ensemble members. The assimilated updates (i.e., difference between assimilated variables and values computed using GR2M original equations as defined in Eq. 8) are shown as dashed orange lines while the predicted updates computed from Eq. 6 are shown as plain blue lines. For both ensemble members, the predicted update captures the general trends of the assimilated updates: both updates are close to 0 during the high flow periods from July to October 1998 while being negative during earlier low flow months. However, the variability of predicted updates appears to be underestimated as can be seen during the low flow period from October 1997 to June 1998. This result reveals the limitations of the regression model used in the update equation which can only explain a part of the variability seen in the assimilated updates. Figure 4.b and Figure 4.d provide a more detailed analysis of the performance of the update equation by plotting predicted (on the horizontal axis) versus assimilated (on the vertical axis) updates for the first two ensembles. In these two plots, the points appear scattered around the 1:1 line (dotted line) in the lower left part of the plot which suggests that the predicted updates exhibit large differences with the assimilated updates for low updates values. The predicted updates are much closer to assimilated updates for large updates with points clustered along the 1:1 line. Overall, the Pearson correlation between assimilated and predicted updates is close to or above 0.4 (shown in bottom right corner of the plots) indicating that the regression captures the main trends of assimilated updates but does not reach a high predictive power. Note that Figure 4.a to Figure 4.d are limited to the first two ensembles. Figure 4.e and Figure 4.f expand the analysis by showing both assimilated and predicted updates for all ensembles. Here again, the predicted updates correlate with the assimilated updates, but lack a high predictive power.

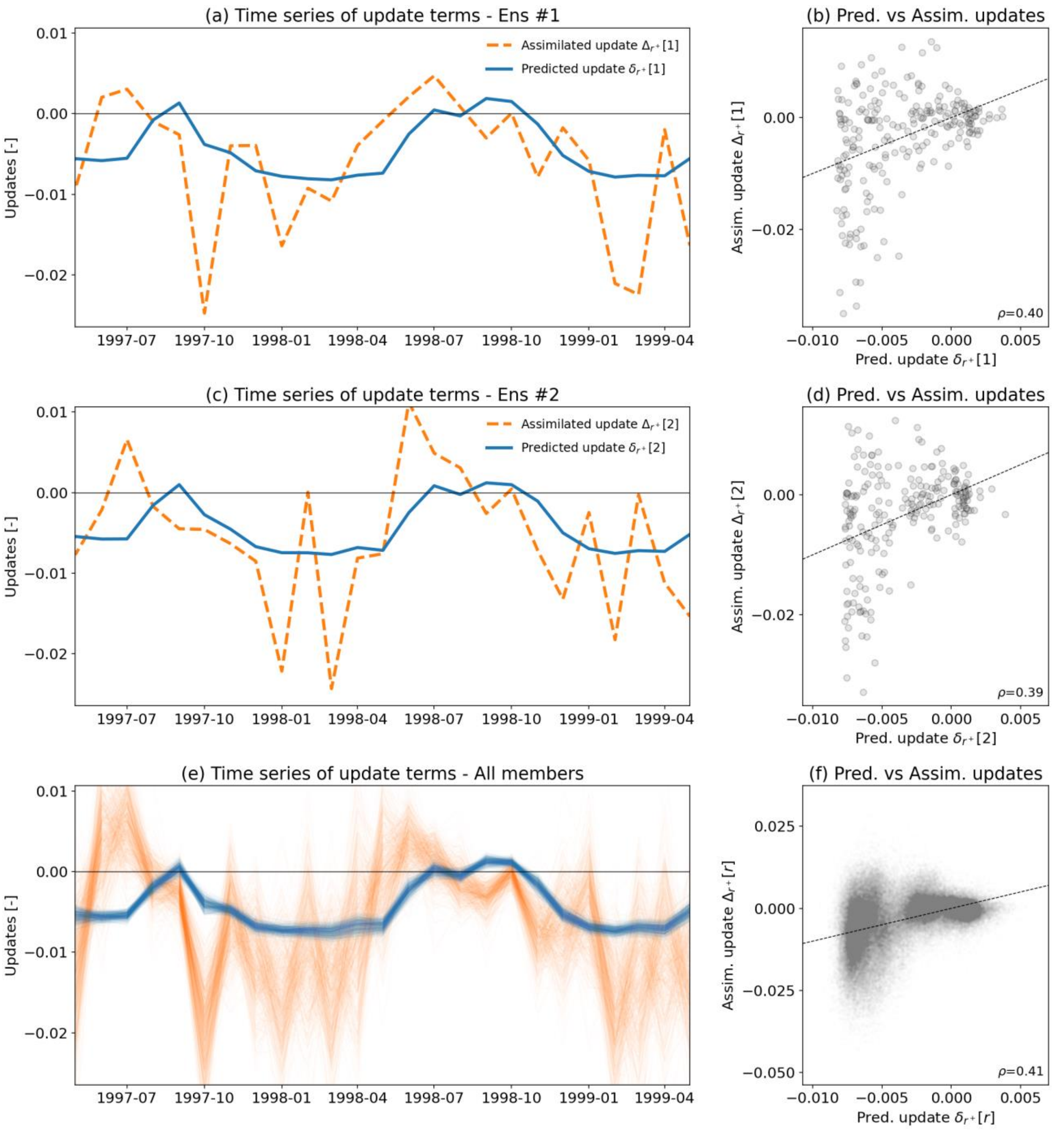

Figure 4: DAISI Step 2 - Assimilated ($\widetilde{\Delta}_{r^+}$) and predicted ($\tilde{\delta}_{r^+}$) update terms for the routing store level (R^+ state variable, see Table 1) and the first two assimilated ensemble members in plots (a) and (b). Plots (b) and (d) show the predicted versus assimilated update for the same ensembles along with the Pearson correlation coefficient between assimilated and predicted updates shown in the lower right corner of each plot. Updates from the 500 ensemble members are shown in plot (e) and (f). Data relates to the Jamieson River at Gerrang Bridge catchment and the first calibration period (1980-1999).

Figure 4.e also highlights the high uncertainty of assimilated updates during the low-flow period between October 1997 and June 1998 during which the updates jump from low to high values following a noisy pattern. This point illustrates the challenge of selecting a suitable update equation able to capture the important trends of the updates without reproducing its noise which is unavoidable in the presence of uncertain data and empirical model structures.

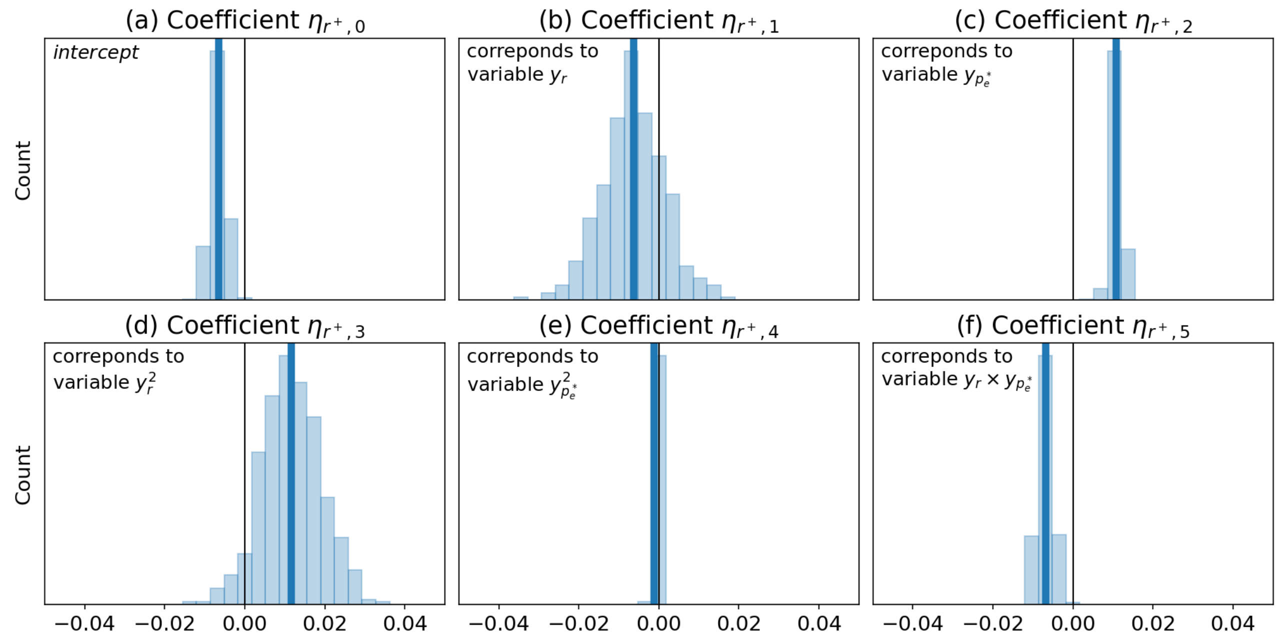

Figure 5: DAISI Step 2 - Distribution of update coefficients $\tilde{\eta}_{r^+}$ for the R^+ state variables (routing store) and the Jamieson River at Gerrang Bridge catchment in the first calibration period (1980-1999). The variable corresponding to the coefficient is given in the top left of each plot. The expected coefficient is shown as a vertical dark blue line.

The distribution of the update coefficients for the routing store (y_{r^+}) resulting from the fitting of the update equation is shown in Figure 5. Plots corresponding to the remaining three state variables are shown in Supplementary Material 1. As indicated in Table 1, variable y_{r^+} depends on two state variables ($y_r, y_{p_e^*}$), hence requiring 6 coefficients to be fitted. The expected value of each coefficient ($\tilde{\eta}_n^a$) is represented by a vertical blue line. The predictor variable corresponding to the coefficient in the update equation is indicated in the top left of each plot. Figure 5 reveals that most coefficients are statistically significantly different from zero with a majority of the probability mass located on either side of 0 (black vertical line), which suggests that most predictors play a significant role in the update equation. There are exceptions: for example, the distribution of coefficient $\eta_{r^+,4}$ (Figure 5.e) is centered around 0, which suggests that the $y_{p_e^*}^2$ variable could be excluded from the update equation without much loss to its predictive power. Such predictor selection could lead to a more parsimonious update equation. It was not undertaken in this paper to keep the fitting of the update equation consistent across all sites and calibration periods.

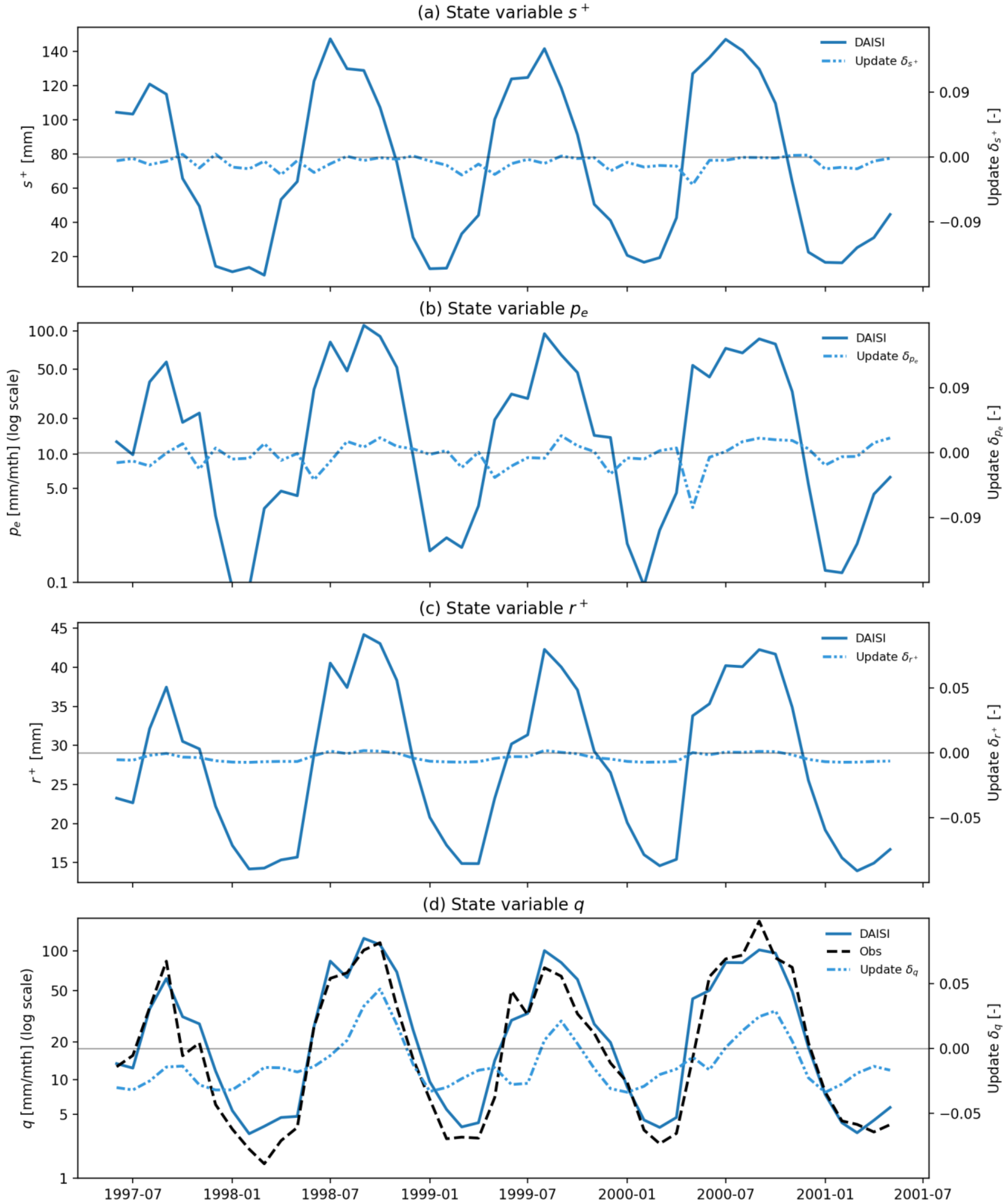

Figure 6: DAISI Step 3 – Updated model simulations (blue line) and update terms (dashed blue line using secondary vertical axis) for the Jamieson River at Gerrang Bridge catchment. The plot covers the last two years of the first calibration period and the first two years of the second validation period. Plot (b) and (d) use a log scale for the vertical axis. The black dotted in plot (d) is the observed streamflow.

4.1.3. Step 3: Model Diagnostic

Once the expected update coefficients are computed (vertical blue line in Figure 5), the updated model can be run and DAISI proceeds to Step 3. We highlight that data assimilation and coefficient fitting are not used in this step and that the updated model runs exactly like GR2M aside from its modified structure.

Figure 6 shows time series of state variables for the updated model along with the update terms δ_n computed from Eq. 6. The data covers the last two years of the first calibration period and the first two years of the following validation period. A comparison between Figure 3 and Figure 6 suggests that the updated model reproduces the behavior of the assimilated ensembles reasonably well. For example, it corrects the GR2M-kge overestimation of low flows with runoff simulations that are closer to observations in Figure 6.d. More important, this finding also applies to the validation period, for example between January 2001 and June 2001 in which the updated model fits the observed flow particularly well. This result demonstrates that the structural changes introduced by DAISI persist beyond the calibration period and can improve simulations during an independent validation period as confirmed by the performance metrics listed in Table 3 and discussed in the following paragraph.

Figure 6 provides further insights on the update terms shown as dotted lines. For example, Figure 6.d shows that the streamflow updates δ_q are negative during most of the period except around high flow peaks (e.g., September1998 and August 1999) where the updates become positive. This means that the updated model reduces low flows and increases high flows compared to GR2M. The updates of the routing store δ_{r^+} shown in Figure 6.c exhibit a much smaller amplitude than δ_q, which indicates that the sum $\delta_{r^+} + \delta_q$ is close to δ_q. As shown in Appendix C, the opposite of this sum is the update term for the GR2M inter-basin exchange flux (amount of water leaving the catchment unaccounted). Consequently, the update of the inter-basin exchange flux is close to $-\delta_q$. In other words, when the updated model increases streamflow compared to GR2M, the inter-basin flux is reduced by the same amount. Considering this explanation, Figure 6.d reveals that the updated model increases the inter-basin flux during low flows (negative δ_q), perhaps to increase losses to ground water. Conversely, the flux is decreased during high flows (positive δ_q). A similar analysis is more complex for the updates related to the production store shown in Figure 6.a and Figure 6.b as the two updates δ_{s^+} and δ_{p_e} are of similar magnitude.

Table 3 displays the four evaluation metrics underlying the diagnostic performed in Step 3 of DAISI computed for the GR2M-kge, GR2M-bc02 and the updated model (DAISI). The three performance metrics (KGE, NSElog and F_B) indicate a significant performance improvement of DAISI compared to both GR2M configurations for both calibration and validation periods. For example, KGE increases from 0.82 and 0.65 for the two GR2M configurations to 0.87 for DAISI when calibrating on Period 1 and evaluating on Period 2, and from 0.86 and 0.74 to 0.92 when calibrating on Period 2 and evaluating on Period 1. Similar metric improvements are seen for both NSElog and F_B metrics with DAISI reaching systematically higher performance.

These improvements, especially when evaluating the model outside of the calibration period, suggest that the updated model is a robust alternative to GR2M. At the same time, the modelled rainfall elasticity ϵ_P is generally higher for the updated model compared to both GR2M configurations. For example, ϵ_P increases from 1.75 for GR2M-kge and 1.93 for GR2M-bc02 to 1.99 for the updated model when calibrating on Period 1 and evaluating on Period 2. Note that Period 1 was significantly wetter than Period 2 with a mean annual rainfall of 1190 mm/year compared to 1093 mm/year for Period 2 as indicated in Table 3, which constitutes a valuable test to explore future climate scenario that are likely to be drier than present condition in South-East Australia (Charles, Chiew et al. 2020). Given that the updated model improves all performance metrics compared to both GR2M configurations, it seems reasonable to assume that these elasticities are closer to the true elasticity, and hence better suited to evaluate the impact of future climate scenario. The high elasticity computed from the updated model suggests that the variability of future runoff projections will increase significantly compared to GR2M-kge, which is an important finding in a catchment contributing to inflows into one of the largest dams in Australia.

For the sake of brevity, the presentation of other diagnostic tools introduced in Section 2.4 is not done for the example catchment. Section 4.3 presents the application of these tools to the whole catchment dataset.

4.2. DAISI Evaluation Metrics Computed for 201 Catchments

Following the application of DAISI Step 1 and 2 to the 201 catchments of our dataset, this section and the next present the diagnostic obtained from DAISI Step 3.

The distribution of the Normalized RMSE ratio N_R measuring the statistical reliability of the assimilated ensembles is presented in Table 4 for the 201 catchments and the two calibration periods. The 25th percentile, median and 75th percentile are 0.70, 0.81 and 0.95. These values are lower than 1, indicating that the assimilated ensemble is slightly too wide for most catchments across the dataset. Supplementary Material S3 suggests that statistical reliability of the assimilated ensemble can be improved by tuning the variance reduction factor α_e in the data assimilation algorithm (see Appendix A). Such tuning was not undertaken here to keep the assimilation scheme as simple as possible and because it does not have a significant impact on performance metrics (see Supplementary Material S3). Overall, this result suggests that the Ensemble Smoother algorithm reaches reasonable performance but could be improved, which is hardly surprising considering the strong linearity assumption underlying this data assimilation algorithm. This point will be further discussed in section 5.3.

Table 4: Distribution of Normalized RMSE ratio (N_R) computed from assimilated ensembles (DAISI Step 1) over the 201 test catchments and the two calibration periods.

Statistic	***Normalized RMSE ratio (N_R)***
Min	0.42
Q25	0.70
Median	0.81
Q75	0.95
Max	1.40

Figure 7 presents the distribution of the four metrics computed for the 402 catchments/periods over independent validation periods. The bar plots presented in the right-hand side of each plot show the percentage of catchments/periods where metrics for the updated model (DAISI) are larger, similar or lower than GR2M-kge and GR2M-bc02 by more than 0.05. Figure 7 reveals that the median value of KGE, NSElog and flow duration curve bias index F_B is systematically higher for the updated model compared to both GR2M configurations, which confirms the superiority of the former over the later. With KGE in Figure 7.a, the increase is small between the updated model (median of 0.69) and GR2M-kge (median of 0.65). However, it is much larger when comparing the updated model against GR2M-bc02 (median of 0.52). For NSElog in Figure 7.b, the increase is large between the updated model (median of 0.76) and GR2M-kge (median of 0.68) but insignificant between the updated model and GR2M-bc02 (median of 0.75). F_B shown in Figure 7.c follows a similar pattern than NSElog. In terms of pairwise comparison, the updated model always obtains similar or better performance than the best of GR2M configuration for a majority of catchments and periods. For example, DAISI reaches a KGE that is significantly better than GR2M-kge in 43% of catchments/periods and similar to GR2M-kge in 46%. Against GR2M-bc02, these figures reach 65% and 20% of catchments/periods. All other metrics provided in the supplementary material S2 confirm these results showing that DAISI leads to a consistent and reliable improvement of performance compared to GR2M across all flow regimes. Even if the performance improvement is modest for certain metrics (e.g., KGE metric when comparing against GR2M-kge), the number of catchments where DAISI is worse than GR2M remains limited which suggests that the updated model does not introduce major structural trade-offs (e.g., favoring a certain type of catchments against another).

Overall, the updated model combines the strength of both GR2M configurations by equaling or exceeding their combined maximum for all performance metrics. This is an important result as it suggests that the DAISI structural updates surpass the performance obtained from alternative objective functions.

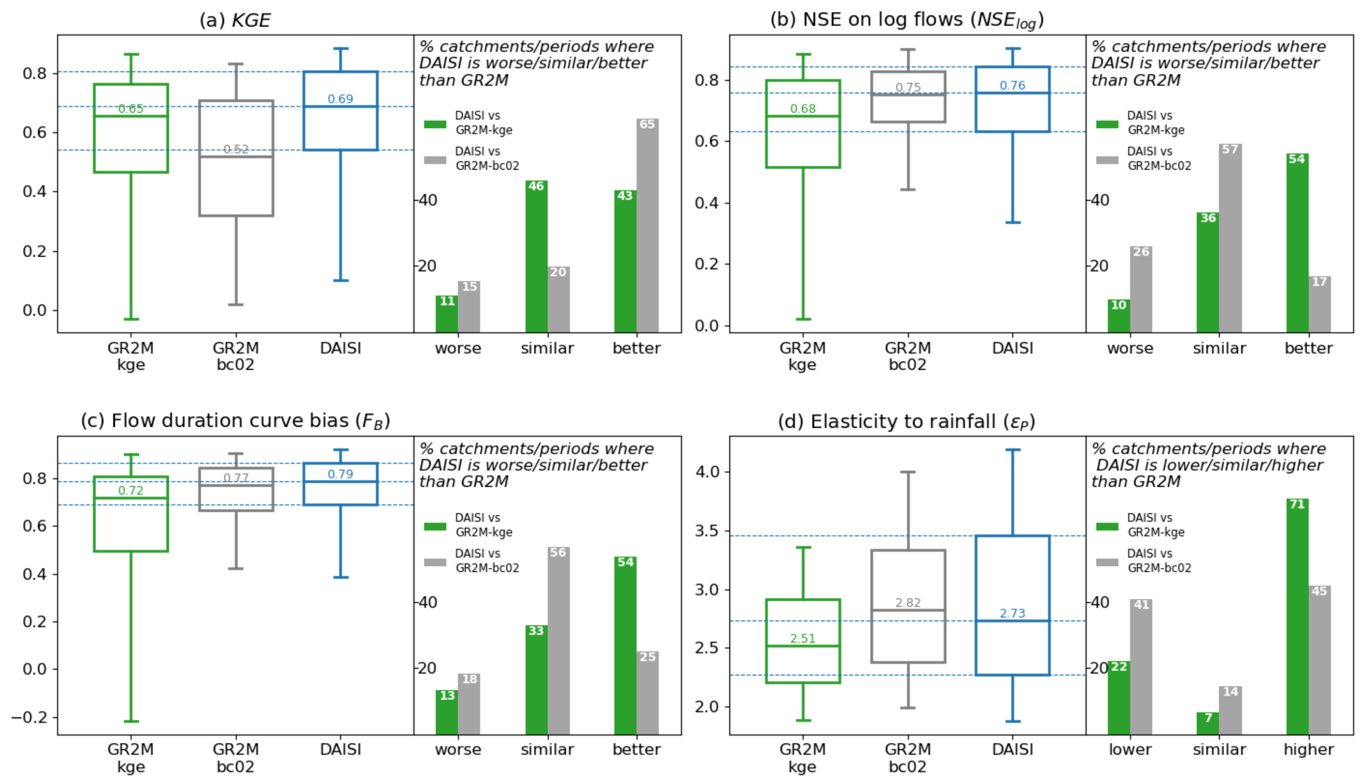

Figure 7: DAISI Step 3 evaluation metrics for GR2M calibrated using the KGE objective function (green), GR2M calibrated using the BC02 objective function (grey) and DAISI updated model (blue) for the 201 catchments. Metric values are computed for the two validation periods for each catchment. The bar charts on the right of each plot indicate the percentage of catchments/periods where DAISI is lower/similar/higher or worse/similar/better than GR2M-kge and GR2M-bc02.

The fourth evaluation metric is the elasticity of modelled runoff to rainfall shown in Figure 7.d. The median elasticity is 2.52 for GR2M-kge which increases to 2.74 for the updated model. Pairwise comparisons confirm this result with updated model elasticity being significantly higher than GR2M-kge in 71% of the site/periods. Comparing the updated model against GR2M-bc02 reveals that both models reach similar elasticity values with equal proportions of sites/periods where one is greater than the other. However, Supplementary Material S4 shows that when using BC02 as an objective function, DAISI obtains a significantly higher elasticity on a majority of sites/periods (median elasticity of DAISI reaches 2.96 in this case). Consequently, it can be said that DAISI generally leads to higher elasticity values across the catchment dataset. Considering that the updated model obtains better or equal performance than GR2M for most performance metrics, the elasticity from the updated model is very likely to be closer to reality than the GR2M elasticity.

Figure 8 explores the evaluation metrics further by showing the spatial distribution of metric averages between the two validation periods for each catchment. The first column of the figure shows the metrics for GR2M-kge, the second the metrics for the updated model and the last column the difference between the two. Figure 8.a, b and c corresponding to the KGE metric reveal that there are strong spatial trends in the performance improvement brought by DAISI which is mostly occurring in catchments located in the Western part of the state of Victoria (WVIC, see Figure 2 for exact location of this region) and the North-Eastern part of the state of New South Wales (NNSW). For these two regions, KGE improvements are greater than +0.10 (dark green triangles in Figure 8.c). Improvement of rainfall-runoff model performance in the WVIC region is important because this region has been reported to suffer from strong rainfall-runoff non-stationarity with long lasting effects from recent drought (Peterson, Saft et al. 2021). Conversely, KGE values for the catchments located in the center of the domain (Eastern Victoria) are comparable between GR2M and the updated model (white dots). A closer inspection of Figure 8.a and Figure 8.b reveals that GR2M reaches its highest KGE values in these catchments (dark blue points in Figure 8.a). As GR2M simulations are of high quality there, it is difficult for DAISI to improve performance significantly. Nonetheless, it is important to note that DAISI does not degrade performance in this region.

The spatial distribution of performance differences for NSElog and F_B metrics resembles the one of KGE as can be seen in Figure 8.f and Figure 8.i. The updated model improves performance over GR2M in the WVIC and NNSW regions with limited gains in the central region. The rainfall elasticity follows the same spatial pattern with higher elasticity for the updated model compared to GR2M in WVIC and NNSW. It is worth noting that the GR2M elasticity in the WVIC and NNSW varies between 2.50 to 3.25 (light to dark blue points in Figure 7.d) which increases by up to +0.50 with the updated model (dark green triangles in Figure 8.l). This represents an increase in elasticity of 15% to 20%.

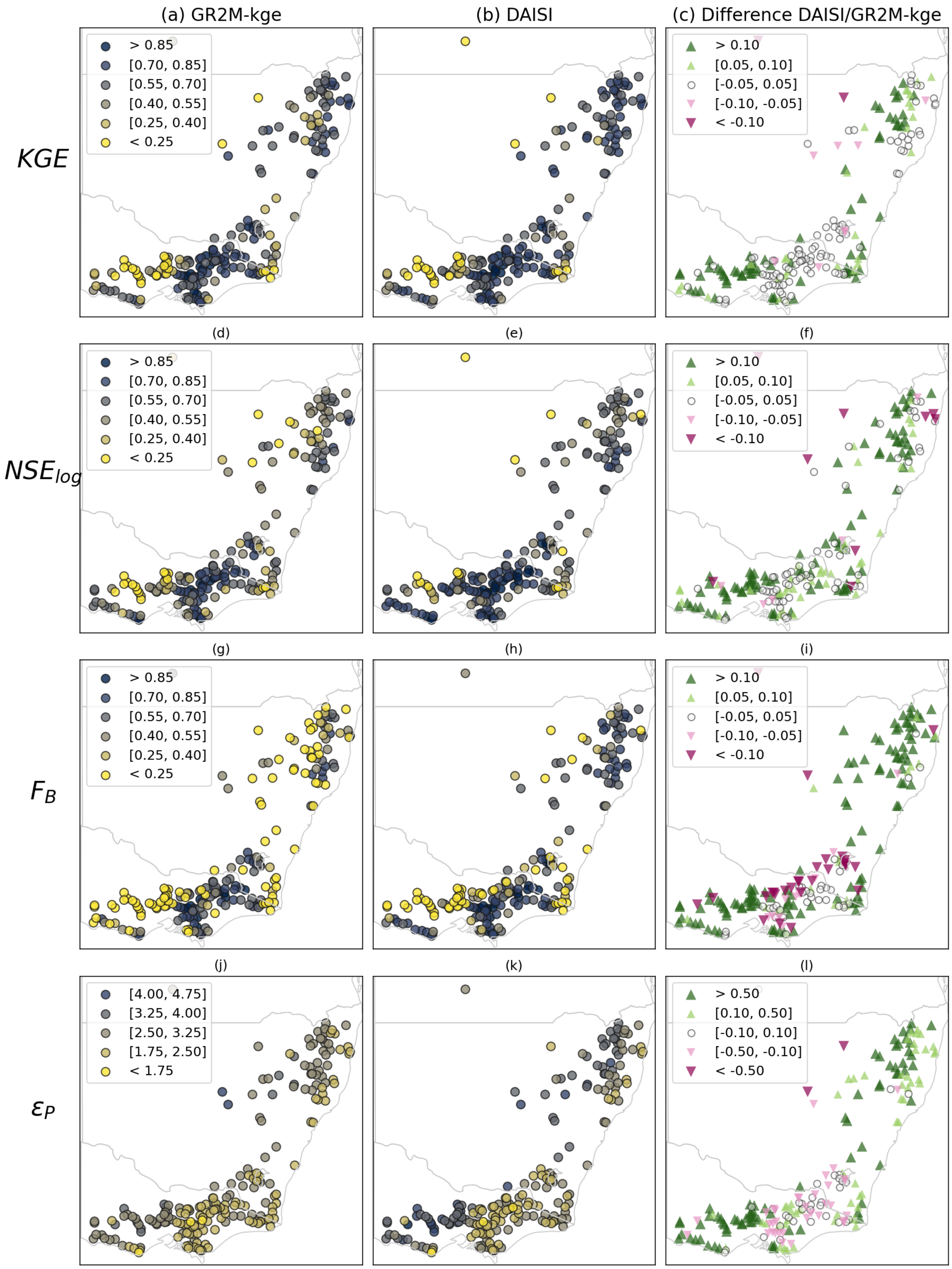

Figure 8: Spatial distribution of the four metrics for GR2M (first column) and updated model (DAISI, second column) over the 201 catchments. Metric values are computed from and averaged over the two validation periods for each catchment. The difference between DAISI and GR2M metrics is shown in the third column with green upper pointing (pink lower pointing) triangles showing catchment with better (worse) performance for DASI versus GR2M.

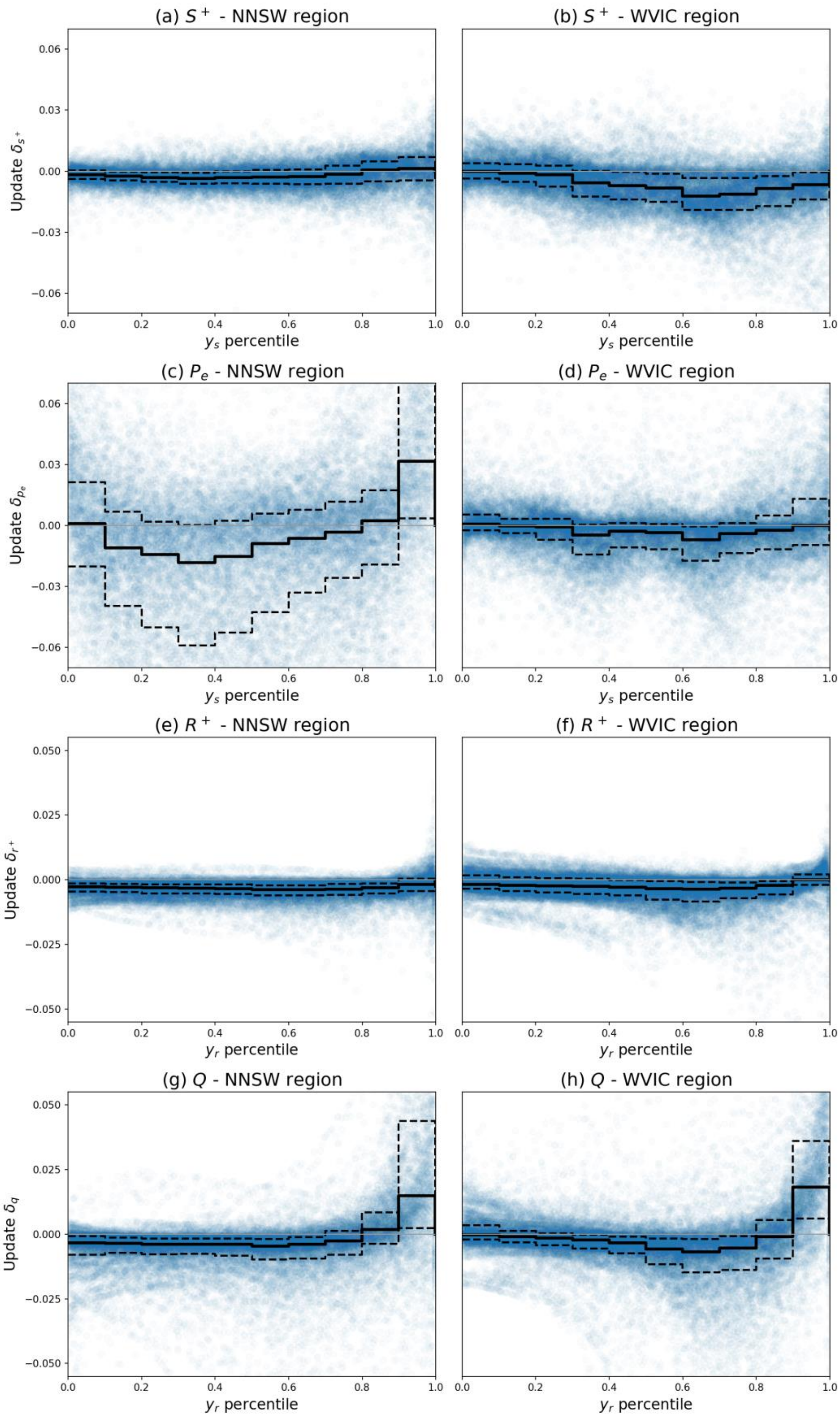

Figure 9: DAISI Step 3 - model structure diagnostic diagrams for four state equations (rows) and for catchments in the NNSW (first column) and WVIC (second column) regions. Data are from both calibration periods. The plots show the update term δ_n on the vertical axis. The horizontal axis shows the percentile rank of y_S for the first two rows and y_R for the last two rows. Medians (black line), 25% and 75% percentiles (dotted lines) of the update term are computed by binning the data according to the variable on the horizontal axis.

4.3. DAISI Model Structure Diagnostic for 201 Catchments

The previous section confirmed the diagnostic undertaken in the example catchment which concluded that the updated model is a better alternative to GR2M over the catchment data set considered in this paper. Building on this result, the second part of the DAISI Step 3 diagnostic can be undertaken using plots described in Section 2.4. Figure 9 shows scatter plots of the update terms on the vertical axis. The horizontal axis displays percentile ranks of y_s (production store) for the first two rows (Figure 9.a to Figure 9.d) and of y_r (routing store) for the last two rows (Figure 9.e to Figure 9.h).

The streamflow updates (δ_q) shown in Figure 9.g and Figure 9.h are similar for both regions with an update that is close to 0 for very low routing store levels, then decreasing to a median value of approximately -0.01 for percentiles of y_r up to 0.7. Above this value, the streamflow update increases rapidly with y_r reaching a positive median greater than +0.02 close to the maximum of y_r. This pattern explains why the updated model improves performance on low flows measured by the NSElog metric discussed in the previous section by lowering simulated flows when the routing store is low, and hence correcting the tendency of GR2M to overestimate low flows (see example in Figure 3). The positive update seen for high values of y_r leads to an increase in streamflow values if y_r is high, explaining the modest increase in mid to high flow performance measured by the KGE metric.

The routing store updates (δ_{r^+}) shown in Figure 9.e and Figure 9.f are of much smaller magnitude than the streamflow updates. This suggests that the sum $\delta_{r^+} + \delta_q$ is largely dominated by the latter which, as explained in section 3.1 and Appendix C, implies that the update term on the inter-basin exchange term (water gained or lost from the surface water catchment) is approximately equal to $-\delta_q$. In other words, when the updated model increases streamflow by δ_q, it decreases the exchange flux by $-\delta_q$.

The behavior of the streamflow and routing store updates appears similar in NNSW and WVIC regions as can be seen by comparing Figure 9.e with Figure 9.f. Conversely, the updates on the production store δ_{s^+} and effective rainfall δ_{p_e} reveal a striking difference between the NNSW and WVIC regions. In NNSW region, variations of δ_{s^+} (Figure 9.a) are negligible compared to those of δ_{p_e} (Figure 9.c) but in WVIC region, they are of similar magnitude (Figure 9.b and Figure 9.d). Based on section 3.1 and Appendix C, this suggests that $\delta_{s^+} + \delta_{p_e} \approx \delta_{pe}$ in NNSW , and hence that the update on actual evapotranspiration is close to $-\delta_{p_e}$. In WVIC we can assume that $\delta_{s^+} \approx \delta_{p_e}$ hence that the update on the actual evapotranspiration is approximately $-2\delta_{p_e}$. Consequently, the structural update affects the actual evapotranspiration twice as much in WVIC as NNSW. This is an important finding to improve the representation of evapotranspiration in the model depending on the modelling region.

To go beyond the previous qualitative analysis of the structural updates, Figure 10 presents the singular value decomposition of the update coefficient matrix introduced in Eq. 18. The results for streamflow (last row in Figure 10) are the easiest to interpret and are commented first. Figure 10.j shows the component of the first two singular vectors along with their confidence intervals and weights (see Eq. 18) in the legend. The total weight for these two vectors is 0.93 (sum of 0.73 and 0.20), which is close to the maximum of 1 and indicates that the update coefficients are well approximated by linear combinations of these two vectors. This is an important result because it suggests that the polynomial used to correct streamflow variable in the update equation (Eq. 5) can be described accurately across the whole dataset with two degrees of freedom only instead of the 6 coefficients used in Eq. 6. Furthermore, the singular vectors show narrow confidence intervals in Figure 10.j which suggests that they are not influenced by the catchment selection and potentially applicable to a wider range of catchments. As seen in Figure 10.j, the components of these vectors are significant for y_r (-0.24 for vectors #1 and -0.52 for vector #2), y_r^2 (0.76 and 0.41) and $y_r \times y_{p_e}$ (-0.60 and 0.73) and close to 0 for the intercept (0.06 and 0.12), y_{p_e} (0.07 and -0.10) and $y_{p_e}^2$ (0.04 and -0.03). More precisely, if we neglect the smallest coefficients, the first singular vector corresponds to the following update polynomial for the streamflow state variable:

$$\delta_q = -0.24 y_r + 0.76 y_r^2 - 0.60 y_r y_{p_e} \qquad \text{Eq. 20}$$

For a fixed value of y_{p_e}, this polynomial is equal to 0 when $y_r = 0$, then decreases with y_r to reach a minimum and finally increases with y_r. This analysis explains the patterns seen in Figure 9 and allows precising the structural diagnostic by clarifying the role of y_{p_e} which was not apparent in Figure 9. In this discussion, the precise numerical values of the coefficients in Eq. 20 are less important than the functional form of the update which narrows considerably the type of form to be considered for future model improvement. In addition, Figure 10.k shows that the first principal component exhibits strong regional trends. This component is negative for catchments in the WVIC and NNSW regions (dark purple triangles) which implies that the update for these catchments has a form similar to the opposite of Eq. 20. The second principal component shown in Figure 10.l is strongly positive for catchments located in the central region (dark green triangles). Consequently, the update equation in these catchments is similar to an equation like Eq. 20 with coefficients taken from the second component. Such information could be used to define different state equations in these regions.

The singular value decomposition of the update coefficient matrix for the R^+ state variable follows similar patterns than Q. The sum of the weights for the first two singular vectors is 0.95 (sum of 0.77 and 0.18, see legend in Figure 10.g) which means that the update equation can be represented accurately by linear combinations of two vectors only in most catchment across the dataset. The polynomial and regional trends (see figures Figure 10.h and Figure 10.i) associated with the singular vectors are similar to the ones for Q.

The singular value decomposition corresponding to S^+(Figure 10.a, b, c) and P_e (Figure 10.d, e, f) is more complex because the weights associated with the first two singular vectors are significantly lower than 1 (sum of 045+0.22=0.67 in Figure 10.a and 0.40+0.32=0.72 in Figure 10.c). For these two states, the low values of the weights reveal that the functional form of the update is more complex than linear combinations of the first two singular vectors in most catchments. In addition, the confidence intervals of the singular vector shown in Figure 10.a and Figure 10.c are relatively wide, suggesting that the component are affected by the catchment selection, and hence less likely to generalize beyond our dataset. The most striking element visible in in Figure 10.b is the strong regional trends shown by the first principal component for S^+. The component clearly differentiates the catchments located in the WVIC (positive component) from the ones in the NNSW region (negative component). Here again, these conclusions suggest that the improvement in state equations are region specific.

Overall, the DAISI diagnostic identified several directions to guide future improvement of the GR2M model including elements related to parameterization of the update and its regional trends. A summary of these directions is provided in section 5.2.

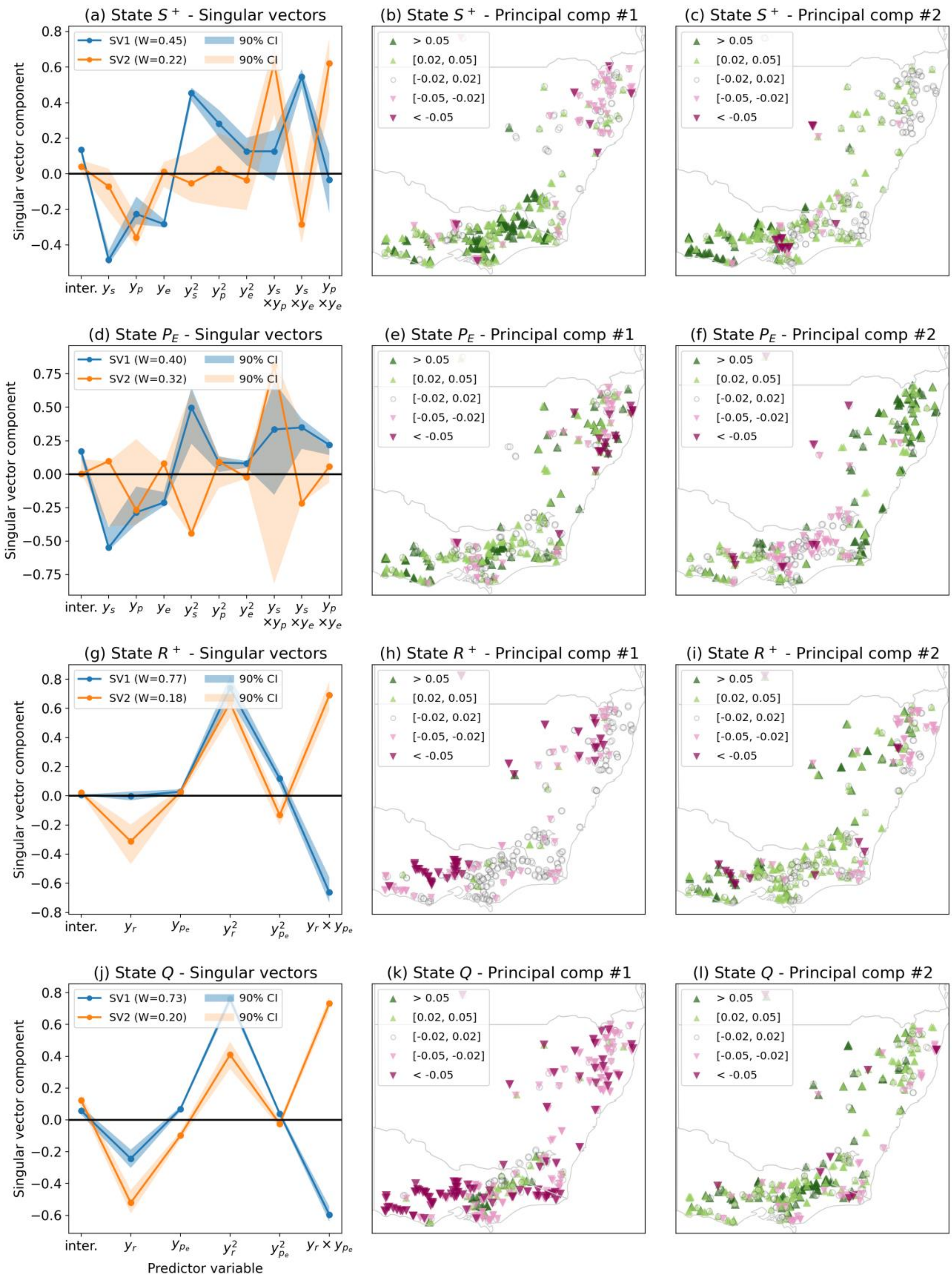

Figure 10: DAISI Step 3 - Reduced singular value decomposition of update coefficient matrices (see Eq. 18) for the four state equations (rows) and for the 201 catchments with data pooled from the two calibration periods. The plots in the first column show the components of the first two singular vectors along with their 90% bootstrap confidence intervals. The plots in the second and third columns show the projection of the update coefficient vectors for each site and calibration period on the first (second column) and second (third column) principal component, respectively.

5. Discussion

5.1. Advantages and Limitations of DAISI

Most existing methods used to improve model structures are based on trial and error using a finite set of structures that is arbitrarily selected by the modeler. Compared to this discrete approach, the first advantage of DAISI is that the exploration of alternative model structures is driven by data through data assimilation (DAISI Step 1) and fitting of the update equation (DAISI Step 2). This process can identify modelling solutions that were not considered a priori due to the complexity of formulating multi-dimensional state equations to represent physical processes that are often poorly quantified at the catchment scale (for example see discussion about the difficulty to close mass balance by Safeeq, Bart et al. 2021, Huang, Wang et al. 2023). DAISI also offers an alternative to trial and error by considering a continuum of model structures generated via the update equation (Eq. 5). The results presented in Sections 4.2 and 4.3 show that the updated equation improves performance significantly compared to the original model including for the simulation of contrasting hydro-climatic conditions, and converges to a reduced number of update configurations with clear regional patterns. At the same time, DAISI does not lose the potential for interpreting model equations based on a physical system understanding, which is the main issue with most machine learning approaches.

The second advantage of DAISI is its generic nature. The first two steps of DAISI, i.e., data assimilation and fitting algorithms, are mostly independent from the model structure and observed data. The only model specific element in DAISI is the normalization of state equations used to remove the influence of model parameters. This normalization is needed to compare the structural updates across different sites as is done in Section 4.3. However, we point out that it is not compulsory if the focus is on a single site or if the model parameters are identical across sites (for example when using a landscape model with same parameter values across a region). Consequently, DAISI is a general method that could be used to guide improvement for a wide range of models. This opens opportunities for applying DAISI to models outside the field of hydrology, for example ecology where state equations are often harder to identify than in hydrology due to the spatial variability and non-linearity of ecological processes (Cressie, Calder et al. 2009). It also allows DAISI to incorporate observed data beyond the traditional climate inputs used in empirical rainfall-runoff models. This has been attempted (for example accounting for artificial storage in the GR4J lumped model by Payan, Perrin et al. 2008) but remains a difficult exercise because model states in this type of model rarely correspond to observed data. In contrast, DAISI can incorporate additional data seamlessly via either the data assimilation algorithm by expanding the observed data vector $\tilde{d}$ (see appendix A) or by adding a predictor to the update equation. This point is further discussed in Section 5.3.

Finally, DAISI is a flexible and modular method where each step is independent from the others. For example, the aim of data assimilation in Step 1 is to evaluate the conditional probability of states given input and observation data via an ensemble. In this paper, the linear Ensemble Smoother described in Appendix A is used because of its limited computing requirements. However, any algorithm generating similar outputs more accurately could be used, which is discussed further in Section 5.3. Once an ensemble of states is generated, DAISI Step 2 fits the update equation to each assimilated ensemble. Here again the approach presented in the paper was chosen because of its parsimony and closed form solution but could be replaced with more powerful fitting techniques.

Despite the qualities highlighted above, the first obvious limitation of DAISI is that it requires an existing model structure to apply the update equation. Early attempts (not shown) of removing the existing state equation (f_n^*) from Eq. 5 and creating a fully data-driven model structure led to poorer performance than the original model, which highlighted the difficulty of producing a model structure without a strong prior knowledge. However, relying on an existing model has benefits including the possibility to remove the structural update completely and revert to the original model if needed. Such a case is discussed in Section 4.2 where the GR2M model was seen to perform well in the central region of our modelling domain leading to structural update becoming negligible (see Figure 8.c, f, i and l). In other words, the updated model identified by DAISI is unlikely to suffer from large reduction of performance against the original structure.

The second limitation of DAISI is the reliance on fixed model parameters obtained from a previous calibration exercise. A simple solution to overcome this limitation is to include parameters in the assimilated variables using the “state augmentation” technique (Vrugt, Diks et al. 2005, Pathiraja, Marshall et al. 2016). This approach was investigated (not shown) but did not lead to significant differences in both performance and diagnostic. Another more radical approach would be to repeat the whole DAISI process using parameters calibrated with different objective functions. This is done in Supplementary Material S4 where the GR2M model is calibrated using a box-cox transformed sum of squared errors following McInerney, Thyer et al. (2017). This exercise confirms that DAISI improves average performance for all metrics considered compared to GR2M but reveals that the largest improvements are obtained for different metrics compared to the ones identified in Section 4.2. This can be explained by the fact that the choice of objective function specializes the model in the simulation of a particular streamflow regime (for example KGE focuses on mid to high flows). Within DAISI, data assimilation and structural updates correct the largest errors, most likely outside of this streamflow range, and consequently improve the corresponding performance metrics (for example low flow metrics in when calibrating the model against KGE). Despite these performance differences, the results shown in Supplementary Material S4 suggest that a change in objective function did not affect most DAISI diagnostic plots, especially the singular value decomposition shown in Figure 10, revealing that the choice of objective function may not be a critical factor in the DAISI diagnostic. These results are encouraging but it is acknowledged that more research is needed to formally incorporate parameter uncertainty in DAISI.

The simplicity of both the data assimilation and fitting algorithms used in this paper is another limitation of DAISI which may constrain the performance of the method in its current form. As shown above, the data assimilation algorithm could be replaced with more flexible approaches. Regarding the fitting algorithm, the lack of physical constraints is an important issue because it leads to update terms that are potentially non-physical (e.g. negative streamflow) and requires truncation when running the updated model as shown in Appendix C. Extensive checks on modelled time series such as the ones presented in Figure 6 along with the computation of multiple performance metrics reported in Section 4.2 did not reveal any obvious non-physical behavior of the updated model. This is likely to be due to the small amplitude of the updates compared to original model values which rarely leads to exceeding physical constraints.

5.2. What have we learnt about the GR2M model?

The DAISI method applied to GR2M, and more specifically the diagnostic conducted in Step 3, identified several elements to guide further improvement of this model. First, extensive analysis of performance metrics computed over a period independent from the calibration period concluded that the updated model improves all metrics, especially the ones related to low flow simulations. The updated model also increases the elasticity of modelled streamflow to rainfall significantly compared to GR2M-kge, which suggests that structural updates produce a more robust model for modelling future streamflow under climate change.

Second, clearly defined structural updates are found for the lower parts of the model including the routing store (R^+) and simulated streamflow (Q). Streamflow values are altered in the updated structure by reducing mid-range values (negative update) while increasing high values (positive update) following a form similar to the polynomial of Eq. 20. The update for the routing store resembles the ones for streamflow but is of much lower magnitude, which lead to the conclusion that the updated model redistributes fluxes between streamflow and the inter-basin exchange (flux entering or leaving the surface water catchment). More precisely, the exchange flux is increased for low to mid-levels of the routing store and decreased for high levels of the routing store. Overall, these findings point to the need to modify the partition between streamflow and exchange flux in GR2M and relate this partition to the routing store level.

Third, the structural updates are less pronounced for the upper parts of the model including the production store (S^+) and effective rainfall (P_e). The updates reduce both variables for mid-range levels of the production store while leaving them unaffected for very low and very high values of the store.

Fourth, there are two regions where the structural updates in the production store differ significantly. In the Northern part of the state of New South Wales (NNSW), the updates of the effective rainfall are of much larger magnitude than updates of the production store level, which suggests that the updated structure introduced an equal redistribution of flux between effective rainfall and actual evapotranspiration compared to GR2M. In the Western part of the state of Victoria (WVIC), a similar flux redistribution is observed, but the modification in actual evapotranspiration is found to be approximately twice the change in effective rainfall. This more aggressive redistribution is likely to reduce production store level in this region, which is a recommendation formulated by Fowler, Knoben et al. (2020) while investigating the cause for poor performance of rainfall-runoff models in this region.

Despite all these findings, it is acknowledged that the updated model generated by DAISI remains heavily parameterized as it depends on the two original GR2M parameters and 32 update coefficients (see Table 1). Incorporating the finding identified above into a compact structure constitutes a logical follow-up of the work presented in this paper.

5.3. How can DAISI be improved?

This paper presented a first version of the DAISI method. As mentioned in the previous sections, it is currently limited by the simplicity of the data assimilation in Step 1 and fitting algorithm used in Step 2. More flexible assimilation algorithms, such as ensemble particle filter (Moradkhani, Sorooshian et al. 2005, Van Delft, El Serafy et al. 2009), could improve the quality of assimilated ensembles and allow the identification of more robust structural updates. In addition, the Ensemble Smoother (ES) data assimilation algorithm used in this paper is applied independently in each catchment, hence neglecting spatial correlation that is likely to exist between observation errors in neighboring catchments. Such an extension is relatively straightforward because ES was originally designed by van Leeuwen and Evensen (1996) to assimilated observations in large spatially explicit models.

The main issue related to the fitting algorithm was raised in section 5.1 with the lack of physical constraints in the fitting of update coefficients. This could be addressed by replacing the ordinary least squares solution introduced in Eq. 14 by a Bayesian regression with a censored predictand defined according to physical constraints (see for example the model developed by Wang, Robertson et al. 2009). However, such statistical models generally lack a closed form solution and rely on sampling methods, which would increase the computing time of DAISI significantly (the fit must be repeated for each assimilated ensemble).

Improving the current algorithms in DAISI as described above is important, but we believe that greater benefits would come from including more observed data. As mentioned in section 5.1, DAISI is flexible enough to incorporate additional observed data in the assimilation algorithm or in the fitting of the update coefficients. In South-East Australia, evapotranspiration has a significant impact on runoff which is expected to grow in future climate (Fowler, Knoben et al. 2020). Adding in-situ or remotely sensed actual evapotranspiration data to DAISI is possible and could lead to improvement in rainfall-runoff model structures for simulating both runoff and evapotranspiration.

Finally, it would be useful to extend the application of DAISI to daily models (for example GR4J) to confirm that the method can be applied to more complex structures and in the presence of delayed response.

6. Conclusion

This paper introduced the Data Assimilation Informed model Structure Improvement (DAISI) method which aims at analyzing and improving a hydrological model structure by combining the Ensemble Smoother data assimilation algorithm with polynomial updates applied to the model state equations. The method is generic, modular and was demonstrated with an application to the GR2M monthly rainfall-runoff model and a dataset of 201 catchments in South-East Australia.

The results show that the updated model generated with DAISI reaches higher median performance across the catchment data set for all metrics considered including KGE, NSE on log transform flow and flow duration curve bias. Performance improvement is largest for metrics measuring low flow performance such as log NSE where the updated model produced significantly higher performance score. In addition, the elasticity of modelled runoff to rainfall was shown to increase from a median of 2.51 for GR2M to 2.80 for the updated model, which is closer to the observed data, suggesting that the structural changes will lead to more robust modelling of future streamflow under climate change. Finally, the DAISI diagnostic identified a reduced number of update configurations in the GR2M structure with clear regional patterns. These configurations correspond to specific polynomials of the inputs to the state equations that could form the basis for the definition of improved equations in a revised model. The regional patterns suggest that the structural updates correspond to distinct functions in three sub-regions of the modelling domain (Western Victoria, central region, and Northern New South Wales).

Several avenues for improvement were proposed starting with the incorporation of additional observed data in DAISI (for example actual evapotranspiration) to better constrain internal model variables.

Other proposed improvements include the incorporation of parameter uncertainty and the testing of DAISI for more complex model structures or shorter simulation time steps.

Acknowledgments

This work was conducted on the traditional lands of the Ngunnawal people, and we pay our respects to their leader past, present and emerging. We also acknowledge the traditional owners of the catchments used in this study. This research was supported through funding from the Australian Government Murray–Darling Water and Environment Research Program, the Victorian Water and Climate Initiative and an INRAE-CSIRO linkage project. We acknowledge the support of the Australian Bureau of Meteorology, New South Wales and Victoria state governments in providing the climate and streamflow data for this research study. We also thank James Bennett, CSIRO, and Charles Perrin, INRAE, for their comments on earlier versions of the manuscript.

Open Research

The hydro-climate data described in Section 3.3 is available from Lerat (2023). The software used to run the four steps of the DAISI method can be accessed from Lerat (2023b).

References

Arsenault, R., A. Poulin, P. Côté and F. Brissette (2014). "Comparison of stochastic optimization algorithms in hydrological model calibration." Journal of Hydrologic Engineering **19**(7): 1374-1384.

Beck, M. B. (1985). "Structures, Failure, Inference and Prediction." IFAC Proceedings Volumes **18**(5): 1443-1448.

Beven, K. (2001). "How far can we go in distributed hydrological modelling?" Hydrology and Earth System Sciences **5**(1): 1-12.

Beven, K. and J. Freer (2001). "Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology." Journal of hydrology **249**(1-4): 11-29.

Box, G. E. P. and G. C. Tiao (2011). Bayesian inference in statistical analysis, John Wiley & Sons.

Bulygina, N. and H. Gupta (2009). "Estimating the uncertain mathematical structure of a water balance model via Bayesian data assimilation." Water Resources Research **45**(12): 0-13.

Bulygina, N. and H. Gupta (2011). "Correcting the mathematical structure of a hydrological model via Bayesian data assimilation." Water Resources Research **47**(5).

Bureau of Meteorology. (2019). "Water Data Online." from http://www.bom.gov.au/waterdata.

Charles, S. P., F. H. S. Chiew, N. J. Potter, H. Zheng, G. Fu and L. Zhang (2020). "Impact of downscaled rainfall biases on projected runoff changes." Hydrology and Earth System Sciences **24**(6): 2981-2997.

Chiew, F. H. S., N. J. Potter, J. Vaze, C. Petheram, L. Zhang, J. Teng and D. A. Post (2014). "Observed hydrologic non-stationarity in far south-eastern Australia: implications for modelling and prediction." Stochastic Environmental Research and Risk Assessment **28**: 3-15.

Chiew, F. H. S., J. Vaze, N. R. Viney, P. W. Jordan, J. M. Perraud, L. Zhang, J. Teng, W. J. Young, J. Peña Arancibia and R. A. Morden (2008). "Rainfall-runoff modelling across the Murray-Darling Basin."

Chiew, F. H. S., W. J. Young, W. Cai and J. Teng (2011). "Current drought and future hydroclimate projections in southeast Australia and implications for water resources management." Stochastic Environmental Research and Risk Assessment **25**: 601-612.

Clark, M. P., D. E. Rupp, R. A. Woods, X. Zheng, R. P. Ibbitt, A. G. Slater, J. Schmidt and M. J. Uddstrom (2008). "Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model." Advances in Water Resources **31**(10): 1309-1324.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener and L. E. Hay (2008). "Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models." Water Resources Research **44**(12).

Coron, L., V. Andréassian, C. Perrin, M. Bourqui and F. Hendrickx (2014). "On the lack of robustness of hydrologic models regarding water balance simulation: a diagnostic approach applied to three models of increasing complexity on 20 mountainous catchments." Hydrology and Earth System Sciences **18**(2): 727-746.

Coron, L., V. Andréassian, C. Perrin, J. Lerat, J. Vaze, M. Bourqui and F. Hendrickx (2012). "Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments." Water Resources Research **48**(5).

Cressie, N., C. A. Calder, J. S. Clark, J. M. V. Hoef and C. K. Wikle (2009). "Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling." ECOLOGICAL APPLICATIONS **19**(3): 553-570.

Ditthakit, P., S. Pinthong, N. Salaeh, F. Binnui, L. Khwanchum and Q. B. Pham (2021). "Using machine learning methods for supporting GR2M model in runoff estimation in an ungauged basin." Scientific Reports **11**(1): 19955.

Doucet, A., S. Godsill and C. Andrieu (2000). "On sequential Monte Carlo sampling methods for Bayesian filtering." Statistics and Computing **10**(3): 197-208.

Efstratiadis, A. and D. Koutsoyiannis (2010). "One decade of multi-objective calibration approaches in hydrological modelling: a review." Hydrological Sciences Journal–Journal Des Sciences Hydrologiques **55**(1): 58-78.

Evensen, G. (2009). Data assimilation: the ensemble Kalman filter, Springer.

Evensen, G. and P. J. van Leeuwen (2000). "An Ensemble Kalman Smoother for Nonlinear Dynamics." Monthly Weather Review **128**(6): 1852-1867.

Fenicia, F., D. Kavetski and H. H. G. Savenije (2011). "Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development." Water Resources Research **47**(11).

Fortin, V., M. Abaza, F. Anctil and R. Turcotte (2014). "Why should ensemble spread match the RMSE of the ensemble mean?" Journal of Hydrometeorology **15**(4): 1708-1713.

Fowler, K., W. Knoben, M. Peel, T. Peterson, D. Ryu, M. Saft, K. W. Seo and A. Western (2020). "Many Commonly Used Rainfall-Runoff Models Lack Long, Slow Dynamics: Implications for Runoff Projections." Water Resources Research **56**(5): e2019WR025286-e022019WR025286.

Fowler, K., M. Peel, A. Western and L. Zhang (2018). "Improved rainfall-runoff calibration for drying climate: Choice of objective function." Water Resources Research **54**(5): 3392-3408.

Frost, A. J., A. Ramchurn and A. Smith (2016). "The bureau's operational AWRA landscape (AWRA-L) Model." Bureau of Meteorology Technical Report.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin (2013). Bayesian data analysis, Third Edition, Chapman and Hall/CRC.

Gharari, S., H. V. Gupta, M. P. Clark, M. Hrachowitz, F. Fenicia, P. Matgen and H. H. G. Savenije (2021). "Understanding the Information Content in the Hierarchy of Model Development Decisions: Learning From Data." Water Resources Research **57**(6): e2020WR027948-e022020WR027948.

Ghorbanidehno, H., A. Kokkinaki, J. Lee and E. Darve (2020). "Recent developments in fast and scalable inverse modeling and data assimilation methods in hydrology." Journal of Hydrology **591**: 125266.

Gong, J., A. H. Weerts, C. Yao, Z. Li, Y. Huang, Y. Chen, Y. Chang and P. Huang (2023). "State updating in a distributed hydrological model by ensemble Kalman filtering with error estimation." Journal of Hydrology **620**: 129450.

Gupta, H. V., H. Kling, K. K. Yilmaz and G. F. 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.

Hapuarachchi, H. A. P., M. A. Bari, A. Kabir, M. M. Hasan, F. M. Woldemeskel, N. Gamage, P. D. Sunter, X. S. Zhang, D. E. Robertson and J. C. Bennett (2022). "Development of a national 7-day ensemble streamflow forecasting service for Australia." Hydrology and Earth System Sciences **26**(18): 4801-4821.

Hastie, T., R. Tibshirani, J. H. Friedman and J. H. Friedman (2009). The elements of statistical learning: data mining, inference, and prediction, Springer.

Huang, P., G. Wang, L. Guo, C. R. Mello, K. Li, J. Ma and S. Sun (2023). "Most Global Gauging Stations Present Biased Estimations of Total Catchment Discharge." Geophysical Research Letters **50**(15): e2023GL104253.

Huard, D. and A. Mailhot (2008). "Calibration of hydrological model GR2M using Bayesian uncertainty analysis." Water Resources Research **44**(2).

Kirchner, J. W. (2009). "Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward." Water Resources Research **45**(2): 2429-2429.

Knoben, W. J. M., J. E. Freer, M. C. Peel, K. J. A. Fowler and R. A. Woods (2020). "A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments." Water Resources Research **56**(9): e2019WR025975-e022019WR025975.

Kuczera, G. and E. Parent (1998). "Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm." Journal of Hydrology **211**(1): 69-85.

Lamb, R. and K. Beven (1997). "Using interactive recession curve analysis to specify a general catchment storage model." Hydrology and Earth System Sciences **1**(1): 101-113.

Lawson, C. L. and R. J. Hanson (1974). Solving least squares problems, SIAM.

Lei, F., C. Huang, H. Shen and X. Li (2014). "Improving the estimation of hydrological states in the SWAT model via the ensemble Kalman smoother: Synthetic experiments for the Heihe River Basin in northwest China." Advances in Water Resources **67**: 32-45.

Lerat, J. (2023). Monthly time series of rainfall, potential evapotranspiration and streamflow for 201 catchments in South-East Australia, https://doi.org/10.5281/zenodo.10065059

Lerat, J. (2023b). A python package to run the Data Assimilation Informed model Structure Improvement (PyDAISI), https://doi.org/10.5281/zenodo.10065353

Lerat, J., M. Thyer, D. McInerney, D. Kavetski, F. Woldemeskel, C. Pickett-Heaps, D. Shin and P. Feikema (2020). "A robust approach for calibrating a daily rainfall-runoff model to monthly streamflow data." Journal of Hydrology **591**: 125129.

Li, Y., D. Ryu, A. W. Western, Q. J. Wang, D. E. Robertson and W. T. Crow (2014). "An integrated error parameter estimation and lag-aware data assimilation scheme for real-time flood forecasting." Journal of Hydrology **519**: 2722-2736.

Makhlouf, Z. and C. Michel (1994). "A two-parameter monthly water balance model for French watersheds." Journal of Hydrology **162**(3-4): 299-318.

McInerney, D., M. Thyer, D. Kavetski, J. Lerat and G. Kuczera (2017). "Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors." Water Resources Research **53**(3): 2199-2239.

Moradkhani, H., C. M. DeChant and S. Sorooshian (2012). "Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov chain Monte Carlo method." Water Resources Research **48**(12).

Moradkhani, H., S. Sorooshian, H. V. Gupta and P. R. Houser (2005). "Dual state-parameter estimation of hydrological models using ensemble Kalman filter." Advances in Water Resources **28**(2): 135-147.

Mouelhi, S., C. Michel, C. Perrin and V. Andréassian (2006). "Stepwise development of a two-parameter monthly water balance model." Journal of Hydrology **318**(1-4): 200-214.

Nearing, G. S., F. Kratzert, A. K. Sampson, C. S. Pelissier, D. Klotz, J. M. Frame, C. Prieto and H. V. Gupta (2021). "What Role Does Hydrological Science Play in the Age of Machine Learning?" Water Resources Research **57**(3): e2020WR028091.

Nelder, J. A. and R. Mead (1965). "A simplex method for function minimization." The computer journal **7**(4): 308-313.

Pathiraja, S., L. Marshall, A. Sharma and H. Moradkhani (2016). "Hydrologic modeling in dynamic catchments: A data assimilation approach." Water Resources Research **52**(5): 3350-3372.

Payan, J. L., C. Perrin, V. Andréassian and C. Michel (2008). "How can man-made water reservoirs be accounted for in a lumped rainfall-runoff model?" Water Resources Research **44**(3).

Perrin, C., C. Michel and V. Andréassian (2001). "Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments." Journal of hydrology **242**(3-4): 275-301.

Perrin, C., C. Michel and V. Andréassian (2003). "Improvement of a parsimonious model for streamflow simulation." Journal of hydrology **279**(1-4): 275-289.

Peterson, T. J., M. Saft, M. C. Peel and A. John (2021). "Watersheds may not recover from drought." Science **372**(6543): 745-749.

Pushpalatha, R., C. Perrin, N. Le Moine and V. Andréassian (2012). "A review of efficiency criteria suitable for evaluating low-flow simulations." Journal of Hydrology **420**: 171-182.

Refsgaard, J. C., H. Madsen, V. Andréassian, K. Arnbjerg-Nielsen, T. A. Davidson, M. Drews, D. P. Hamilton, E. Jeppesen, E. Kjellström, J. E. Olesen, T. O. Sonnenborg, D. Trolle, P. Willems and J. H. Christensen (2014). "A framework for testing the ability of models to project climate change and its impacts." Climatic Change **122**(1-2): 271-282.

Royer-Gaspard, P., V. Andreássian and G. Thirel (2021). "Technical note: PMR - A proxy metric to assess hydrological model robustness in a changing climate." Hydrology and Earth System Sciences **25**(11): 5703-5716.

Safeeq, M., R. R. Bart, N. F. Pelak, C. K. Singh, D. N. Dralle, P. Hartsough and J. W. Wagenbrenner (2021). "How realistic are water-balance closure assumptions? A demonstration from the Southern Sierra critical zone observatory and kings river experimental watersheds." Hydrological Processes **35**(5): e14199.

Seo, D. J., L. Cajina, R. Corby and T. Howieson (2009). "Automatic state updating for operational streamflow forecasting via variational data assimilation." Journal of Hydrology **367**(3-4): 255-275.

Thiboult, A. and F. Anctil (2015). "On the difficulty to optimally implement the Ensemble Kalman filter: An experiment based on many hydrological models and catchments." Journal of Hydrology **529**: 1147-1160.

Van Delft, G., G. Y. El Serafy and A. W. Heemink (2009). "The ensemble particle filter (EnPF) in rainfall-runoff models." Stochastic Environmental Research and Risk Assessment **23**: 1203-1211.

Van Esse, W. R., C. Perrin, M. J. Booij, D. C. M. Augustijn, F. Fenicia, D. Kavetski and F. Lobligeois (2013). "The influence of conceptual model structure on model performance: a comparative study for 237 French catchments." Hydrology and Earth System Sciences **17**(10): 4227-4239.

van Leeuwen, P. J. and G. Evensen (1996). "Data Assimilation and Inverse Methods in Terms of a Probabilistic Formulation." Monthly Weather Review **124**(12): 2898-2913.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten and J. M. Verstraten (2005). "Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation." **41**: 1017-1017.

Vrugt, J. A. and C. J. F. Ter Braak (2011). "DREAM (D): an adaptive Markov Chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems." Hydrology and Earth System Sciences **15**(12): 3701-3713.

Wang, Q. J., D. E. Robertson and F. H. S. Chiew (2009). "A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites." Water Resources Research **45**(5).

Wi, S. and S. Steinschneider (2022). "Assessing the Physical Realism of Deep Learning Hydrologic Model Projections Under Climate Change." Water Resources Research **58**(9): e2022WR032123.

Zheng, H., F. H. S. Chiew and L. Zhang (2022). "Can Model Parameterization Accounting for Hydrological Nonstationarity Improve Robustness in Future Runoff Projection?" Journal of Hydrometeorology **23**(11): 1831-1844.

Appendix A: Ensemble Smoother algorithm

The linear Ensemble Smoother (ES, van Leeuwen and Evensen 1996, Evensen 2009) implemented in this paper starts by transforming model and input variables so that their distribution becomes closer to normal following Clark, Rupp et al. (2008). The transformation adopted are the log transform for rainfall and $BC02$ transform for streamflow (similar to what was used in Eq. 19) with other variables left untransformed. Subsequently, ES perturbs observed data, input and state variables to obtain R ensembles for each time step t, noted $\tilde{d}_t[r]$, $\tilde{u}_t[r]$ and $\tilde{x}_t[r]$, respectively, where $r = 1,..,R$. In this paper, independent perturbations are added to transformed data and input vectors as follows (Moradkhani, Sorooshian et al. 2005, Pathiraja, Marshall et al. 2016):

$$\tilde{d}_t[r] = \tilde{d}_t + \tilde{e}_t^d[r]$$ Eq. 21

$$\tilde{u}_t[r] = \tilde{u}_t + \tilde{e}_t^u[r]$$ Eq. 22

Where $\tilde{e}_t^d[r]$ and $\tilde{e}_t^u[r]$ are sampled from multivariate normal distributions. The perturbed observed vectors $\tilde{d}[r]$ are then collated into a matrix D of dimension $OT \times R$. Subsequently, the original model is run using perturbed inputs $\tilde{u}[r]$ as forcings to the state equations (see Eq. 1) combined with. another perturbation to represent the model error. The perturbed states $\tilde{x}_t^f[r]$ (or "forecast" states to follow the data assimilation terminology) are computed as follows:

$$\tilde{x}_{t+1}^f[r] = f\left(\tilde{u}_t^f[r], \tilde{x}_t^f[r], \tilde{\theta}\right) + \tilde{e}_t^x[r]$$ Eq. 23

Where $\tilde{e}_t^x[r]$ is the state error (also referred to as "model" error in data assimilation terminology) sampled from a multivariate normal distribution. Finally, the perturbed ensembles $\tilde{x}^f[r]$ are collated into matrix X^f of dimension $VT \times R$. A subset of this matrix of dimension $OT \times R$, referred to as HX^f, contains the model outputs.

The perturbation scheme presented above has been the subject of a numerous studies (Lei, Huang et al. 2014, Gong, Weerts et al. 2023) with potentially complex parameterization. A pragmatic approach is adopted here by using perturbations with mean 0 and covariance defined similarly for the three vectors $\tilde{e}_t^d[r]$, $\tilde{e}_t^u[r]$ and $\tilde{e}_t^x[r]$ as follows:

$$\Sigma_e^v = \alpha_e^2 \Sigma_v$$ Eq. 24

Where v is either d (observations), u (inputs) or x (state variables), α_e is a scaling factor set to 0.1 and Σ_v is the sample covariance matrix of variable v computed from the original model simulation run over the calibration period. The value chosen for α_e remains subjective and based on values generally reported for the uncertainty in hydrological data (Vrugt, Diks et al. 2005, Seo, Cajina et al. 2009) where an error rate of $\pm 10\%$ is common. Alternative values of α_e have been tested with results reported in supplementary material S3.

The ensemble smoother updates the perturbed ensemble X^f to produce what is referred to as "analysed" states X^a computed as (see Section 9.5 in Evensen 2009):

$$X^a = X^f + K(D - HX^f) \quad \text{Eq. 25}$$

Where K is the Kalman gain matrix defined as

$$K = \Sigma_{XHX}(\Sigma_D + \Sigma_{HX})^{-1} \quad \text{Eq. 26}$$

with Σ_D and Σ_{HX} the sample covariances of the perturbed observations D and model outputs HX, respectively, and Σ_{HXH} the sample covariance between perturbed states and model outputs. These three matrices are computed from ensemble data as

$$E_A = A - \mu_A \, \mathbf{1}_R^T \quad \text{for } A = D, X^f, HX^f \quad \text{Eq. 27}$$

$$\Sigma_D = \frac{E_D E_D^T}{R-1}, \Sigma_{HX} = \frac{E_{HX} E_{HX}^T}{R-1}, \Sigma_{HXH} = \frac{E_{HX} E_X^T}{R-1} \quad \text{Eq. 28}$$

Where μ_A is the column mean of matrix A of dimension $R \times 1$ and $\mathbf{1}_R$ is the unity vector of same dimension.

It is important to note that the updating process of Eq. 25 is only done once as opposed to what is done in the Ensemble Kalman Smoother in which the update is recomputed sequentially for every observation (Evensen and van Leeuwen 2000).

Appendix B: GR2M Model Structure

The GR2M model was introduced by Mouelhi et al. (2006). In this appendix, the reference to a particular time t is dropped to simplify notations. Using the notations introduced in Section 3.1, the model runs as follow (Mouelhi, Michel et al. 2006):

$$S_1 = \frac{\tanh(P/\theta_1)\,\theta_1 + S}{1 + \tanh(P/\theta_1)\dfrac{S}{\theta_1}} \quad \text{Eq. 29}$$

$$S_2 = \frac{S_1(1 - \tanh(E/\theta_1))}{1 + \left(1 - \dfrac{S_1}{\theta_1}\right)\tanh(E/\theta_1)} \quad \text{Eq. 30}$$

$$S^+ = \frac{S_2}{\left(1 + \left(\frac{S_2}{\theta_1}\right)^3\right)^{1/3}}$$ Eq. 31

$$P_e = P + S - S_1 + S_2 - S^+$$ Eq. 32

$$R_2 = \theta_2\,(R + P_e)$$ Eq. 33

$$Q = \frac{R_2^2}{R_2 + \theta_r}$$ Eq. 34

$$R^+ = R_2 - Q$$ Eq. 35

The four state equations listed in Table 1 correspond to equations Eq. 31 (production store), Eq. 32 (effective rainfall), Eq. 34 (routing store) and Eq. 35 (streamflow). Dividing both sides of Eq. 29 by X_1 leads to a form of the production store equation that is independent of θ_1:

$$y_{s_1} = \frac{\tanh(y_p) + y_s}{1 + \tanh(y_p)\,y_s}$$ Eq. 36

Where $y_{s_1} = S_1/\theta_1,\ y_s = S/\theta_1, y_p = P/\theta_1$. The same approach can be applied to equations Eq. 30 to Eq. 32, suggesting that one can obtain transformed state equations for states S^+ and P_e that are independent of θ_1 when introducing the normalized variables $y_s = S/\theta_1,\ y_{s^+} = S^+/\theta_1, y_{p_e} = P_e/\theta_1, y_e = E/\theta_1$. Using such variables leads to the first two transform state equations:

$$y_{s^+} = \frac{y_{s_2}}{\sqrt[3]{1 + {y_{s_2}}^3}}$$ Eq. 37

$$y_{p_e} = y_p + y_s - y_{s_1} + y_{s_2} - y_{s^+}$$ Eq. 38

Where

$$y_{s_2} = \frac{y_{s_1}(1 - \tanh(y_e))}{1 + (1 - y_{s_1})\tanh(y_e)}$$ Eq. 39

Similar approach can be used for equations Eq. 33 to Eq. 35 by introducing $y_r = R\frac{\theta_2}{\theta_r},\ y_{r^+} = \frac{R^+}{\theta_r}, y_{p_e^*} = P_e\frac{\theta_2}{\theta_r}, y_q = \frac{Q}{\theta_r}$. Using these variables, the two states equations Eq. 34 and Eq. 35 become independent from parameter θ_2 and constant θ_r as follows:

$$y_{r^+} = \frac{y_r + y_{p_e^*}}{1 + y_r + y_{p_e^*}} \quad \text{Eq. 40}$$

$$y_q = \frac{\left(y_r + y_{p_e^*}\right)^2}{1 + y_r + y_{p_e^*}} \quad \text{Eq. 41}$$

Overall, Equations Eq. 37, Eq. 38, Eq. 40 and Eq. 41 constitute the four normalized state equations of GR2M.

It is worth noting that the state variables mentioned in Eq. 29 to Eq. 35 do not include actual evapotranspiration and inter-basin exchange (flux gained from or lost to neighboring catchments, see extensive discussion about this flux by Mouelhi, Michel et al. (2006)). The reason for this omission is that the variables listed above are sufficient to describe the model dynamic completely. In the case of the production store for example, once the store level at the start (S) and end (S^+) of the time step are known along with the effective rainfall (P_e), the actual evapotranspiration AE can be computed as a mass balance residual equal to

$$AE = S + P - S^+ - P_e \quad \text{Eq. 42}$$

A similar approach applied to the routing store leads to the computation of the inter-basin exchange F counted positively if water leaves the catchment as

$$F = R + P_e - R^+ - Q \quad \text{Eq. 43}$$

Appendix C: GR2M Updated Model Structure

The updated GR2M model structure operates similarly to the original structure except that update terms are added to states equations as per Eq. 5. Mass balance constraints are also included to avoid non-physical values. In the following equations, the four state functions f_s, f_{p_e}, f_r and f_q represent the right-hand side of equations Eq. 37, Eq. 38, Eq. 40 and Eq. 41, respectively.

Introducing the notation $clip(x_0, x_1, x) = \max(x_0, \min(x_1, x))$ and dropping the reference to a particular time step t like in Appendix B, the updated model structure becomes:

$$\hat{y}_{s^+} = f_s\left(y_s, y_p, y_e\right) + \delta_{s^+} \quad \text{Eq. 44}$$

$$S^+ = clip(0, \min(S + P, \theta_1), \theta_1 \hat{y}_{s^+}) \quad \text{Eq. 45}$$

$$\hat{y}_{p_e} = f_{p_e}\left(y_s, y_p, y_e\right) + \delta_{p_e} \quad \text{Eq. 46}$$

$$P_e = clip(0, S + P - S^+, \theta_1 \hat{y}_{p_e}) \quad \text{Eq. 47}$$

$$\hat{y}^{*}_{p_e} = P_e \frac{\theta_2}{\theta_r} \tag{Eq. 48}$$

$$\hat{y}_{r^+} = f_r\left(y_r, \hat{y}^{*}_{p_e}\right) + \delta_r \tag{Eq. 49}$$

$$R^+ = clip(0, \theta_r, \theta_r \hat{y}_{r^+}) \tag{Eq. 50}$$

$$\hat{y}_q = f_q\left(y_r, \hat{y}^{*}_{p_e}\right) + \delta_q \tag{Eq. 51}$$

$$Q = max(0, \theta_r \hat{y}_q) \tag{Eq. 52}$$

Where δ_n stands for the update term for state variable n computed from Eq. 6. For example, δ_{s^+} is computed as follows in Eq. 44:

$$\begin{aligned}\delta_{s^+} &= \eta_{s,0} + \eta_{s,1}\, y_s + \eta_{s,2}\, y_p + \eta_{s,3}\, y_e + \eta_{s,4} y_s^2 + \eta_{s,5} y_p^2 + \eta_{s,6} y_e^2 + \eta_{s,7}\, y_s y_p \\ &\quad + \eta_{s,8}\, y_s y_e + \eta_{s,9}\, y_p y_e\end{aligned} \tag{Eq. 53}$$

The mass balance constraints introduced in Eq. 44 and Eq. 47 ensure that the store level is bounded within $[0, \theta_1]$, and that the effective rainfall and actual evapotranspiration (see Eq. 43) remain positive. Consequently, the maximum imposed to S^+ in Eq. 45 is the lowest of the store capacity θ_1 and the sum of S with precipitation P. This maximum is reached if actual evapotranspiration and effective rainfall becomes 0. In turn, Eq. 47 ensures that the effective rainfall P_e remains below the sum of the change in store level $(S - S^+)$ with P, which is reached if actual evapotranpiration is 0. The mass balance constraints associated with the routing store are simpler to obtain because GR2M allows for water to leave or enter the catchment via the inter-basin exchange term computed from Eq. 43 (Mouelhi, Michel et al. 2006). Consequently, the only constraints required are that the routing store level is bounded within $[0, \theta_r]$ (Eq. 50), and that simulated streamflow remains positive (Eq. 51).

Additional comments can be made on Eq. 42 and Eq. 43 to better understand the nature of the update terms for S^+ and R^+. Starting with S^+ by combining Eq. 42 with Eq. 45 and Eq. 47 while ignoring mass balance constraints, we obtain:

$$AE = S + P - \theta_1 \hat{y}_{s^+} - \theta_1 \hat{y}_{p_e} \tag{Eq. 54}$$

Combining this equation further with Eq. 44 and Eq. 46 and rearranging leads to

$$AE = S + P - \theta_1 \left(f_s\left(\hat{y}_s, y_p, y_e\right) + f_{p_e}\left(\hat{y}_s, y_p, y_e\right) \right) + \theta_1\left(-\delta_{s^+} - \delta_{p_e}\right) \tag{Eq. 55}$$

In the right-hand side of this equation, all terms except the last one are derived from the GR2M structure while the last term is related to the update terms. Consequently, the opposite of the sum

$\delta_{s^+} + \delta_{p_e}$ can be considered as the update term for actual evapotranspiration. Similar manipulations for the routing store equations lead to

$$F = R + P_e - \theta_r \left(f_r\left(\hat{y}_r, \hat{y}_{p_e}^*\right) + f_q\left(\hat{y}_r, \hat{y}_{p_e}^*\right) \right) + \theta_r\left(-\delta_r - \delta_q\right) \qquad \text{Eq. 56}$$

As a result, the opposite of the sum $\delta_r + \delta_q$ can be considered as the update term for the inter-basin exchange flux. The findings derived from Eq. 54 and Eq. 56 provide a way to relate the four update terms to actual evapotranspiration and inter-basin exchange flux.