Parameter-state ensemble thinning for short-term hydrological prediction

Bruce Davison[1], Vincent Fortin[2], Alain Pietroniro[1], Man K. Yau[3], and Robert Leconte[4]

[1]Environment and Climate Change Canada, Saskatoon, Saskatchewan, Canada.
[2]Environment and Climate Change Canada, Montreal, Quebec, Canada.
[3]McGill University, Montreal, Quebec, Canada.
[4]Université de Sherbrooke, Sherbrooke, Quebec, Canada.

Correspondence: Bruce Davison (bruce.davison@canada.ca)

Abstract. The main sources of uncertainty in hydrological modelling can be summarized as structural errors, parameter errors, and data errors. Operational modellers are generally more concerned with predictive ability than model errors, and this paper presents a new, simple method to improve predictive ability. The method is called Parameter-State Ensemble Thinning (P-SET). P-SET takes a large ensemble of continuous model runs and applies screening criteria to reduce the size of the ensemble. The goal is to find the most promising parameter-state combinations for analysis during the prediction period. Each prediction period begins with the same large ensemble, but the screening criteria is free to select a different sub-set of simulations for each separate prediction period. The case study is from June to October, 2014 for a small (1 324 km^2) watershed just north of Lake Superior in Ontario, Canada using a Canadian semi-distributed hydrologic land-surface scheme. The study examines how well the approach works given various levels of certainty in the data; beginning with certainty in the streamflow and precipitation, followed by uncertainty in the streamflow and certainty in the precipitation, and finally uncertainty in both the streamflow and precipitation. The approach is found to work in this case when streamflow and precipitation is fairly certain, while being more challenging to implement in a forecasting scenario where future streamflow and precipitation is much less certain. The main challenge is determined to be related to parametric uncertainty and ideas for overcoming this challenge are discussed. The approach also highlights model structural errors, which are also discussed.

1 Introduction

A fundamental problem making good streamflow predictions in process-based models rests with the various sources of uncertainty in modelling the flow. These sources of uncertainty have been described in a number of papers (e.g. Beck, 1987; Krzysztofowicz, 2001; Vrugt et al., 2005; Liu and Gupta, 2007; Velázquez et al., 2009). In particular, Liu and Gupta (2007) consider a general framework of seven model components which include the system boundary (B), inputs (u), initial states (x_0), parameters (Θ), structure (M), states (x), and outputs (y). Five of these model components (B, u, x_0, Θ, and M) must be predefined and their uncertainties cascade to x and y. Since the inputs, initial states, and observations used to verify the model outputs can often be considered as data errors, and the system boundary can be considered a source of structural uncertainty, the main sources of errors in hydrologic modelling can be summarized as structural errors, parameter errors, and data errors.

Operational hydrological models are generally more concerned with predictive ability than correctness of the model structure (Gupta et al., 2008, p 3804). As such, parameter and data errors are often the focus for operational hydrological predictions. Within this context of managing parameter and data uncertainty, it is the purpose of this paper to propose a novel approach to short-term hydrological prediction in a relatively small, data-sparse watershed (1,324 km^2). The approach involves using a hydrologic land-surface-scheme (H-LSS) and simultaneous estimation of parameters and state variables by reducing, or "thining", a large ensemble of continuous model simulations.

DA is one way to improve hydrological predictions by merging models with observations, and DA methods can be categorized in a number of different ways (Liu and Gupta, 2007; Rakovec et al., 2015; Sun et al., 2016; Asch et al., 2016). One way of categorizing DA methods is by variational or sequential (statistical) methods. Variational approaches minimize the differences between observations and model output over a period of time, while sequential approaches assimilate observations as they are obtained. Another way of categorizing DA methods is by their time dependence. Usually, smoothing problems attempt to make predictions for the past, filtering problems attempt make predictions for the present, and forecasting problems attempt to make predictions for the future. DA can include methods that help resolve problems related to estimating states, assessing parameters and identifying the appropriate model structure (Liu and Gupta, 2007). Most applications of DA focus on merging state variables in a model with corresponding observations, while a few methods combine state and parameter estimation to improve predictions (e.g. Vrugt et al., 2005; Moradkhani et al., 2005a, b; Drécourt et al., 2006; Labarre et al., 2006; Qin et al., 2009; Nie et al., 2011; Xie and Zhang, 2013; Bi et al., 2014). However, none of these methods ensure that parameters and states are compatible with one-another due to the statistical nature of parameter and state estimation applied in the approaches.

In this paper, a new and very simple method, that ensures parameter and state compatibility, is presented for short-term hydrological ensemble prediction (up to 3 days). We call this approach Parameter-State Ensemble Thinning (P-SET). The approach is described with the intention of making clear how to implement it with a wide variety of models in data-rich or data-sparse watersheds, and examined here using a parameter-intensive, deterministic hydrologic land-surface scheme in a data sparse watershed. The case studies include model structural and parameter errors, which are inevitable regardless of the model being used or the basin being modelled, to evaluate the ability of the P-SET approach.

The intial case studies are hindcasting exercises that reduce data and model structural uncertainty as much as possible, followed by a more realistic forecasting example (albeit in hindcasting mode) that incorporates data input uncertainty using Environment and Climate Change Canada's (ECCC's) meteorological Regional Ensemble Prediction System (REPS) to drive the model.

2 Methodology

2.1 The Parameter-State Ensemble Thinning (P-SET) Approach

P-SET works using a deterministic model in the following manner, as illustrated in Figure 1. First, a number of parameter sets (M) are pre-defined to be used for continuous simulation with the model. Screening criteria are used to determine which of the parameter sets and their associated state variables will be used to generate an ensemble of streamflows for analysis in a

projection period. The analysis is completed for the projection period and the process repeated for the next appropriate time-step in the continuous simulations. Note that the M model simulations continue through the projection period in the continuous simulations. The method simply chooses which of the M continuous simulations to select for the projection analysis. In this manner, due to the fact that a deterministic model is used and the parameters and states are tied to one-another (given the same initial conditions and model structure), both the parameters and states are drawn from the entire M simulations for the projection period.

For a single screen-projection period, this approach is also described by Algorithm 1. This Algorithm has similarities with Approximate Bayesian Computation (ABC, Biau et al., 2015). However, as described by Sadegh and Vrugt (2013), ABC can only be used with a stochastic operator. Sadegh and Vrugt (2013) use a deterministic equivalent of ABC to determine an initial distribution of suitable parameter values. The case study presented in this paper uses Latin Hypercube Sampling (LHS, McKay et al., 1979) to determine an initial distribution of parameter values. The P-SET approach, however, reduces an initial set of parameter values (and their associated state variables) for analysis during a prediction period. Each prediction period begins with the same initial ensemble, but the screening criteria is free to select a different sub-set of simulations for each separate prediction period. The interested reader is directed to the following publications for applications of ABC and its deterministic equivalent in hydrology (Nott et al., 2012; Sadegh and Vrugt, 2013; Vrugt and Sadegh, 2013; Sadegh and Vrugt, 2014; Sadegh et al., 2015; Vrugt and Beven, 2018). The approach also has similarities to Bayesian Recursive Estimation (BaRE, Thiemann et al., 2001) and the parameter identification method based on the localization of information (PIMLI, Vrugt et al., 2002). However, P-SET does not use any stochastic perturbation.

Algorithm 1 Pseudo-code of the P-SET algorithm - adapted from Biau et al. (2015)

Require: A positive integer M and an integer k_M between 1 and M.

for i = 1 to M **do**

generate y_i from the each member (θ_i) of the initial ensemble of parameter sets.

end for

return The θ_i's such that $s(y_i)$ is among the k_M-nearest neighbors of s_0.

In the context of the P-SET approach for hydrological prediction, θ_i is the i^{th} parameter set. In Biau et al. (2015), only the k_M-nearest neighbors between the statistic representing the observations, s_0, and simulations, $s(y_i)$, for the screening period are kept for analysis. The case study presented below alters the selection criteria by using a distance function (Root Mean Squared Error) to determine the discrepancy between the observations and the simulations, rather than independent statistical properties (such as mean and standard deviation) of the two.

In the P-SET approach, we can disregard the notion of finding a distribution of parameter sets that fits the entire streamflow record of interest. Instead, we look for a collection of plausible parameter sets, locate a certain number of these that generate the "best" results for the screening period under consideration, and then evaluate how well these parameters and states perform for the projection period. The process is then repeated to find new parameter sets and states for consideration in successive projection periods. There are a number of ways in which screen and projection periods can be formulated. A number of such

formulations are described in section 2.7, which should clarify the generic process described in this paragraph. It is important to note that if the screen period is a hindcast, the projection period can be analysed for its forecasting potential. Otherwise, the projection period can be analysed to help identify model structural errors.

2.2 Case Study Basin Description

The study watershed is 1,324 km^2 and is drained by the Little Pic River near Coldwell, Ontario, Canada, just north of Lake Superior (Figure 2). The streamflow gauge (02BA003) is between the communities of Terrace Bay and Marathon and has been operated by the Water Survey of Canada from 1972 to the present. There are no precipitation measurements in the basin, but the surrounding region's annual precipitation ranges from 654 to 879 mm, with the mean summer rainfall ranging from 231 to 298 mm (Crins et al., 2009). The mean annual temperature ranges from -1.7 to 2.1 °C (Mackey et al., 1996). The sub-surface sits on Precambrian Shield with significant amounts of volcanic rock, greenstone, siltstone and shale (Sutcliffe, 1991). The dominant landcover in the basin is mixed forest, followed by coniferous forest, water, sparse forest and deciduous forest (Crins et al., 2009). The streamflow regime is characterized by frozen conditions through the winter months (November to April), but has been known to produce a spring freshet as early as March. Summer and autumn peaks can be on the same order of magnitude as the spring freshet, but are more often smaller. The peak flow is usually in May and the highest daily discharge recorded is 269 m^3/s on June 30, 2008.

2.3 The Semi-Distributed Hydrologic Land-Surface Scheme

The model used to simulate the streamflow is the semi-distributed hydrological land-surface scheme MESH (Pietroniro et al., 2007), configured with the Canadian Land-Surface Scheme (CLASS, Verseghy, 1991; Verseghy et al., 1993), the hydrologic routing from WATFLOOD (Kouwen et al., 2002), and additional hydrological processes to better simulate surface and sub-surface lateral flow across the landscape to the river (Soulis et al., 2000, 2011).

The basin geophysical characteristics needed for MESH include a digital elevation model (DEM), landcover classification, and soil information. The DEM comes from the Canadian Digital Elevation Data (CDED) at a scale of 1:50,000 and based on the NAD83 horizontal reference datum (Natural Resources Canada, 2015). The landcover classification comes from the LCC2000-V product originating from classified Landsat 5 and Landsat 7 satellite images and the soils information comes from the ecodistricts classification of the national ecological framework for Canada (Agriculture and Agri-Food Canada, 2015). The basin fits within ecodistrict 389 - Long Lake.

Table 1 shows the estimated percentages of each landcover present in the basin as defined by the LCC2000-V product. Based on this classification, the two dominant landcovers are coniferous and broadleaf forest, which are often mixed. Without knowing more specific information about the landcover, the mixed forests are assumed to be fifty percent coniferous and fifty percent broadleaf, resulting in an estimate of forty-eight percent coniferous and thirty-nine percent broadleaf. These values are then arbitrarily rounded up to fifty percent coniferous and forty percent broadleaf in the model representation of landcover. The remaining ten percent of landcover inevitably includes parametric uncertainty due to the model's inability to properly represent the eight percent of the basin that is covered by small lakes.

Figure 2 illustrates a) the location of the basin, b) ecodistrict boundary and model grid, c) river network and gauge location, and d) landcover. Sub-grid variability of each grid is handled via the CLASS tile with each grid being represented by the same ecodistrict.

2.4 Forcing Data

The meteorological inputs for MESH include incoming shortwave radiation, incoming longwave radiation, precipitation, temperature, barometric pressure, specific humidity and wind speed. The timestep of the model is set to 30 minutes. For the first two case studies minimizing input data uncertainty, most of these meteorological inputs were derived from ECCC's Global Environmental Multi-scale (GEM) Numerical Weather Prediction (NWP) model (Côté et al., 1998a), stitching together the 6–17 hour UTC forecasts from twice-daily runs beginning in January, 2002 and linearly interpolating between hours to obtain half-hourly values. Precipitation is obtained from the Canadian Precipitation Analysis (CaPA, Mahfouf et al., 2007; Lespinas et al., 2015), which is an assimilation of ground-based observations and GEM precipitation forecasts. For the third case study including forcing input data uncertainty, an ensemble of meteorological inputs is obtained from ECCC's Meteorological Regional Ensemble Prediction System (REPS). The REPS provides 72 hour forecasts twice daily.

2.5 Parameter Selection

H-LSS's contain many parameters and there is a large body of scientific literature examining various techniques for effectively estimating parameters (for a brief review, see Matott et al., 2009). The method that was used in this study is Latin Hypercube Sampling (LHS, McKay et al., 1979). Twenty-eight parameters were perturbed based on the results of a simple study (not shown) comparing the perturbation of 6, 15 and 28 parameters. Table 2 shows the parameter values that were fixed during the simulations while Table 3 shows the ranges for parameters that were perturbed. The parameters that were perturbed were based on the lead author's experience with the model. Parameter intervals were set based on the ranges found in sources identified under the source column of Table 3. In the case of user specified parameters, these were set by the lead author.

It is worth noting up-front that this approach to parameter perturbation is very inefficient. Sampling via LHS is a variation of uniform random sampling that is traditionally used in the GLUE methodology (Beven and Binley, 1992). Tolson and Shoemaker (2008) provide a very thorough explanation of the limitations of LHS and other methods of combatting the inefficiency of the traditional GLUE uniform random sampling. The purpose of this study, however, is to examine the P-SET methodology. Implications of the parameter sampling methodology are examined in the discussion after the results are presented.

2.6 Projection Periods for Short-term Hydrological Prediction

This paper is focused on short-term hydrological ensemble prediction (up to 3 days), with an interest in using the ECCC meteorological REPS to force a more comprehensive Hydrological-Ensemble Prediction System (H-EPS). As such, projection periods are defined as the three-day windows of time from the beginning of each ECCC-REPS run at 0 UTC and 12 UTC. The

red and pink bars in Figure 3 illustrate twelve projection periods beginning on 0 UTC, July 20 to 12 UTC, July 25, 2014. The remainder of Figure 3 is described in section 2.7.2.

2.7 Ensemble Selection Methodologies

The total population of model runs is generated by setting M to $10,000$ in Algorithm 1 and using LHS to generate the 10,000 parameter sets from a uniform prior distribution of 28 parameters based on the ranges shown in Table 3. MESH is run with each of the $10,000$ parameter sets in a continuous simulation mode to generate streamflow values (y) for the period of June 2002 to November 2014.

The "thinning" is performed by screening the total population of $10,000$ model runs to generate an ensemble of the "best" model runs to be used in the subsequent projection period. Algorithm 2 represents the specific implementation of Algorithm 1 used in this paper.

Algorithm 2 Pseudo-code of the P-SET algorithm implementation in this paper

Require: A positive integer $10,000$ (M) and an integer (k_M) between 1 and M (In this study, a sensitivity analysis of k_M is performed by setting $k_M = 5, 10, 20, 30, 40,$ and 50).

for i = 1 to $10,000$ **do**

generate streamflow i (y_i) using the model (MESH)

end for

return The parameter sets with the lowest k_M RMSE values between observed and simulated streamflow values from the $10,000$ model runs for the screening period.

The following P-SET configurations are examined in this study and will be described shortly:

1. Optimal hindcast of 3-day projections
2. Preceding streamflow screen
3. Hindsight parameter constraint and preceding 3-day streamflow screen
4. Hydrologic-Ensemble Prediction System (H-EPS)

An initial evaluation of the P-SET screens requires some sources of uncertainty to be minimized. In particular, streamflow observations and precipitation uncertainty are considered; and questions around the model's ability to manage snow processes are simply avoided.

The quality of a streamflow observation is commonly known to be affected by ice, the occurrence of which is noted in the Water Survey of Canada database of streamflow. For the Little Pic River watershed, ice on the river can occur as early as November and as late as April. In addition, prior to the spring freshet, the streamflow contains almost no information that could assist in predicting future streamflow. The future flow primarily depends on other land-surface characteristics such as snow water equivalent and frozen ground. Because the immediate interest is in evaluating the P-SET screens when snow or ice

on the ground is not present, the analysis is only performed from June 1 to October 31 for 2014, with some qualitative analysis beginning on May 1, 2014.

Each of the four P-SET configurations is described below.

2.7.1 Optimal Hindcast of 3-day Projections

Since there are no precipitation observations in the basin, but there are some precipitation gauges nearby which are used in the generation of CaPA, the approach is initially tested with reduced precipitation uncertainty by forcing the model with CaPA.

In addition, streamflow uncertainty is minimized by screening with known streamflow in a hindcasting excercise. In other words, the resulting ensemble is used to determine if the parameter selection methodology (LHS) allows the model to produce streamflow values that match observations given advanced knowledge of precipitation and streamflow. This process is illustrated in Figure 4. In this case, 10,000 simulations are run continuously through the MESH model, of which the approach chooses a number (k_M) for the hindcast analysis. The process is then repeated for subsequent screening periods. Note that in this hindcasting exercise, the screening periods correspond to the projection periods.

2.7.2 Preceding Streamflow Screen

Figure 5 illustrates the preceding streamflow screen. In this study, 10,000 simulations are run continuously through the MESH model, of which the screen chooses a number (k_M) from which to analyze for a projection period. The process is then repeated for subsequent screening periods, noting that the M simulations run continuously through the previous projection periods even though they are not all selected for the previous projection period analysis.

To give more detail to the sequencing within the hindcast-projection cycle, Figure 3 illustrates the preceding streamflow screen considered for a 3-day screening period (other screening period lengths are examined in the results). In this Figure, twelve screen periods are shown in orange and green for July 17 to July 25, 2014. The first screen period is represented by the orange bar near the top of the Figure beside the PSET1 label. This screen period runs from 0 UTC on July 17 to 0 UTC on July 20, 2014. During this three-day period, the "best" parameter sets are selected based on how the model simulates the observed streamflow. The simulations for these top performing parameter sets are then extended for three more days, which is considered to be the projection period.

This process is then repeated 12 hours later. The second screening period is represented by the orange bar illustrated just below the first screening period, and labeled PSET2 in the Figure. This screening period runs from 12 UTC on July 17 to 12 UTC on July 20, 2014. During this three-day period, the "best" parameter sets are selected based on how the model simulates the observed streamflow. Although there is considerable overlap between the first and second screening periods, the second screening period begins and ends 12 hours after the first screening period, producing a new ensemble of "best" parameter sets. The simulations for these new parameter sets are then extended for three days, which is considered to be a new projection period that also begins and ends 12 hours after the previous projection period.

The process is then continually repeated every 12 hours as shown by the remainder of the bars shown in the rows labeled PSET3 to PSET12. Each instance of the screen and projection periods represents a single application of P-SET, which is why

each row is labeled as such. We call this approach the preceding streamflow screen because the projection periods shown in red and pink occur immediately after the screening periods.

One important consideration, that becomes relevant in the analysis, is the six hours of precipitation that occurred on July 22 from 14 UTC to 20 UTC. This is illustrated by the small blue bar at the top and near the middle of Figure 3. Some of the projection periods "see" this precipitation event (illustrated by the green bars) and some do not (illustrated by the orange bars). As will be explained more fully in section 2.8.2, the analysis is split according to the sub-periods of the screen and projection periods that a) occur during and just after the precipitation event (the light green and pink bars); and b) that occur when it is otherwise rain-free (orange, dark green and red bars).

To properly examine the effectiveness of the P-SET approach, a sensitivity analysis is performed for the length of the screen period as well as for k_M in Algorithms 1 and 2. The length of the screen period is tested for 3, 10, 20, 30 and 40 days while k_M is tested for values of 5, 10, 20, 30, 40 and 50.

2.7.3 Hindsight Parameter Constraint and Preceding 3-Day Streamflow Screen

The third ensemble considered is very similar to the preceding streamflow screen with a screening period of 3 days. However, to constrain the parameter space further than determined by LHS, the population of parameter sets is reduced by selecting a sub-set from the initial population of 10,000 parameter sets. The sub-set is selected by confining the parameter values based on model simulations that respond well during precipitation events in 2014. Figure 6 illustrates this screen and more details are provided in section 3.3 of the results. This ensemble represents an approach that cannot be used in a forecasting context, but does represent a proxy for other parameter-constraining methods that are explored in the discussion and helps to examine model structural errors.

2.7.4 Hydrologic-Ensemble Prediction System (H-EPS)

Of course it is not often known for sure if precipitation will occur in the future, and certainly not the amount of precipitation that will occur. As a result, ECCC's Meteorological REPS is also used with the June 1 to October 31, 2014 data in a hindcasting mode to examine how the P-SET approach can be used in a true forecasting context. At 00 UTC and 12 UTC every 12 hours from May 1 to October 31, 2014, the top 10 parameter-state pairs from the screen period and different projection periods are run for 3-days using the forcing data from the 20 members of the REPS, for a total of 200 H-EPS members. This analysis is performed for a hind-sight parameter constraint and preceding 3-day streamflow screen and k_M value of 10 for illustrative purposes.

The only other use of ECCC's REPS as a part of an H-EPS is found in Abaza et al. (2013), in which the Canadian operational meteorological Global Ensemble Prediction System (GEPS) was compared with the REPS and the deterministic 15 km GEM NWP forcing of the province of Quebec's operational streamflow forecasting system. The study found that both the GEPS and REPS outperformed a deterministic run for eight watersheds ranging in size from 355 to 5820 km^2. The REPS was also found to be superior to the GEPS in terms of its ability to predict forecast uncertainty.

One issue highlighted in the conclusion of Abaza et al. (2013) is that the REPS was found to produce unusually high precipitation spikes. This issue of excessive precipitation was, in many cases, determined to be caused by the physics perturbation scheme that was used to generate the ensemble (Erfani et al., 2014) and was fixed in the version of the REPS that was officially released on December 4, 2013. The update to the REPS is one of the main reasons for focusing on 2014 as a period of interest.

2.8 Verification

Demargne et al. (2010) differentiates diagnostic verification for evaluating the performance of a system from real-time verification for helping end-users make decisions about the future. The verification performed here is done for the first of these objectives; evaluating the performance of a system.

2.8.1 Verification of the Ensemble Selection Methodologies

First, a qualitative analysis is undertaken to take advantage of the human brain's ability to synthesize information. The results are then quantitatively verified using the Ensemble Verification System (EVS, Brown et al., 2010). To examine the quality of the ensemble mean when compared with the corresponding observation, the mean error (ME) is calculated. Then the quality of the ensemble distribution is calculated using rank histograms. Finally, the skill relative to using the current streamflow as the forecast is calculated using the mean Continuous Ranked Probability Skill Score ($\overline{\text{CRPSS}}$). The reference forecast in this study is taken to be the measured streamflow at 00 UTC and 12 UTC each day as the forecast for the next 72 hours. This reference forecast is a persistence forecast, which assumes the streamflow is persistent for the forecast period.

The ME measures the average difference between a set of forecasts and corresponding observations. In this case, it measures the average difference between the mean average of the ensemble forecast ($\overline{Y}$) and the observation (x) as follows:

$$\text{ME} = \frac{1}{n}\sum_{i=1}^{n}(\overline{Y}_i - x_i)$$

The ME may be positive, zero, or negative. A positive value represents an ensemble mean that is positively biased while a negative error represents an ensemble mean that is negatively biased. A value of zero represents an absence of bias in the ensemble mean.

The rank histogram measures the reliability of an ensemble forecasting system. It involves counting the fraction of observations that fall between any two ranked ensemble members in the forecast distribution. For an ensemble forecast containing m ensemble members ranked in ascending order, there are $m-1$ spaces between any two ranked ensemble members and two spaces at the ends (above and below the ensemble forecast range) for a total of $m+1$ spaces ($s_1, \ldots, s_{m+1}$). The corresponding observation h for each ensemble forecast will fall within one of the spaces.

$$h_i = \frac{1}{n}\sum_{j=1}^{n} 1\{x_j \in s_{ij}\}$$

where h_i is the fraction in the i^{th} bin, x_j is the j^{th} observed value, s_{ij} is the i^{th} gap associated with the j^{th} forecast, and $1\{\cdot\}$ is a step function that gives a value of 1 if the condition is met and 0 otherwise.

The mean Continuous Ranked Probability Skill Score ($\overline{\text{CRPSS}}$) measures the performance of one forecasting system compared to another forecasting system in terms of the mean Continuous Ranked Probability Score ($\overline{\text{CRPS}}$). The $\overline{\text{CRPS}}$ measures the average square eror of a probability forecast across all possible event thresholds. The $\overline{\text{CRPSS}}$ comprises a ratio of the $\overline{\text{CRPS}}$ for the forecasting system to be evaluated $\overline{\text{CRPS}}_{\text{EVAL}}$, and the $\overline{\text{CRPS}}$ for a reference forecasting system, $\overline{\text{CRPS}}_{\text{REF}}$.

$$\overline{\text{CRPSS}} = \frac{\overline{\text{CRPS}}_{\text{REF}} - \overline{\text{CRPS}}_{\text{EVAL}}}{\overline{\text{CRPS}}_{\text{REF}}}$$

As a measure of the average square error in probability, values for the $\overline{\text{CRPS}}$ approaching zero are better. As a result, values for the $\overline{\text{CRPSS}}$ closer to one are better as this illustrates that $\overline{\text{CRPS}}_{\text{EVAL}} < \overline{\text{CRPS}}_{\text{REF}}$.

2.8.2 Splitting the Analysis Based on Rainfall

The qualitative analysis shown in the following Results section illustrates that there is a significant difference in the abilities of the P-SET configurations to effectively project streamflow when it is raining and when it is not raining. As a result, the quantitative analysis is split into two parts: 1) for periods in which it is raining and just afterwards (for the remainder of each respective screen or projection period), and 2) for periods in which it is otherwise not raining.

Note that these periods do not necessarily correspond to the rising-limb and recession periods of the hydrograph since the river does not always respond strongly to the precipitation for the time period of study in this basin. As a result, for lack of better terminology, these periods are hereafter referred to as "rain-influenced" and "rain-free". It would be more correct to say "periods during and immediately after the rainfall within the 3-day period" and "otherwise rain-free," but this terminology would be cumbersome throughout the remainder of the paper. Furthermore, it is also important to note that the terms "rain-influenced" and "rain-free" only refer to a time period rather than the discharge of the river. The time periods that these terms refer to are the stretch of time under consideration in the analysis.

Recall the description for Figure 3 in Section 2.7.2. for illustrating the difference between "rainfall" and "rain-free". The screen periods are rain-free for the July 22 rain event in PSET1 through to PSET12 as shown by the orange and dark green bars, while the screen periods are rain-influenced in PSET7 through to PSET12 as shown in the light green bars. Similarly, the projection periods are rain-influenced in PSET1 to PSET6 as shown by the pink bars and rain-free in PSET1 to PSET12 as shown in the red bars. In both cases (screen and projection), the rain-influenced period is considered to be from the beginning of the rainfall (July 22, 9 EST in Figure 3) to the end of the corresponding screen or projection period that "sees" the precipitation.

Recall that MESH is run in a continuous simulation mode for the period of June 2002 to November 2014, with a detailed analysis of the ensemble selection methodologies from June 1 to October 31, 2014. Within this time period, there are five significant precipitation events. The beginning and ending of the precipitation events are considered as follows:

- July 22, 14 UTC (9 EST) to 20 UTC (15 EST)

- August 11, 14 UTC (9 EST) to 20 UTC (15 EST)
- September 10, 08 UTC (3 EST) to September 11, 02 UTC (September 10, 21 EST)
- September 19, 20 UTC (15 EST) to September 20, 20 UTC (15 EST)
- October 3, 08 UTC (3 EST) to 20 UTC (15 EST)

For the rain-influenced and rain-free periods, the quality of the ensemble mean, distribution and skill are compared. The June 3 rain-influenced period is not assessed due to the fact that it was not a projection period in the preceding streamflow screen.

2.8.3 Verification of the H-EPS

As with the earlier analysis when precipitation uncertainty is minimized, the mean error and CRPSS are calculated for streamflow for both rain-influenced and rain-free periods as determined by precipitation events in the basin. The overall mean error and CRPSS is also calculated.

The mean error, rank histograms and CRPS of the REPS precipitation ensemble mean are calculated using CaPA as the observation. The CRPSS is also calculated using the June to October CaPA "climatology" as the reference forecast. The mean error, CRPS and CRPSS are calculated above and below the ninetieth percentile of CaPA precipitation (0.42 mm/hr).

3 Results

The results are ordered according to the ensemble selection methodologies. In all cases, the projection time period of interest is short-term, which is defined here as 3 days.

3.1 Optimal Hindcast of 3-day Projections

In determining the effectiveness of the P-SET approach, it is necessary to see if the method has the possibility of succeeding with prior knowledge of streamflow and precipitation data. For this purpose, the method is applied for June 1 to October 31, 2014 and compared with hourly streamflow observations. Figure 7a shows precipitation from CaPA. Figure 7b shows the observed streamflow (black) and corresponding optimal model runs (red). Figure 7c shows the corresponding basin-average water storage state variables for each of the parameter-state pairs chosen by the optimal hindcasts. These storage results will be discussed later.

For this study, the qualitative results in Figure 7b) illustrates that CaPA precipitation cascades to reasonable streamflow values for the time period examined. A quantitative analysis compares these results to the other ensemble selection methodologies, but a qualitative analysis is first performed for each of the remaining approaches.

3.2 Preceding Streamflow Screen

One manner in which to screen the parameter sets (and associated states) is to consider only the preceding streamflow. In this study, the best RMSE values from the preceding days of streamflow are used to determine the parameter sets to use for the prediction of the subsequent 3 days of streamflow. This process is repeated twice daily at 0 UTC and 12 UTC (19 and 5 local time for the basin in question) for June 1 to October 31, 2014.

Figures 7d, e and f show the overall results. Qualitatively, the screen produces good results when there is negligible precipitation. However, the results degrade when it rains, particularly for the 3-day screen. To illustrate this aspect of the screen in more detail, Figures 8a and 8b show how the 3-day screen reacts for a single rain event on July 22, 2014. In Figure 8a, which is the equivalent of PSET6 in Figure 3, the screening period does not "see" the rain and the projected streamflows resulting from the screened parameter sets overestimate the actual streamflow. In Figure 8b (equivalent to PSET7 in Figure 3), the rain event occurs during the screen period and the subsequent projected streamflows are much more closely aligned with the observations. This result is consistent with all significant precipitation events, with the screen choosing parameter sets that overestimate streamflow when the precipitation event is not "seen" by the screen. However, Figure 7e and f show that the impact of not seeing the precipitation event is reduced with a longer screen period.

3.3 Hindsight Parameter Constraint and Preceding 3-Day Streamflow Screen

The third screen explored here is one in which the top simulations are selected based on the preceding streamflow and parameter ranges that are proven to be important during the 2014 precipitation events. Figure 9 shows parameters that are particularly sensitive during six precipitation events. Each parameter is normalized between 0 and 1. Based on a subjective visual analysis of these box-plots, the 10,000 parameter sets are reduced to 91 parameter sets by confining the values of the normalized parameters as follows: KS1 < 0.1, WF_R2 > 0.6, CLAY11 > 0.5, CLAY12 > 0.5, SDEP1 > 0.2. Using the preceding streamflow screen with these 91 parameter sets to obtain the top 10 runs for each 3-day period yields Figure 10. This approach of identifying parameter ranges based on periods of hydrological significance is similar to the DYNIA approach described by Wagener et al. (2003), albiet much less rigorous. With the exception of the June 3 precipitation event, these results are clearly much better than those found in unconstrained 3-day screen shown in Figure 7c. Although this method clearly cannot be used in a forecasting context, the significance of these findings are examined in the discussion.

3.4 A Quantitative Comparison of Screens

Table 4 shows the mean error of the ensemble mean for the previously defined rain-free and rain-influenced periods for the reference forecast, the optimal hindcast, the preceding streamflow screen (with various lengths of time for the screen period), and the 3-day screen with constrained parameters. All of the methods, including the reference forecast, provide reasonable results for the rain-free periods. For the rain-influenced periods, which are the real periods of concern for this study, the optimal hindcast is capable of finding parameter sets that have a low mean error. The 3-day screen performs the worst in terms of overpredicting streamflow in rain-influenced periods, with results improving as the length of the screen period increases.

The 3-day screen with constrained parameters performs close to the optimal hindcast with only a slight over-prediction of the observed flows.

The number of "top" runs selected (k_M) does not appear to have much influence over these mean error results. As a consequence the remainder of the analysis if performed with a value of $k_M = 10$.

Using the current streamflow as the reference forecast. Table 5 shows the skill of the optimal hindcast, 20-day screen and 3-day screen with constrained parameters. The optimal hindcast exhibits a relatively high skill for rain-free and rain-influenced periods. For the rain-free periods, the 20-day screen shows some skill for the 48 and 72 hour forecast, while the 3-day screen with constrained parameters shows no skill. For the rain-influenced periods, the only screen that shows any skill is the 3-day projection with constrained parameters. These results quantify the qualitative analysis shown in Figures 7 and 10.

3.5 H-EPS

To address the question of how this data assimilation approach could be used in a forecasting context, a full H-EPS is used to force selected parameter-state ensemble members with ECCC's Meteorological Regional Ensemble Prediction System (REPS), as described in the methodology section. Two sets of parameter-state ensembles are selected to see how the REPS performs. The ensembles are based on 1) the optimal hindcast of 3-day projections and 2) the hindsight parameter constraint and preceding 3-day streamflow screen. These ensembles were selected because they were the only methods that showed any skill in the rain-influenced periods. Of course, neither of these ensembles can be used in operational forecasting, so they are used as a proxy for illustrative purposes assuming that the limitations of the preceding streamflow screen can be addressed as explored in the discussion. Although these two ensembles are "unforecastable," performing this analysis provides a more meaningful mechanism to examine model structural and forcing errors.

Figure 11a) shows CaPA (reddish brown) and the 20 REPS precipitation members (blue). The resulting 200 streamflow ensembles (recall that $k_M = 10$) are shown in Figure 11b, with the black line representing the observed streamflow, the orange lines coming from the hindsight constrained parameter and preceding 3-day streamflow screen, and the green lines coming from the optimized hindcast. Even for the optimal hindcast, which shows near-perfect alignment with the observed streamflow when forced with GEM and CaPA, the REPS members that overestimate the precipitation have an impact on the resulting ensemble of streamflows.

Table 6 shows the mean error for streamflow and Table 7 shows the CRPSS, for both rain-free and rain-influenced periods. The overall mean error and CRPSS are also calculated.

The mean error results show that the H-EPS ensemble mean overestimates streamflow in all cases. The CRPSS scores show that the H-EPS fails to show skill during key time periods for many of the ensembles when compared to using the current streamflow as the forecasted streamflow. This lack of skill will be considered in the discussion. To examine these findings with respect to the precipitation; the mean error, rank histograms and CRPS of the REPS precipitation ensemble mean are calculated using CaPA as the observation. The CRPSS is also calculated using the June to October CaPA "climatology" as the reference forecast. The mean error, CRPS and CRPSS shown in Table 8 are calculated above and below the ninetieth percentile of CaPA precipitation (0.42 mm/hr). Below this threshold, the REPS mean precipitation over-estimates the CaPA precipitation.

In the top 10 percent of CaPA precipitation values, however, the REPS mean under-estimates the CaPA precipitation. The rank histograms (not shown) indicate that the ensemble members tend to underestimate precipitation, although some REPS members do over-estimate the higher CaPA precipitation values. The CRPS shows the highest (worst) values, and the CRPSS shows the least skill, for the highest precipitation rates.

4 Discussion

The discussion is organized around three questions. The first question looks at whether-or-not the P-SET approach is capable of reproducing observed streamflow, which corresponds to the optimized hindcast. The second question considers the effectiveness of the remaining screening approaches. The third question revolves around the more realistic example of using the approach in a full H-EPS. Finally, advantages and limitations of the approach are discussed.

4.1 Given maximum data certainty, can the P-SET approach reproduce observed streamflow?

Although the P-SET approach could be applied to a hydrological model with few parameters, the Canadian MESH model is used with many parameters perturbed, which increases the dimensionality of the problem. Although much simpler models tend to dominate the operational hydrological modelling community, part of the motivation behind using a hydrologically-enhanced land-surface scheme in the case study is to begin laying some foundation for using such parameter-intensive models for operational ensemble hydrological forecasting.

One major limitation to the way in which MESH is applied in this study is the use of the relatively inefficient Latin Hypercube Sampling to determine the prior distribution of parameter sets to be used with the P-SET approach. Despite this limitation, however, the results clearly show that the approach can, with confidence in the precipitation forcing and streamflow, find parameter-state sets that match the observed hydrograph over successive periods of a few days. One possible way of dealing with the uncertainty in precipitation is to perturb the CaPA precipitation field as is examined by Carrera et al. (2015).

The widely varying nature of the simulated basin storage for the selected runs for each 3-day period also highlights a limitation with the study. This limitation is in only using streamflow as the state variable to determine the top parameters each time. Consider the following water balance equation for the basin: $P - E = R + \mathrm{d}S/\mathrm{d}t$, where P is precipitation, E is evapotranspiration, R is runoff and $\mathrm{d}S/\mathrm{d}t$ is the change in basin storage over time. Over the short time-periods of a few days in short-term hydrological prediction, E can generally be ignored, leaving only $P = R + \mathrm{d}S/\mathrm{d}t$. In the hindcasting exercise presented in this part of the study, P and R are considered to be known and the only remaining term is $\mathrm{d}S/\mathrm{d}t$. So why does the analysis show such a wide range of basin storage terms for the best matching assimilated streamflow? The answer lies in the fact that it is not the basin storage that balances the equation, but rather the change in storage over the time period of interest. The model is capable of releasing or storing the appropriate amount of water in both rain-influenced and rain-free scenarios, and the model determines $\mathrm{d}S/\mathrm{d}t$ based on the interaction of existing storage, model physics and parameters.

The issue of widely-varying simulated basin storage (Figure 7c) also highlights the issue of equifinality, which is defined here as the idea that many different model simulations can produce acceptable results (Beven, 1993). The model is able to find

many parameter-state sets that fit the streamflow for short periods of time. If only streamflow observations are available, the selected simulations are equifinal. However, including the state of basin storage clearly shows that the parameter-state sets are not equal. If soil moisture observations are also available and used, then the selected simulations could be further constrained. Of course, including soil moisture observations to further constrain the selection of simulations would not remove equifinality. It would simply make it more likely that the model is more accurately predicting both streamflow and soil moisture.

One assumption in most environmental modelling exercises is that the parameters do not vary with time, or at least they vary slowly or if the system is disturbed in some way such as land-use change (Bard, 1974; Wagener et al., 2003; Liu and Gupta, 2007). Wagener et al. (2003) indicate that the inability of a single parameter set to simulate an entire streamflow record provides evidence of model structural error. It is incorrect to assume that MESH has a perfect model structure, so the results indicate that any model structural errors can be compensated for by the parameter sets. One can also presume that data errors can also be hidden by the selection of certain parameter sets. Clearly the model needs further constraints to give the results a more solid foundation. One of these constraints could be multi-objective calibration, or else further "thining", based on some aspects of storage in the model. One such possibility would be to examine the usefulness of the soil moisture and ocean salinity (SMOS) satellite (Mecklenburg et al., 2012; Jackson et al., 2012; Ridler et al., 2014).

Related to model structural error, the unresponsive streamflow in this study is likely due to "fill-and-spill" dynamics (Spence, 2010). Being on the Precambrian Shield, and the starting point of many streams in the basin being small lakes, there are many parts of the basin that need to be filled-up before they contribute to streamflow. This physical process, especially with respect to the headwater lakes, is not represented in the version of MESH used in this study. Future work should focus on this aspect more closely.

4.2 How well do different P-SET configurations work?

The issues of parameter time-invariance and the most appropriate model structures are generally secondary considerations in forecasting. The focus shifts from the exercise of improving the model and its parameterization to the exercise of making a more accurate prediction. The results presented from the various configurations tested, however, indicate that some thought is required to determine the appropriate parameter sets at the appropriate times.

The only configurations in this study that show any skill in predicting streamflow when it rains are a) the optimal hindcast of 3-day projections, and b) the hindsight parameter constraint and preceding 3-day streamflow screen. The manner in which the second of these approaches is applied in this study reveals that 91 of the original 10,000 LHS parameter sets can be used effectively with the P-SET approach to perform short-term predictions in the basin for the months of July to October, 2014. The reduction of 10,000 parameter sets to 91 parameter sets is noteable.

The fact that constraining the parameter sets allows for the approach to produce reasonable results throughout the period provides some assurance that the method has the possibility of being able to predict streamflow with some skill in a forecasting context. The key, at least in part, is expected to be in using a method other than LHS to determine the initial ensemble of parameter sets. Alternative approaches could use algorithms such as Dynamically Dimensioned Search - Approximation of Uncertainty (DDS-AU) which have been shown to be more efficient than GLUE (Tolson and Shoemaker, 2008). The prior

can also be obtained by looking for parameter sets that perform well for different hydrological signatures (e.g. Zhang et al., 2014; Shafii and Tolson, 2015) or different hydrological scenarios which might include streamflow responses to snow-melt, runoff over frozen ground, rain during wet conditions, rain during dry conditions, or whatever else can be considered a relevant hydrological event affecting streamflow.

As shown in this study, increasing the length of the screening period has a positive impact on the scores. The gains in mean error values do not improve after 20 day screening periods, indicating that there is a limit to the value of longer screen periods. In this study, the 20 day screening period allows the method to see the previous precipitation event in all cases examined. As such, the ability of the screen to capture important hydrological responses is critical to improving results. The downside to having a longer screen period, however, is that the forecaster must wait longer to apply the approach. This screen-period time limitation for the forecaster may not be true for basins where snow, ice and frozen ground are not dominant processes. We expect that different basins will have different optimal screen period lengths depending on the important hydrological processes in the basin.

If given more information about the state of the basin (other than streamflow), different hydrological scenarios could also be used in determining the appropriate parameter-state sets to screen. For example, if the SMOS satellite indicates that the basin is dry, the streamflow observation is relatively low and a certain amount of precipitation is expected in the near future, then past scenarios that fit this description could be used to screen the parameter sets. As a result, parameter sets that fit both the current state of the basin as well as the expected forcing could be screened, if both the current basin state and expected precipitation has been previously experienced and observations are available.

Such an approach is very similar to the well-established k-nearest neighbor (k-nn) bootstrap method as described by Lall and Sharma (1996). In its simplest form, the k-nn approach finds k similar patterns in the past data and uses this information to make a prediction about the next data point. The P-SET preceding streamflow screen essentially does the same thing, except that it looks for similar patterns in an ensemble of model runs rather than in a time series of data points. By including criteria beyond streamflow as suggested in the previous paragraph, one could (for example) look for past parameter sets that successfully simulated the streamflow when the basin exhibited a certain threshold of upper-layer soil moisture from SMOS, a given streamflow, and a specified amount of precipitation. This approach requires a relatively long time series of observational data with model simulations and could provide an interesting comparison between the model-centric P-SET approach and purely data-driven analogue methods.

4.3 How can this approach be used in a forecasting context (including precipitation uncertainty)?

The mean error results for the H-EPS ensemble mean streamflow, forced with the REPS (Table 6), are similar in nature to the mean ensemble streamflow forced with GEM and CaPA (Table 4). However, an important finding is drawn from the CRPSS scores in Table 7. Overall, the 3-day screen with constrained parameters does not show skill, with the exception being hour 72 for the rain-influenced periods.

These findings illustrate that the H-EPS contains too much uncertainty to be used with any skill for this particular study. It is important to note that the same lack of skill may not be true for other time periods or different basins. For this particular study,

it is not surprising that the REPS does not show any skill when compared to using the current streamflow as the forecast. For this basin and the time period considered, the streamflow is not very responsive to the precipitation input for much of the time. Situations when the river is not responsive to precipitation favor the approach of using the current streamflow as the forecast.

A resulting question is whether or not the lack of skill in the H-EPS is due to the uncertainty in the REPS precipitation, or the unresponsive behaviour of the streamflow to precipitation during this period. Looking more closely at the REPS precipitation mean error compared to CaPA (Table 8) indicates that the REPS tends to overestimate the bottom 90 percent, and underestimate the top 10 percent, of CaPA values, which are taken to be as close to observed as is possible in the basin. The only noticeable trend in time is that the underestimation in the top 10 percent of precipitation becomes more pronounced with time.

Returning to the question of whether-or-not the lack of skill in the H-EPS is due to the uncertainty in the REPS precipitation, or the unresponsive behaviour of the streamflow to precipitation during this period, it seems that both factors contribute to the overall lack of skill. As Figure 11, shows, however, relatively small differences in precipitation result in large changes to streamflow, indicating that the land-surface physical processes (e.g. fill-and-spill) that determine the responsiveness of the streamflow to precipitation, are probably the more important of the two for this particular study.

4.4 Advantages and Challenges of the Approach

One key benefit of the P-SET approach is that it is conceptually straight-forward. In plain language, the idea is to setup a series of continuous simulations and draw the most appropriate runs from these simulations for making a projection or forecast. This concept is very easy to understand and implement. In an operational forecasting environment, this simplicity is desirable.

Another advantage is that the parameters and state variables are always consistent with one another. This cannot be said for other approaches such as the dual Particle Filter or dual Ensemble Kalman Filter.

As the examples provided in this study have shown, the approach is also flexible. It can be used in the more traditional manner of hydrologic model calibration by selecting multi-year screen periods (Davison et al., 2017), which has not been shown here, or in other unique ways that have been examined and discussed throughout the paper. It can be seen as a more general approach to model calibration.

Two challenges with the approach are 1) how to determine the initial ensemble of parameter sets to run in continuous simulations, and 2) how to select the most appropriate runs for making a projection or forecast. This study uses a parameter-intensive H-LSS and deals with the first challenge by using LHS to determine the initial ensemble, and deals with the second challenge by comparing different screen period lengths to select the appropriate runs. There are likely better ways of dealing with these challenges than have been explored here, and one proxy method (the hindsight parameter constrined and preceding 3-day streamflow screen) has been explored in lieu of these other potential methods.

Fortunately, there is an exhaustive body of research and a number of existing tools that can be used to overcome these challenges. Possible solutions to determine a better initial ensemble of parameters include: 1) selecting parameter sets based on more than just streamflow, 2) selecting parameter sets based on different hydrological signatures or aspects of the streamflow 3) using k-nn type approach of looking for parameter sets that worked in similar circumstances in the past, 4) using more efficient algorithms than LHS. Any or all of these methods can be used together to improve the determination of the initial

ensemble. In terms of selecting the most effective simulations once the initial ensemble has been established, one method that can be explored is to use more than streamflow to select the top runs with the P-SET method. The length of the screening period is also a consideration that needs further exploration.

For both determining a better initial ensemble and selecting the most effective simulations once the initial ensemble has been established, remote sensing offers such opportunities to gather information on the watershed state (e.g. soil moisture, snow) that can complement the limited information that streamflow provides. This approach would better constrain the model in the parameter and state estimation process. Using different hydrological signatures, or segmenting the hydrographs for different parameters (e.g. groundwater parameters during low flows), are also ideas worth exploring.

The effectiveness of these methods requires further study.

5 Conclusions

The main contribution of this work is the introduction of a new method (P-SET). The method always returns to an initial ensemble of simulations and removes the need to resample the parameter space between each model run. The weighting of each simulation from the original ensemble is then determined by assigning each simulation a value of zero or one based on a screening criteria.

In this study, one approach is to use P-SET for the preceding days of streamflow every 12 hours (preceding streamflow screen). It is shown that increasing the length of time for the screening period generally improves the results, up to a point (in this study example, 20 days). A second approach is the same as the first approach with a parameter-constrained subset of the original 10,000 runs (preceding 3-day streamflow screen with parameter constraints). The parameter constraints are determined from an analysis of the results during rain-influenced periods.

The optimal hindcast results clearly show that the model and LHS method of sampling 10,000 prior parameter sets is capable of simulating the streamflow for any three-day period where the precipitation input is reasonable. The methods tested to select the most appropriate runs, however, show that making a projection is more complicated. The only method that consistently shows reasonable projections in this work is the preceding streamflow screen with parameter constraints. The problem with this approach is that it is not immediately clear how such a screen can be used in a forecasting context. Something more is needed to provide better parameter estimates if the P-SET method is to be useful in an operational forecasting setting. Fortunately, there are a number of approaches that can be explored to provide superior guidance on the parameters, either in pre-determining the prior or in selecting the most appropriate runs from the prior.

In addition to introducing P-SET, a fuller H-EPS is presented that includes forcing uncertainty from ECCC's REPS. For this particular basin and time-period, the resulting H-EPS is shown overall to be less skillful than using the current streamflow as the forecast for the future streamflow, likely due to model structural errors in MESH. This result is not generally applicable as one should expect the current streamflow to be a fairly good indicator of future streamflow when the stream is relatively unresponsive to precipitation inputs, as is the case in this study. It is expected that the REPS precipitation in an H-EPS would

exhibit more skill in more responsive basins without the same fill-and-spill physical processes or for more responsive time periods in this basin.

Acknowledgements. We gratefully acknowledge Environment and Climate Change Canada, the Natural Sciences and Engineering Research Council (NSERC) and Hydro Quebec for funding this study. We also thank Ethan Johnson for his editorial support and the thesis committee for their comments on an earlier draft of this paper that is a part of the first author's PhD dissertation. We also thank Dr. Ousmane Seidou for his helpful discussion on the topic , and Dr. Gérard Biau and Dr. Christian Robert for their input. Finally, we thank Dr. Jasper Vrugt, 2 anonymous reviewers and the journal editor for their helpful comments.

References

Abaza, M., Anctil, F., Fortin, V., and Turcotte, R.: A comparison of the Canadian global and regional meteorological ensemble prediction systems for short-term hydrological forecasting, Monthly Weather Review, 141, 3462–3476, 2013.

Agriculture and Agri-Food Canada: National Ecological Framework, digital media. [Available online at http://sis.agr.gc.ca/cansis/nsdb/ecostrat/index.html.], 2015.

Asch, M., Bocquet, M., and Nodet, M.: Data assimilation: methods, algorithms, and applications, vol. 11, SIAM, 2016.

Bard, Y.: Nonlinear parameter estimation, vol. 513, Academic Press New York, 1974.

Beck, M. B.: Water quality modeling: a review of the analysis of uncertainty, Water Resources Research, 23, 1393–1442, 1987.

Beven, K.: Prophecy, reality and uncertainty in distributed hydrological modelling, Advances in water resources, 16, 41–51, 1993.

Beven, K. and Binley, A.: The future of distributed models: model calibration and uncertainty prediction, Hydrological Processes, 6, 227–246, 1992.

Bi, H., Ma, J., Qin, S., and Zhang, H.: Simultaneous estimation of soil moisture and hydraulic parameters using residual resampling particle filter, Science China Earth Sciences, 57, 824–838, 2014.

Biau, G., Cérou, F., and Guyader, A.: New insights into Approximate Bayesian Computation, Annales de l'Institut Henri Poincaré, Probabilités et Statistiques, 51, 376–403, 2015.

Brown, J. D., Demargne, J., Seo, D.-J., and Liu, Y.: The Ensemble Verification System (EVS): A software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations, Environmental Modelling & Software, 25, 854–872, 2010.

Burnash, R. J., Ferral, R. L., and McGuire, R. A.: A generalized streamflow simulation system, conceptual modeling for digital computers, 1973.

Carrera, M. L., Bélair, S., and Bilodeau, B.: The Canadian Land Data Assimilation System (CaLDAS): Description and Synthetic Evaluation Study, Journal of Hydrometeorology, 16, 1293–1314, 2015.

Côté, J., Gravel, S., Méthot, A., Patoine, A., Roch, M., and Staniforth, A.: The operational CMC-MRB global environmental multiscale (GEM) model. Part I: Design considerations and formulation, Mon. Wea. Rev., 126, 1373–1395, 1998a.

Crins, W. J., Gray, P. A., Uhligh, P. W., and Wester, M. C.: The Ecosystems of Ontario, Part 1: Ecozones and Ecoregions, Tech. Rep. SIB TER IMA TR-01, Ontario Ministry of Natural Resources, Peterborough, Ontario, 2009.

Davison, B., Fortin, V., Pietroniro, A., Yau, M. K., and Leconte, R.: Parameter-state ensemble data assimilation using Approximate Bayesian Computing for short-term hydrological prediction, Hydrology and Earth System Sciences Discussions, 2017, 1–38, https://doi.org/10.5194/hess-2017-482, https://www.hydrol-earth-syst-sci-discuss.net/hess-2017-482/, 2017.

Demargne, J., Brown, J., Liu, Y., Seo, D.-J., Wu, L., Toth, Z., and Zhu, Y.: Diagnostic verification of hydrometeorological and hydrologic ensembles, Atmospheric Science Letters, 11, 114–122, 2010.

Dingman, S. L.: Physical Hydrology, Prentice-Hall Inc., 2 edn., 2002.

Drécourt, J.-P., Madsen, H., and Rosbjerg, D.: Calibration framework for a Kalman filter applied to a groundwater model, Advances in Water Resources, 29, 719–734, 2006.

Erfani, A., Frenette, R., Gagnon, N., Charron, M., Beauregard, S., Giguère, A., and Parent, A.: The New Regional Ensemble prediction System (REPS) at 15 km horizontal grid spacing (from version 1.1.0 to 2.0.1), Tech. re., Meteorological Research Branch, National Predictions Development, and National Operations Divistions at the Canadian Meteorological Center, Environment Canada, 2014.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrological Processes, 22, 3802–3813, 2008.

Jackson, T. J., Bindlish, R., Cosh, M. H., Zhao, T., Starks, P. J., Bosch, D. D., Seyfried, M., Moran, M. S., Goodrich, D. C., Kerr, Y. H., et al.: Validation of soil moisture and ocean salinity (SMOS) soil moisture over watershed networks in the US, Geoscience and Remote Sensing, IEEE Transactions on, 50, 1530–1543, 2012.

Kouwen, N., Mousavi, S., Singh, V., Frevert, D., et al.: WATFLOOD/SPL9 hydrological model & flood forecasting system., Mathematical models of large watershed hydrology, pp. 649–685, 2002.

Krzysztofowicz, R.: The case for probabilistic forecasting in hydrology, Journal of Hydrology, 249, 2–9, 2001.

Labarre, D., Grivel, E., Berthoumieu, Y., Todini, E., and Najim, M.: Consistent estimation of autoregressive parameters from noisy observations based on two interacting Kalman filters, Signal Processing, 86, 2863–2876, 2006.

Lall, U. and Sharma, A.: A nearest neighbor bootstrap for resampling hydrologic time series, Water Resources Research, 32, 679–693, 1996.

Lespinas, F., Fortin, V., Roy, G., Rasmussen, P., and Stadnyk, T.: Performance Evaluation of the Canadian Precipitation Analysis (CaPA), Journal of Hydrometeorology, 16, 2045–2064, 2015.

Liu, Y. and Gupta, H. V.: Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, Water Resources Research, 43, n/a–n/a, https://doi.org/10.1029/2006WR005756, 2007.

Mackey, B. G., McKenney, D. W., Yang, Y.-Q., McMahon, J. P., and Hutchinson, M. F.: Erratum: Site regions revisited: a climatic analysis of Hills' site regions for the province of Ontario using a parametric method, Canadian Journal of Forest Research, 26, 1112, 1996.

Mahfouf, J.-F., Brasnett, B., and Gagnon, S.: A Canadian precipitation analysis (CaPA) project: Description and preliminary results, Atmos.–Ocean, 45, 1–17, 2007.

Matott, L. S., Babendreier, J. E., and Purucker, S. T.: Evaluating uncertainty in integrated environmental models: A review of concepts and tools, Water Resources Research, 45, 2009.

McKay, M. D., Beckman, R. J., and Conover, W. J.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics, 21, 239–245, 1979.

Mecklenburg, S., Drusch, M., Kerr, Y. H., Font, J., Martin-Neira, M., Delwart, S., Buenadicha, G., Reul, N., Daganzo-Eusebio, E., Oliva, R., et al.: ESA's soil moisture and ocean salinity mission: Mission performance and operations, Geoscience and Remote Sensing, IEEE Transactions on, 50, 1354–1366, 2012.

Moradkhani, H., Hsu, K.-L., Gupta, H., and Sorooshian, S.: Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, Water Resources Research, 41, n/a–n/a, https://doi.org/10.1029/2004WR003604, 2005a.

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

Natural Resources Canada: Land Cover, circa 2000 - Vector, digital media. [Available online at http://wmsmir.cits.rncan.gc.ca/index.html/pub/geobase/official/lcc2000v_csc2000v/doc/Land_Cover.pdf.], 2015.

Nie, S., Zhu, J., and Luo, Y.: Simultaneous estimation of land surface scheme states and parameters using the ensemble Kalman filter: identical twin experiments, Hydrology and Earth System Sciences, 15, 2437–2457, 2011.

Nott, D. J., Marshall, L., and Brown, J.: Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection?, Water Resources Research, 48, 2012.

Pietroniro, A., Fortin, V., Kouwen, N., Neal, C., Turcotte, R., Davison, B., Verseghy, D., Soulis, E., Caldwell, R., Evora, N., et al.: Development of the MESH modelling system for hydrological ensemble forecasting of the Laurentian Great Lakes at the regional scale, Hydrology and Earth System Sciences, 11, 1279–1294, 2007.

Qin, J., Liang, S., Yang, K., Kaihotsu, I., Liu, R., and Koike, T.: Simultaneous estimation of both soil moisture and model parameters using particle filtering method through the assimilation of microwave signal, Journal of Geophysical Research: Atmospheres (1984–2012), 114, n/a–n/a, https://doi.org/10.1029/2008JD011358, 2009.

Rakovec, O., Weerts, A., Sumihar, J., and Uijlenhoet, R.: Operational aspects of asynchronous filtering for flood forecasting, Hydrology and Earth System Sciences, 19, 2911, 2015.

Ridler, M.-E., Madsen, H., Stisen, S., Bircher, S., and Fensholt, R.: Assimilation of SMOS-derived soil moisture in a fully integrated hydrological and soil-vegetation-atmosphere transfer model in Western Denmark, Water Resources Research, 50, 8962–8981, 2014.

Sadegh, M. and Vrugt, J.: Bridging the gap between GLUE and formal statistical approaches: approximate Bayesian computation, Hydrology and Earth System Sciences, 17, 4831–4850, 2013.

Sadegh, M. and Vrugt, J. A.: Approximate bayesian computation using markov chain monte carlo simulation: Dream (abc), Water Resources Research, 50, 6767–6787, 2014.

Sadegh, M., Vrugt, J. A., Xu, C., and Volpi, E.: The stationarity paradigm revisited: Hypothesis testing using diagnostics, summary metrics, and DREAM (ABC), Water Resources Research, 51, 9207–9231, 2015.

Shafii, M. and Tolson, B. A.: Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives, Water Resources Research, pp. 3796–3814, https://doi.org/10.1002/2014WR016520, 2015.

Soulis, E., Snelgrove, K., Kouwen, N., Seglenieks, F., and Verseghy, D.: Towards closing the vertical water balance in Canadian atmospheric models: coupling of the land surface scheme CLASS with the distributed hydrological model WATFLOOD, Atmosphere-Ocean, 38, 251–269, 2000.

Soulis, E., Craig, J., Fortin, V., and Liu, G.: A simple expression for the bulk field capacity of a sloping soil horizon, Hydrological Processes, 25, 112–116, 2011.

Spence, C.: A paradigm shift in hydrology: Storage thresholds across scales influence catchment runoff generation, Geography Compass, 4, 819–833, 2010.

Sun, L., Seidou, O., Nistor, I., and Liu, K.: Review of the Kalman-type hydrological data assimilation, Hydrological Sciences Journal, 61, 2348–2366, 2016.

Sutcliffe, R.: Proterozoic geology of the Lake Superior area, in: Geology of Ontario, edited by Thurston, P., Williams, H., RH, S., and Stott, G., vol. 4 Part 1, pp. 627–658, Ontario Geological Survey, 1991.

Thiemann, M., Trosset, M., Gupta, H., and Sorooshian, S.: Bayesian recursive parameter estimation for hydrologic models, Water Resources Research, 37, 2521–2535, 2001.

Tolson, B. A. and Shoemaker, C. A.: Efficient prediction uncertainty approximation in the calibration of environmental simulation models, Water Resources Research, 44, n/a–n/a, https://doi.org/10.1029/2007WR005869, 2008.

Velázquez, J. A., Petit, T., Lavoie, A., Boucher, M.-A., Turcotte, R., Fortin, V., and Anctil, F.: An evaluation of the Canadian global meteorological ensemble prediction system for short-term hydrological forecasting, Hydrology and Earth System Sciences, 13, 2221–2231, https://doi.org/10.5194/hess-13-2221-2009, http://www.hydrol-earth-syst-sci.net/13/2221/2009/, 2009.

Verseghy, D.: CLASS—The Canadian land surface scheme (Version 3.5). Technical Documentation (Version 1), 2011.

Verseghy, D., McFarlane, N., and Lazare, M.: CLASS—A Canadian land surface scheme for GCMs, II. Vegetation model and coupled runs, International Journal of Climatology, 13, 347–370, 1993.

Verseghy, D. L.: CLASS—A Canadian land surface scheme for GCMs. I. Soil model, International Journal of Climatology, 11, 111–133, 1991.

Vrugt, J. A. and Beven, K. J.: Embracing equifinality with efficiency: Limits of Acceptability sampling using the DREAM (LOA) algorithm, Journal of Hydrology, 559, 954–971, 2018.

Vrugt, J. A. and Sadegh, M.: Toward diagnostic model calibration and evaluation: Approximate Bayesian computation, Water Resources Research, 49, 4335–4345, 2013.

Vrugt, J. A., Bouten, W., Gupta, H. V., and Sorooshian, S.: Toward improved identifiability of hydrologic model parameters: The information content of experimental data, Water Resources Research, 38, 2002.

Vrugt, J. A., Diks, C. G., Gupta, H. V., Bouten, W., and Verstraten, J. M.: Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, Water Resources Research, 41, n/a–n/a, https://doi.org/10.1029/2004WR003059, 2005.

Wagener, T., McIntyre, N., Lees, M., Wheater, H., and Gupta, H.: Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis, Hydrological Processes, 17, 455–476, https://doi.org/10.1002/hyp.1135, 2003.

Xie, X. and Zhang, D.: A partitioned update scheme for state-parameter estimation of distributed hydrologic models based on the ensemble Kalman filter, Water Resources Research, 49, 7350–7365, 2013.

Zhang, Y., Vaze, J., Chiew, F. H., Teng, J., and Li, M.: Predicting hydrological signatures in ungauged catchments using spatial interpolation, index model, and rainfall–runoff modelling, Journal of Hydrology, 517, 936 – 948, 2014.

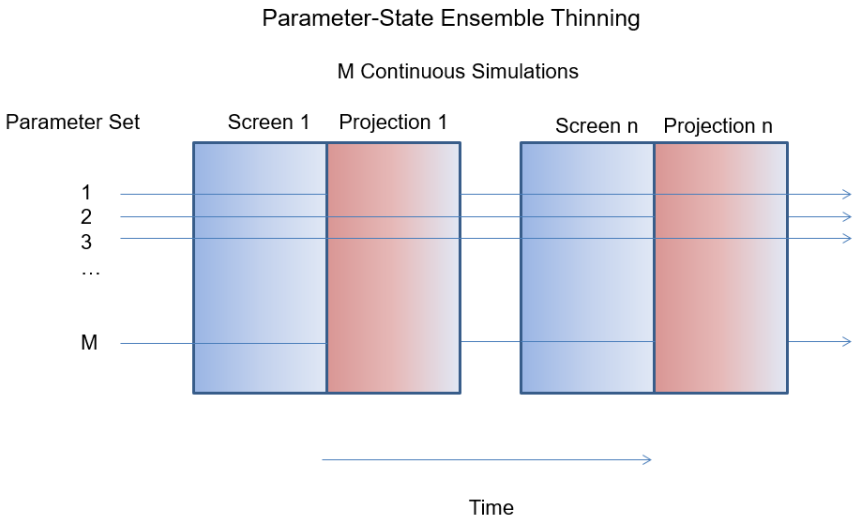

Figure 1. Schematic of the P-SET approach. M simulations are run continuously from a model, of which the screen chooses a number from which to analyze a projection. The process is then repeated for subsequent screening periods, noting that the M simulations run continuously through the previous projection periods even though they are not all selected for the previous projection period analysis.

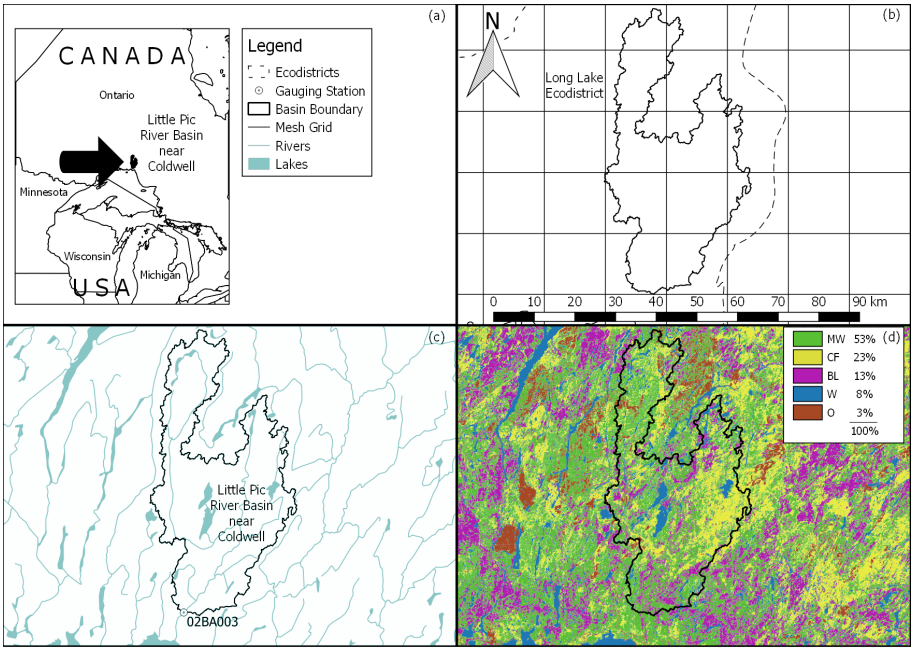

Figure 2. Little Pic River basin near Coldwell, Ontario, Canada. a) Location of the basin and legend, b) basin outline with respect to ecodistrict, c) river network and gauge location (02BA003), and d) landcover (MW is Mixed Wood, CF is Coniferous Forest, BL is Broadleaf Forest, W is Water, and O is other).

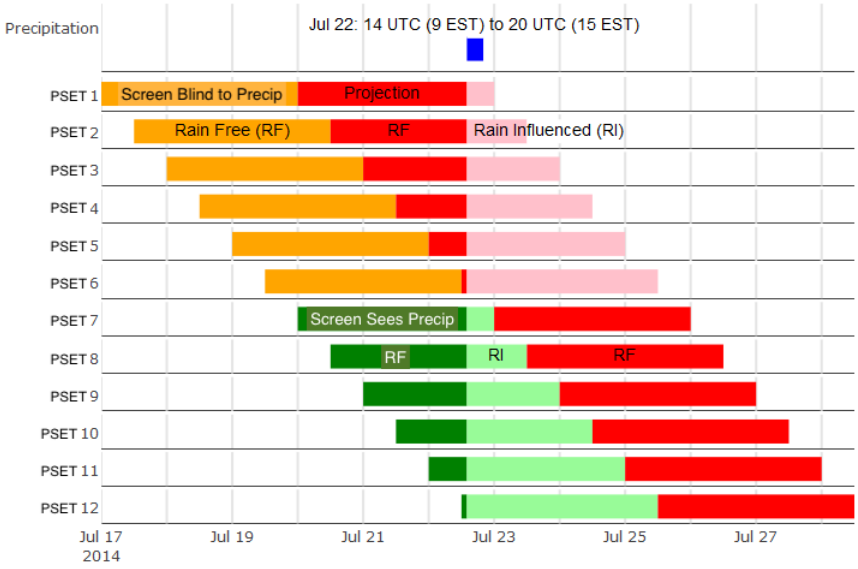

Figure 3. Preceding Streamflow Screen and Projection Periods for July 17 to July 28, 2014 using Hourly Streamflow. This Figure is fully explained in sections 2.7.2 and 2.8.2.

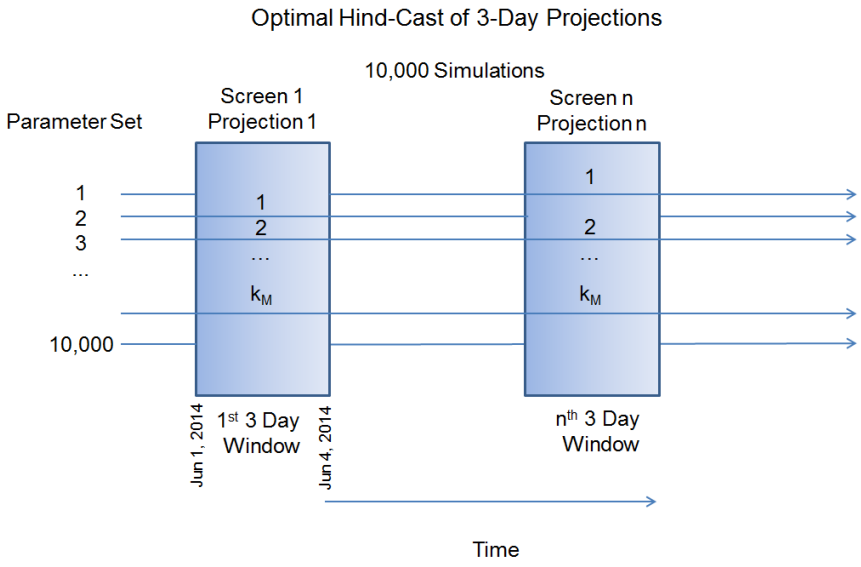

Figure 4. Schematic of P-SET for the optimal hindcast of 3-day projections used in this study. 10,000 simulations are run continuously through the MESH model, of which the screen chooses a number (k_M) for the hindcast analysis. The process is then repeated for subsequent screening periods.

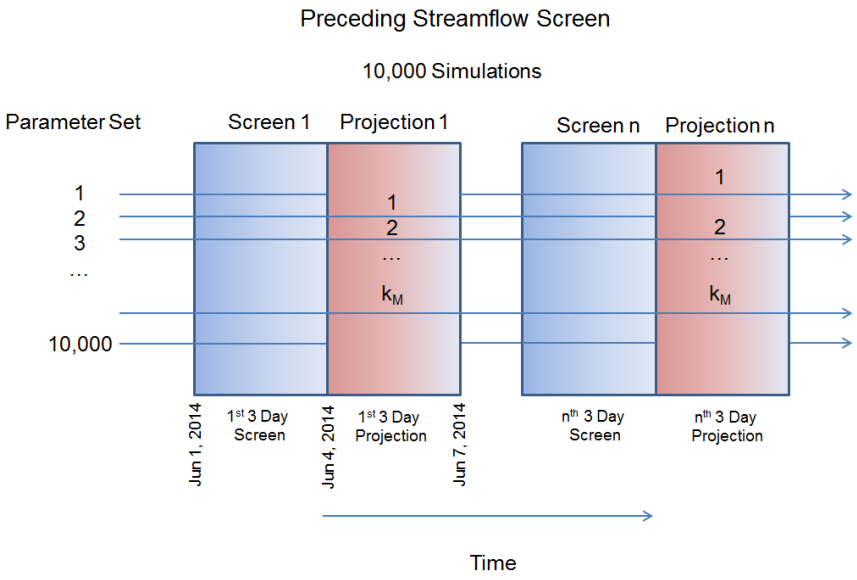

Figure 5. Schematic of P-SET for the preceding streamflow screen. 10,000 simulations are run continuously through the MESH model, of which the screen chooses a number (k_M) from which to analyze a projection. The process is then repeated for subsequent screening periods, noting that the M simulations run continuously through the previous projection periods even though they are not all selected for the previous projection period analysis.

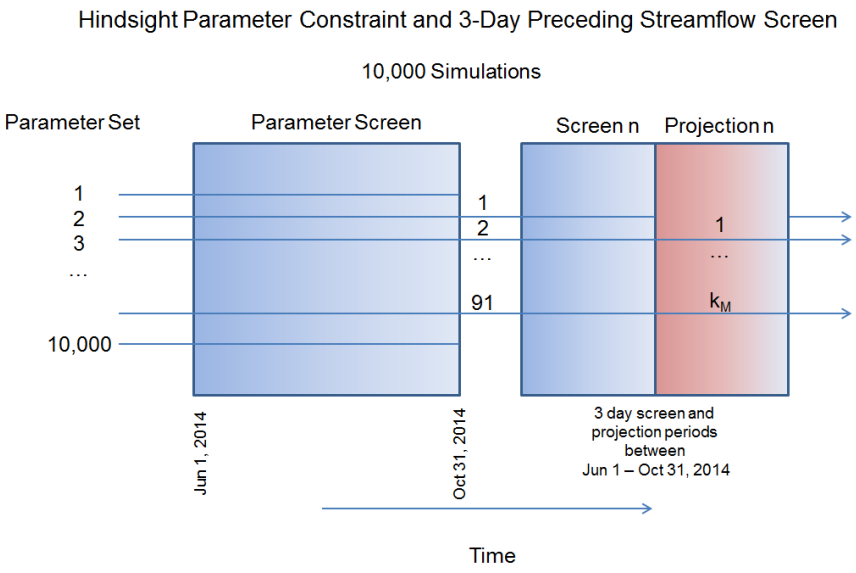

Figure 6. Schematic of P-SET for the hindsight parameter constraint and preceding 3-day streamflow screen used in this study. 10,000 simulations are run continuously through the MESH model. The original 10,000 parameter sets are reduced to 91 parameter sets based on a hindsight analysis of parameters that are shown to be important during precipitation events between June 1 and October 31, 2014. These remaining 91 parameter sets are then selected for analysis in the preceding streamflow screen as if $M = 91$ in Algorithm 1.

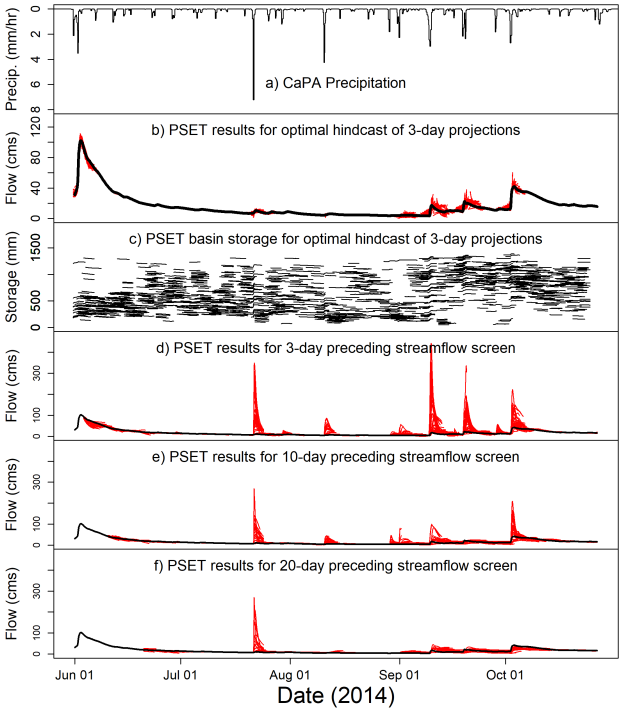

Figure 7. CaPA precipitation (a) for all simulations shown in this Figure. (b) Observed streamflow (black) and top 10 streamflow values (red) for the optimal hindcast of 3-day projections. (c) Corresponding basin-wide storage values for the optimal hindcast of 3-day projections. (d) Top 10 preceding streamflow projections for each of the 3-day screen periods. (e) Top 10 preceding streamflow projections for each of the 10-day screen periods. (f) Top 10 preceding streamflow projections for each of the 20-day screen periods. The black lines in (d), (e) and (f) show observed streamflow with different y-axis scaling than in (b).

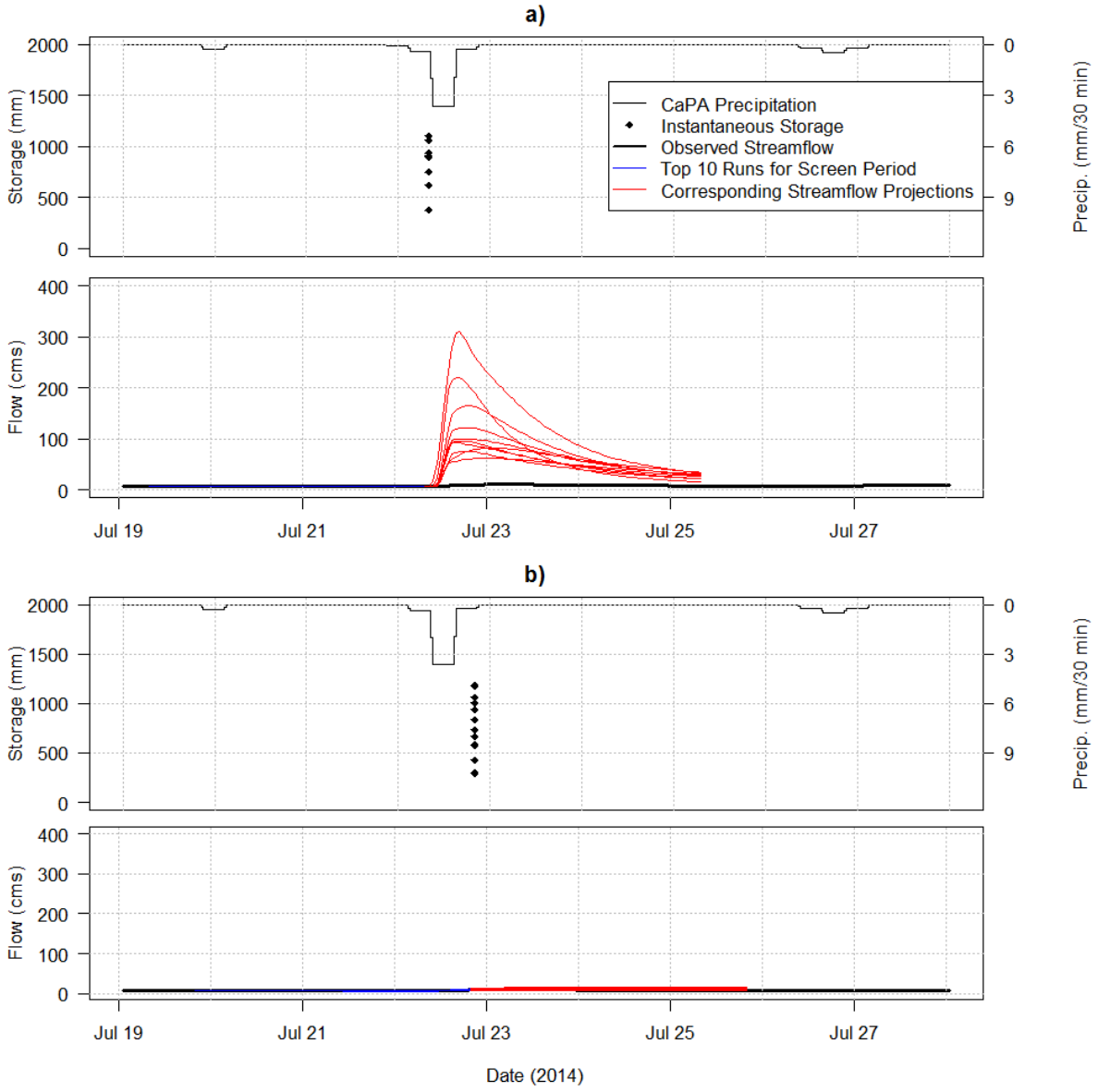

Figure 8. A single screen-projection period for two neighboring time periods. For the two sub-plots in a) the projection begins at 7:00 local time, July 22, 2014 (12 UTC). For the two sub-plots in b) the projection begins at 19:00 local time, July 22, 2014 (0 UTC, July 23). The upper plot of each sub-figure shows CaPA precipitation and instantaneous storage. The lower plot shows observed streamflow (black), the top 10 runs for the screening period (blue), and the corresponding streamflow projections (red).

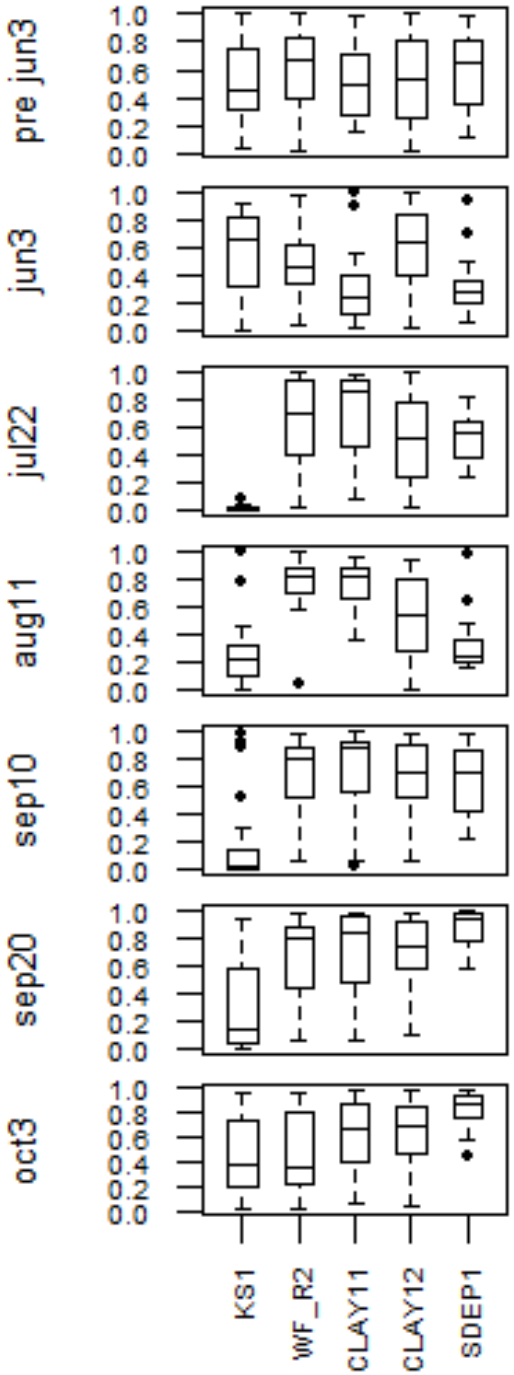

Figure 9. Importance of (normalized) parameters that see precipitation. The top set of box-plots shows that none of the top parameter sets have identifiable parameter values prior to the June 3, 2014 precipitation event. This result is similar to all parameter sets immediately prior to precipitation events that do not see the events. The remainder sets of box-plots show the parameter ranges for the top simulations during precipitation events.

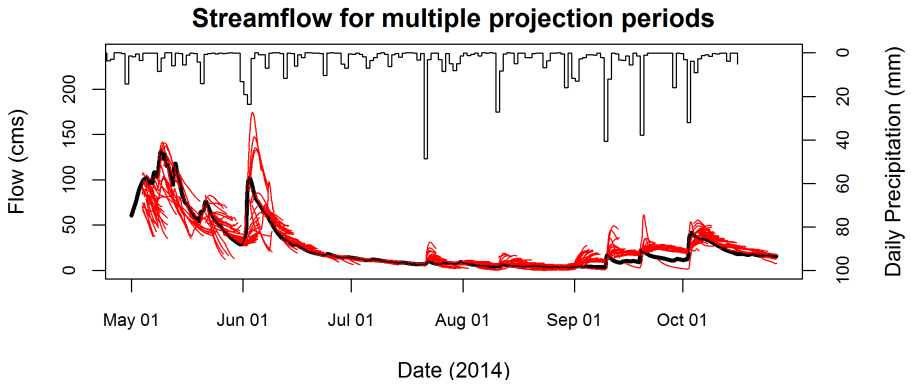

Figure 10. Projection period results after screening based on parameter values and preceding streamflow.

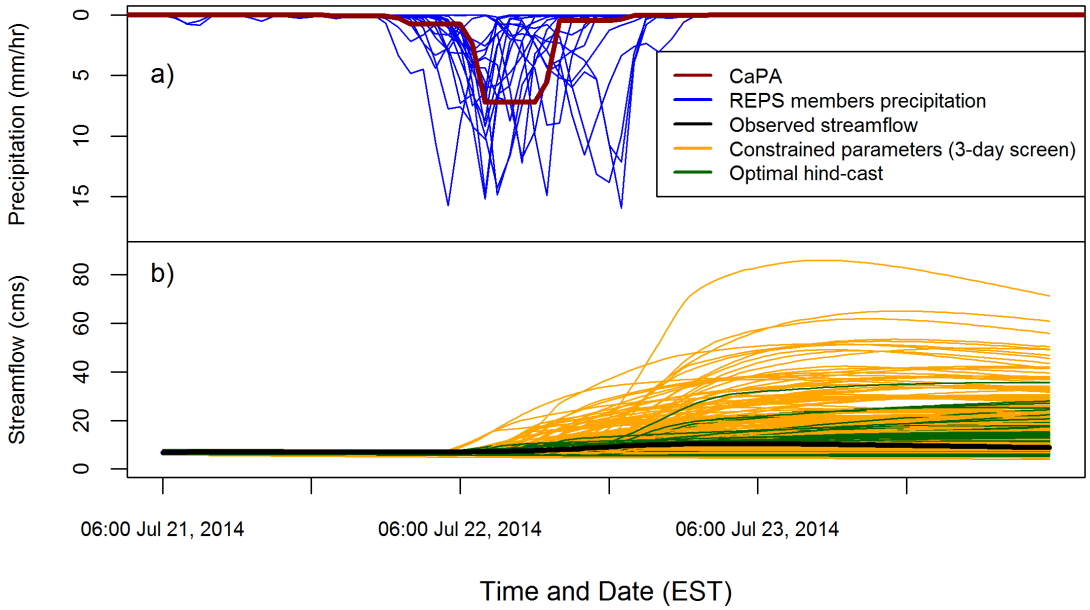

Figure 11. Results of the H-EPS for the 3-day period beginning at 6 Eastern Standard Time (EST) on July 21, 2014. The reddish brown line in sub-figure a) is the Canadian Precipitation Analysis (CaPA) while the blue lines represent the 20 Regional Ensemble Prediction System (REPS) precipitation traces. The single black line in sub-figures b) is the observed streamflow. The orange and green lines show the 200 H-EPS streamflow traces for the projection periods of the constrained parameter with a preceding 3-day streamflow screen, and the optimized hindcast.

Table 1. Landcover percentages based on LCC2000-V Landsat product.

Landcover	Percentage
Water	8
Coniferous Dense Forest	23
Broadleaf Dense Forest	13
Mixed Wood	53
Other	3

Table 2. Fixed landcover parameters.

Parameter Name	Description	Units	Value	Source
QA50 - NL	Reference value of incoming shortwave radiation used in stomatal resistance formula (Needleleaf)	[W m^{-2}]	30	Verseghy (2011)
QA50 - BL	Reference value of incoming shortwave radiation used in stomatal resistance formula (Broadleaf)	[W m^{-2}]	40	Verseghy (2011)
VPDA - NL	Vapour pressure deficit coefficient used in stomatal resistance formula (Needleleaf)	[ ]	0.65	Verseghy (2011)
VPDA - BL	Vapour pressure deficit coefficient used in stomatal resistance formula (Broadleaf)	[ ]	0.5	Verseghy (2011)
VPDB - NL	Vapour pressure deficit coefficient used in stomatal resistance formula (Needleleaf)	[ ]	1.05	Verseghy (2011)
VPDB - BL	Vapour pressure deficit coefficient used in stomatal resistance formula (Broadleaf)	[ ]	0.6	Verseghy (2011)
PSGA - NL	Soil moisture suction coefficient used in stomatal resistance formula (Needleleaf)	[ ]	100	Verseghy (2011)
PSGA - BL	Soil moisture suction coefficient used in stomatal resistance formula (Broadleaf)	[ ]	100	Verseghy (2011)
PSGB - NL	Soil moisture suction coefficient used in stomatal resistance formula (Needleleaf)	[ ]	5	Verseghy (2011)
PSGB - BL	Soil moisture suction coefficient used in stomatal resistance formula (Broadleaf)	[ ]	5	Verseghy (2011)
ROOT - NL	Root depth (Needleleaf)	[m]	0.05	User Selected
ROOT - BL	Root depth (Broadleaf)	[m]	0.05	User Selected
DDEN	Drainage density, equal to the length of the stream divided by area drained by the stream (Basin wide)	[km km^{-2}]	50	Dingman (2002)
XSLP	Average overland slope.	[rise/run]	grid-based	Calculated from Digital Elevation Model
GRKF	Ratio of saturated horizontal hydraulic conductivity at a depth of 1 metre to the saturated horizontal hydraulic conductivity at the surface (Basin wide)	[ ]	0.01	User defined

Table 3. Ranges for the perturbed parameters.

Parameter Name	Description	Units	Lower Limit	Upper Limit	Source
MANN	Manning's n for overland flow.	[m $s^{-1/3}$]	0.02	0.16	Dingman (2002)
KS	Saturated surface horizontal soil conductivity.	[m s^{-1}]	0.00001	0.1	User specified
ZSNL	Limiting snow depth below which coverage is less than one-hundred percent.	[m]	0.1	1	User specified
SDEP	Soil permeable depth, set to greater than model soil depth to simulate fully permeable soil.	[m]	0.1	4.2	User specified
WF-R2	River roughness factor that incorporates a channel shape and width to depth ratio as well as Manning's n.	[$m^{0.5}$ s^{-1}]	0.3	1	User specified
RSMN-NL	Minimum stomatal resistance (Needleleaf)	[s m^{-1}]	175	225	Verseghy (2011)
RSMN-BL	Minimum stomatal resistance (Broadleaf)	[s m^{-1}]	100	150	Verseghy (2011)
SAND-L1	Sand in soil layer 1.	[%]	35	58	Ecodistrict based
SAND-L2	Sand in soil layer 2.	[%]	35	58	Ecodistrict based
SAND-L3	Sand in soil layer 3.	[%]	35	58	Ecodistrict based
CLAY-L1	Clay in soil layer 1.	[%]	0	37	Ecodistrict based
CLAY-L2	Clay in soil layer 2.	[%]	0	37	Ecodistrict based
CLAY-L3	Clay in soil layer 3.	[%]	0	37	Ecodistrict based
LANZ0-NL	Natural log of roughness length (Needleleaf).	[ln(m)]	-0.7	1.1	Verseghy (2011)
LANZ0-BL	Natural log of roughness length (Broadleaf).	[ln(m)]	-0.7	1.1	Verseghy (2011)
ALVC-NL	Visible albedo (Needleleaf).	[ ]	0.02	0.09	Verseghy (2011)
ALVC-BL	Visible albedo (Broadleaf).	[ ]	0.02	0.09	Verseghy (2011)
ALIC-NL	Near infrared albedo (Needleleaf).	[ ]	0.1	0.5	Verseghy (2011)
ALIC-BL	Near infrared albedo (Broadleaf).	[ ]	0.1	0.5	Verseghy (2011)
LAMAX-NL	Maximum leaf area index (Needleleaf).	[ ]	1.8	2.2	Verseghy (2011)
LAMAX-BL	Maximum leaf area index (Broadleaf).	[ ]	4	10	Verseghy (2011)
LAMIN-NL	Minimum leaf area index (Needleleaf).	[ ]	1.4	1.8	Verseghy (2011)
LAMIN-BL	Minimum leaf area index (Broadleaf).	[ ]	0.2	4	Verseghy (2011)
MAXMASS-NL	Standing biomass density (Needleleaf).	[kg m^{-2}]	5	40	Verseghy (2011)
MAXMASS-BL	Standing biomass density (Broadleaf).	[kg m^{-2}]	5	40	Verseghy (2011)
ZPLS	Maximum water ponding depth for snow-covered areas.	[m]	0.1	0.5	User specified
ZPLG	Maximum water ponding depth for snow-free areas.	[m]	0.1	0.5	User specified
DRN	Drainage index, set to 1.0 to allow the soil physics to model drainage or to a value between 0.0 and 1.0 to impede drainage.	[m]	0	1	User specified

Table 4. Mean error ($m^3\ s^{-1}$) as an assessment of the ensemble mean streamflow for the reference forecast, the optimal hindcast, and various configurations of P-SEDA. The value of k_M from Algorithms **??** and 1 varies from 5 to 50 as shown. Rain-influenced and rain-free periods from June to October, 2014 as described in the text.

	Rain Free						Rain Influenced					
	k_M						k_M					
	5	10	20	30	40	50	5	10	20	30	40	50
Reference Forecast	2	2	2	2	2	2	-8	-8	-8	-8	-8	-8
Optimal hindcast	0	0	0	0	0	0	3	3	3	4	4	4
3-day screen	1	2	2	2	2	2	41	42	45	46	46	46
10-day screen	1	1	2	2	2	2	11	10	12	12	12	13
20-day screen	1	1	1	2	2	2	4	5	7	8	9	9
30-day screen	2	1	2	2	3	3	4	6	8	9	9	9
40-day screen	2	2	3	3	3	4	6	7	8	9	9	9
3-day screen with constrained parameters	3	4	4	4	5	5	2	4	4	5	5	6

Table 5. Mean continuous ranked probability skill score (CRPSS) as an assessment of the ensemble skill from P-SEDA for rain-influenced and rain-free periods from June to October, 2014. The configurations considered here are the optimal hindcast, the 3-day projection for the 20-day screen, and the 3-day projection for the 3-day screen with constrained parameters. The reference forecast is the measured streamflow at 00 UTC and 12 UTC each day as the forecast for the next 72 hours. The value of $k_m = 10$ from Algorithms **??** and 1 in all cases.

	Rain Free			Rain Influenced		
	Forecast Hour			Forecast Hour		
	24	48	72	24	48	72
Optimal hindcast	0.85	0.92	0.90	0.76	0.83	0.70
20-day screen	-0.09	0.27	0.42	-0.37	-0.15	-0.08
3-day screen with constrained parameters	-0.74	-0.22	-0.02	0.41	0.45	0.42

Table 6. Mean error (mean H-EPS streamflow - observed) ($m^3\ s^{-1}$) as an assessment of the ensemble mean streamflow from the H-EPS (200 members) for rain-influenced, rain-free and overall periods from June to October, 2014.

	Rain Free			Rain Influenced			Overall		
	Forecast Hour			Forecast Hour			Forecast Hour		
	24	48	72	24	48	72	24	48	72
optimized hindcast	1	2	3	1	2	7	1	2	3
3-day screen with constrained parameters	2	3	3	1	2	4	2	3	3

Table 7. Mean continuous ranked probability skill score (CRPSS) as an assessment of the ensemble skill from the H-EPS for rain-influenced, rain-free and overall periods from June to October, 2014. The reference low-skill forecast is the measured streamflow at 00 UTC and 12 UTC each day as the forecast for the next 72 hours.

	Rain Free			Rain Influenced			Overall		
	Forecast Hour			Forecast Hour			Forecast Hour		
	24	48	72	24	48	72	24	48	72
optimized hindcast	0.77	0.76	0.67	-0.33	0.12	0.01	0.71	0.70	0.60
3-day screen with constrained parameters	-0.77	-0.31	-0.14	-1.2	-0.15	0.11	-0.80	-0.29	-0.11

Table 8. Mean error (mean REPS precipitation - CaPA), CRPS and CRPSS (with JJASO, 2014 "climatology" as reference forecast) for June 1 to October 31, 2014.

	Mean Error			CRPS			CRPSS		
Threshold	Forecast Hour			Forecast Hour			Forecast Hour		
	24	48	72	24	48	72	24	48	72
$Pr \leq 0.9(0.42 \text{ mm h}^{-1})$	0.05	0.06	0.09	0.03	0.03	0.04	0.30	0.19	0.08
$Pr > 0.9(0.42 \text{ mm h}^{-1})$	-0.14	-0.21	-0.19	0.45	0.49	0.56	0.38	0.32	0.22
all	0.03	0.03	0.06	0.08	0.08	0.10	0.36	0.28	0.18