Hybridizing Bayesian and variational data assimilation for robust high-resolution hydrologic forecasting

Felipe Hernández, Xu Liang

Civil and Environmental Engineering Department, University of Pittsburgh, Pittsburgh, 15213, United States of America

Correspondence to: Xu Liang (xuliang@pitt.edu)

List of changes to the manuscript

We would like to once again thank the two anonymous referees and the editor for their valuable comments and suggestions which lead to significant modifications to the manuscript. We believe implementing their comments and suggestions has improved the presentation, clearness, and accuracy of our work. Below is the list of changes made in the revision. Specific pages and lines on the modified manuscript are referenced by page and line number as: (p<page number>, l<line number(s)>).

- (p1, l14-16) clarify that two different case studies were used, each with a different modelling engine and a different spatial resolution.
- (p1, l16-19) and (p1, l21-24) summarize the new comparison results in light of the addition of the continuous forecasting experiment and summarize the advantages of OPTIMISTS over existing methods.
- The last two paragraphs of the introduction (p3, l9-28) and the first paragraph of section 2 (p3-4) now include the differences between OPTIMISTS and other hybrid data assimilation approaches.
- The last two paragraphs of the introduction (p3, l9-28) explain some of the contrasts presented in Table 1 to improve the ties between the text and the table.
- Figure 1 was expanded to illustrate the entire functional structure of the algorithm and the first paragraph in subsection 2.1 (p4) ties the presented structure with the text.
- References to subsection 2.2, which deals with the computations with the state probability distribution, are found in subsubsections 2.1.2 (p5) and 2.1.4 (p6) to remind the reader that the details regarding the distributions will be discussed at a later point.
- The fifth paragraph of subsection 3.1 (p12, l14-23) summarizes how the factorial experiments were set up and its purpose in determining the best assignments to OPTIMISTS parameters.
- The second-to-last paragraph in subsection 3.1 (p12, l24-29) discusses how OPTIMISTS deals with observation errors in contrast with well-known methods.
- The original Figure 4 was removed to avoid the confusing details of the analysis of the results of Experiment 1. The main take-away is summarized in (p15, l14-17).

- Table 5 and Figure 4 show the results of the newly-included continuous forecasting experiment that compares OPTIMISTS and a particle filter in a year-long time frame. The experimental setup is described in subsection 3.2 (p12) and the results are discussed in the last two paragraphs of subsection 4.1 (p15, l18-32).
- Table 5 includes the probabilistic CRPS forecast performance metric in addition to the previous three deterministic performance metrics.
- Figure 4 shows the distribution of the streamflow forecast and not only its mean.
- (p16, l30-p17, l2) discuss why the errors in the scenario-based experiments seem large.
- A section named "Previous versions" was added before the references (p20) that references the 2016 version of the manuscript published in HESSD and mentions the differences between that version and the current one.

Hybridizing Bayesian and variational data assimilation for robust high-resolution hydrologic forecasting

Felipe Hernández, Xu Liang

Civil and Environmental Engineering Department, University of Pittsburgh, Pittsburgh, 15213, United States of America

Correspondence to: Xu Liang (xuliang@pitt.edu)

Abstract. The success of real-time estimation and forecasting applications based on geophysical models has been possible thanks to the two main existing frameworks for the determination of the models' initial conditions: Bayesian data assimilation and variational data assimilation. However, while there have been efforts to unify these two paradigms, existing attempts struggle to fully leverage the advantages of both in order to face the challenges posed by modern high-resolution models—mainly related to model indeterminacy and steep computational requirements. In this article we introduce a hybrid algorithm called OPTIMISTS (Optimized PareTo Inverse Modeling through Integrated STochastic Search) which is targeted at non-linear high-resolution problems and that brings together ideas from particle filters, 4-dimensional variational methods, evolutionary Pareto optimization, and kernel density estimation in a unique way. Streamflow forecasting experiments were conducted to test which specific configurations/parameterizations of OPTIMISTS ledcan lead to higher predictive accuracy. The experiments analysedwere conducted on two watersheds, one with a : the Blue River (low–resolution) using the VIC (Variable Infiltration Capacity) model and one with a the Indiantown Run (high-resolution) using the DHSVM (Distributed Hydrology Soil Vegetation Model). By selecting kernel-based non-parametric sampling, non-sequential evaluation of candidate particles, and through the multi-objective minimization of departures from the streamflow observations and from the background states, OPTIMISTS was shown to outperformefficiently produce probabilistic forecasts with higher or similar accuracy than those produced using a particle filter and a 4D variational method. Moreover, the experiments demonstrated that OPTIMISTS scales well in high-resolution cases without imposing a significant computational overhead and that it was successful in mitigating the harmful effects of overfitting. With thesethe combined advantages of allowing for fast, non-Gaussian, non-linear, high-resolution prediction, the algorithm shows the potential to increase the accuracy and efficiency offor operational prediction systems for the improved management of natural resources.

1 Introduction

Decision support systems that rely on model-based forecasting of natural phenomena are invaluable to society (Adams et al., 2003; Penning-Rowsell et al., 2000; Ziervogel et al., 2005)(Adams et al., 2003; Penning-Rowsell et al., 2000; Ziervogel et al., 2005). However, despite increasing availability of Earth-sensing data, the problem of estimation or prediction in geophysical systems remains as underdetermined as ever because of the growing complexity of such models (Clark et al., 2017)(Clark et al., 2017). For example, taking advantage of distributed physics and the mounting availability of computational power, modern

models have the potential to more accurately represent impacts of heterogeneities on eco-hydrological processes (Koster et al., 2017)~~.~~. This is achieved through the replacement of lumped representations with distributed ones, which entails the inclusion of numerous parameters and state variables. The price to pay for ~~this abandonment of~~thus forsaking parsimony is the added uncertainty in the evaluation of these additional unknowns. Therefore, in order to be able to rely on these high-resolution models for critical real-time and forecast applications, considerable improvements on ~~traditional~~ parameter and initial state estimation techniques must be made with two main goals: First, to allow for an efficient management of the huge number of unknowns; and second, to mitigate the harmful effects of overfitting—i.e., the loss of forecast skill due to an over-reliance on the calibration/training data (Hawkins, 2004)~~(Hawkins, 2004).~~. Because of the numerous degrees of freedom associated with these high-resolution distributed models, overfitting is a much bigger threat due to the phenomenon of equifinality (Beven, 2006)~~(Beven, 2006).~~.

There exists a plethora of techniques to initialize the state variables of a model through the incorporation of available observations, and they possess overlapping features that make it difficult to develop clear-cut classifications. However, two main "schools" can be fairly identified: Bayesian data assimilation and variational data assimilation. Bayesian data assimilation creates probabilistic estimates of the state variables in an attempt to also capture their uncertainty. These state probability distributions are adjusted sequentially to better match the observations using Bayes' theorem. While the ~~seminal~~ Kalman ~~Filter~~filter (KF) is constrained to linear dynamics and Gaussian distributions, ~~Ensemble~~ensemble Kalman ~~Filters~~filters (EnKF) can support non-linear models (Evensen, 2009)~~(Evensen, 2009),~~ and ~~Particle Filters~~particle filters (PF) can also manage non-Gaussian estimates for added accuracy (Smith et al., 2013)~~(Smith et al., 2013).~~. The stochastic nature of these Bayesian filters is highly valuable because equifinality can rarely be avoided and ~~due to~~because of the benefits of quantifying the uncertainty in forecasting applications (Verkade and Werner, 2011; Zhu et al., 2002)~~(Verkade and Werner, 2011; Zhu et al., 2002).~~. While superior in accuracy, PFs are usually regarded as impractical for high-dimensional applications (Snyder et al., 2008)~~(Snyder et al., 2008),~~ and thus recent research has focused on improving their efficiency (van Leeuwen, 2015)~~(van Leeuwen, 2015).~~.

On the other hand, variational data assimilation is more akin to traditional calibration approaches (Efstratiadis and Koutsoyiannis, 2010)~~(Efstratiadis and Koutsoyiannis, 2010)~~ because of its use of optimization methods. It seeks to find a single/deterministic initial state variable combination that minimizes the departures (or "variations") of the modelled values from the observations (Reichle et al., 2001)~~(Reichle et al., 2001)~~ and, commonly, from their history. One- to three- dimensional variants are also employed sequentially, but the paradigm lends itself easily to evaluating the performance of candidate solutions throughout an extended time window in four-dimensional versions (4D-Var). If the model's dynamics are linearized, the optimum can be very efficiently found in the resulting convex search space through the use of gradient methods. While this feature has made 4D-Var very popular in meteorology and oceanography (Ghil and Malanotte-Rizzoli, 1991)~~(Ghil and Malanotte-Rizzoli, 1991),~~ its application in hydrology has been less widespread because of the difficulty of linearizing land-surface physics (Liu and Gupta, 2007)~~(Liu and Gupta, 2007).~~. Moreover, variational data assimilation requires the inclusion of computationally-expensive adjoint models if one wishes to account for the uncertainty of the state estimates (Errico, 1997)~~(Errico, 1997).~~.

Traditional implementations from both schools have interesting characteristics and thus the development of hybrid methods has received considerable attention (Bannister, 2016)~~(Bannister, 2016).~~ For example, Bayesian filters have been used as adjoints in 4D-Var to enable probabilistic estimates (Zhang et al., 2009)~~(Zhang et al., 2009).~~ Moreover, some Bayesian approaches have been coupled with optimization techniques to select ensemble members (Dumedah and Coulibaly, 2013; Park et al., 2009)~~(Dumedah and Coulibaly, 2013; Park et al., 2009).~~ 4DEnVar (Buehner et al., 2010)~~(Buehner et al., 2010),~~ a fully-hybridized algorithm, is gaining increasing attention for weather prediction (Desroziers et al., 2014; Lorenc et al., 2015)~~(Desroziers et al., 2014; Lorenc et al., 2015).~~ It is especially interesting that some algorithms have defied the traditional choice between sequential and "extended-time" evaluations. Weak-constrained 4D-Var allows state estimates to be determined at several time steps within the assimilation time window and not only at the beginning (Ning et al., 2014; Trémolet, 2006)~~(Ning et al., 2014; Trémolet, 2006).~~ Conversely, modifications to EnKFs and PFs have been proposed to extend the analysis of candidate members/particles to span multiple time steps (Evensen and van Leeuwen, 2000; Noh et al., 2011)~~(Evensen and van Leeuwen, 2000; Noh et al., 2011).~~ The success of these hybrids demonstrates that there is a balance to be sought between the allowed number of degrees of freedom and the amount of information to be assimilated at once.

Following these promising paths, in this article we introduce OPTIMISTS (Optimized PareTo Inverse Modelling through Integrated STochastic Search), a hybrid data assimilation algorithm ~~that incorporates the most valuable features from both Bayesian and variational methods while mitigating the deficiencies or disadvantages of the original approaches (e.g., the linearity and determinism of 4D-Var and the limited scalability of PFs). The choice of the selected features and the design of their interactions were~~whose design was guided by the two stated goals: to allow for practical scalability to high-dimensional models, and to enable balancing the imperfect observations and the imperfect model estimates to minimize overfitting. Table 1~~This unique set of interactions is unlike those of any other method in the current literature.~~

~~Table 1~~ summarizes the main characteristics of typical Bayesian and variational approaches, and their contrasts with ~~OPTIMISTS.~~those of OPTIMISTS. Our algorithm incorporates the features that the literature has found to be the most valuable from both Bayesian and variational methods while mitigating the deficiencies or disadvantages associated with these original approaches (e.g., the linearity and determinism of 4D-Var and the limited scalability of PFs): Non-Gaussian probabilistic estimation and support for non-linear model dynamics have been long held as advantageous over their alternatives (Gordon et al., 1993; van Leeuwen, 2009)~~(Gordon et al., 1993; van Leeuwen, 2009)~~ and, similarly, meteorologists favour extended-period evaluations over sequential ones (Gauthier et al., 2007; Rawlins et al., 2007; Yang et al., 2009)~~(Gauthier et al., 2007; Rawlins et al., 2007; Yang et al., 2009).~~ As shown in the table, OPTIMISTS can readily ~~adopts~~adopt these proven strategies~~, but~~. However, there are ~~others in~~other aspects of the assimilation problem for which no single combination of features has demonstrated its superiority. For example, is the consistency with previous states better achieved through the minimization of a cost function that includes a background error term (Fisher, 2003)~~(Fisher, 2003),~~ as in variational methods, or through limiting the exploration to samples drawn from that background state distribution, as in Bayesian methods? Table 1~~While~~ shows that in these cases OPTIMISTS allows for flexible configurations, ~~our~~and it is an additional objective ~~is also~~of this study to test which set of ~~features~~feature interactions allows for more accurate forecasts when using highly-distributed models.

While many of the concepts utilized within the algorithm have been proposed in the literature before, their combination and the broad range of configurations available did not exist before. In other words, the concepts used in OPTIMISTS are unlike those of other methods, including existing hybrids which have mostly been developed around ensemble Kalman filters and convex optimization techniques (Bannister, 2016)—and therefore limited to Gaussian distributions and linear dynamics.

2 Data assimilation algorithm

In this section we describe OPTIMISTS, our proposed data assimilation algorithm which combines advantageous features from several Bayesian and variational methods. As will be explained in detail ~~in~~for each of the steps of the algorithm, these features were selected with the intent of mitigating the limitations of existing methods. ~~While most of the components utilized within the algorithm have been proposed in the literature before, their unique combination is innovative and distinct from the typical approaches used for hybridized data assimilation.~~

OPTIMISTS allows selecting a flexible data assimilation time step Δt—i.e., the time window in which candidate state configurations are compared to observations. It can be as short as the model time step, or as long as the entire assimilation window. For each assimilation time step at time t a new state probability distribution $\boldsymbol{S}^{t+\Delta t}$ is estimated from the current distribution $\boldsymbol{S}^{t}$, the model, and one or more observations $\boldsymbol{o}_{\text{obs}}^{t:t+\Delta t}$. For hydrologic applications, as those explored in this article, these states $\boldsymbol{S}$ include land surface variables within the modelled watershed such as soil moisture, snow cover/snow water equivalent, and stream water volume; and observations $\boldsymbol{o}$ are typically of streamflow at the outlet (Clark et al., 2008)~~(Clark et al., 2008),~~ soil moisture (Houser et al., 1998)~~(Houser et al., 1998),~~ and/or snow cover (Andreadis and Lettenmaier, 2006)~~(Andreadis and Lettenmaier, 2006).~~ However, the description of the algorithm will use field-agnostic terminology not to discourage its application in other disciplines.

~~The~~State probability distributions $\boldsymbol{S}$ in OPTIMISTS are determined from a set of weighted "root" or "base" sample states $\boldsymbol{s}_i$ using multivariate weighted kernel density estimation (West, 1993)~~(West, 1993).~~ This form of non-parametric distributions ~~stand~~stands in stark contrast with those from KFs and EnKFs in their ability to model non-Gaussian behaviour—an established advantage of PFs. Each of these samples or ensemble members $\boldsymbol{s}_i$ is comprised of a value vector for the state variables. The objective of the algorithm is then to produce a set of n samples $\boldsymbol{s}_i^{t+\Delta t}$ with corresponding weights w_i for the next assimilation time step to determine the target distribution $\boldsymbol{S}^{t+\Delta t}$.

This process is repeated iteratively each assimilation time step Δt until the entire assimilation time frame is covered, at which point the resulting distribution can be used to perform the forecast simulations. In subsection 2.1 we describe the main ideas and steps involved in the OPTIMISTS data assimilation algorithm; details regarding the state probability distributions, mainly on how to generate random samples and evaluate the likelihood of particles, are explained in subsection 2.2; and modifications required for high-dimensional problems are described in subsection 2.3.

2.1 Description of the OPTIMISTS data assimilation algorithm

Let a "particle" $\boldsymbol{P}_i$ be defined by a "source" (or initial) vector of state variables $\boldsymbol{s}_i^t$~~,~~ (which is a sample of distribution $\boldsymbol{S}^t$), a corresponding "target" (or final) state vector $\boldsymbol{s}_i^{t+\Delta t}$~~,~~ (a sample of distribution $\boldsymbol{S}^{t+\Delta t}$), a set of output values $\boldsymbol{o}_i^{t:t+\Delta t}$ (those that have corresponding observations $\boldsymbol{o}_{\text{obs}}^{t:t+\Delta t}$), a set of fitness metrics $\boldsymbol{f}_i$, a rank r_i, and a weight w_i. Note that the denomination ~~of~~ "particle" stems from the PF literature and is analogous to the "member" term in EnKFs. The fitness metrics $\boldsymbol{f}_i$ are used to compare particles with each other in the light of one or more optimization objectives. The algorithm consists of the following steps, whose motivation and details are included in the subsubsections below and their interactions illustrated in Figure 1~~. Table 2~~. Table 2 lists the meaning of each of the seven global parameters ~~involved in the algorithm~~ (Δt, n, w_{root}, p_{samp}, $k_{\text{F-class}}$, n_{evo}, and g).

1. Drawing: draw root samples $\boldsymbol{s}_i^t$ from $\boldsymbol{S}^t$ in descending weight order until $\sum w_i \geq w_{\text{root}}$
2. Sampling: randomly sample $\boldsymbol{S}^t$ until the total number of samples in the ensemble is $p_{\text{samp}} * n$
3. Simulation: compute $\boldsymbol{s}_i^{t+\Delta t}$ and $\boldsymbol{o}_i^{t:t+\Delta t}$ from each non-evaluated sample $\boldsymbol{s}_i^t$ using the model
4. Evaluation: compute the fitness values $\boldsymbol{f}_i$ for each ~~sample/~~particle $\boldsymbol{P}_i$
5. Optimization: create additional samples using evolutionary algorithms and return to 3 (if number of samples is below n)
6. Ranking: assign ranks r_i to all particles $\boldsymbol{P}_i$ using non-dominated sorting
7. Weighting: compute the weight w_i for each particle $\boldsymbol{P}_i$ based on its rank r_i

2.1.1 Drawing step

While traditional PFs draw all the root (or base) samples from $\boldsymbol{S}^t$ (Gordon et al., 1993)~~(Gordon et al., 1993),~~ OPTIMISTS can limit this selection to a subset of them. The root samples with the highest weight—those that are the "best performers"—are drawn first, then the next ones in descending weight order, until the total weight of the drawn samples $\sum w_i$ reaches w_{root}. w_{root} thus controls what percentage of the root samples to draw, and, if set to one, all of them are selected.

2.1.2 Sampling step

In this step the set of root samples drawn is complemented with random samples. The distinction between root samples and random samples is that the former are those that define the probability distribution $\boldsymbol{S}^t$ (that serve as centroids for the kernels), while the latter are generated stochastically from the kernels. Random samples are generated until the size of the combined set reaches $p_{\text{samp}} * n$~~.~~ by following the equations introduced in subsection 2.2. This second step contributes to the diversity of the ensemble in order to avoid sample impoverishment as seen on PFs (Carpenter et al., 1999)~~(Carpenter et al., 1999),~~ and serves as a replacement for traditional resampling strategies (Liu and Chen, 1998)~~(Liu and Chen, 1998).~~ The parameter w_{root} therefore controls the intensity with which this feature is applied to offer users some level of flexibility. ~~Generating~~Here generating random samples at the beginning, instead of resampling those that have been already evaluated, could lead to

discarding degenerate particles (those with high errors) early on and contribute to improved efficiency, given that the ones discarded are mainly those with the lowest weight as determined in the previous assimilation time step.

2.1.3 Simulation step

In this step, the algorithm uses the model to compute the resulting state vector $\boldsymbol{s}_i^{t+\Delta t}$ and an additional set of output variables $\boldsymbol{o}_i^{t:t+\Delta t}$ for each of the samples (it is possible that state variables double as output variables~~, though).~~). The simulation is executed starting at time t for the duration of the assimilation time step Δt (not to be confused with the model time step which is usually shorter). Depending on the complexity of the model, ~~this~~the simulation step can be the one with the highest computational requirements. In those cases, parallelization of the simulations would ~~help~~ greatly ~~to reduce~~help in reducing the total footprint of the assimilation process. The construction of each particle $\boldsymbol{P}_i$ is started by assembling the corresponding values computed so far: $\boldsymbol{s}_i^t$~~,~~ (drawing, sampling, and optimization steps), and $\boldsymbol{s}_i^{t+\Delta t}$~~,~~ and $\boldsymbol{o}_i^{t:t+\Delta t}$~~.~~ (simulation step).

2.1.4 Evaluation step

In order to determine which initial state $\boldsymbol{s}_i^t$ is the most desirable, a two-term cost function J is typically used in variational methods that simultaneously measures the resulting deviations of modelled values $\boldsymbol{o}_i^{t:t+\Delta t}$ from observed values $\boldsymbol{o}_{\text{obs}}^{t:t+\Delta t}$ and the departures from the background state distribution $\boldsymbol{S}^t$ (Fisher, 2003)~~(Fisher, 2003).~~. The function usually has the form shown in Eq. (1)~~(1):~~:

$$J_i = c_1 \cdot J_{\text{background}}(\boldsymbol{s}_i^t, \boldsymbol{S}^t) + c_2 \cdot J_{\text{observations}}\left(\boldsymbol{o}_i^{t:t+\Delta t}, \boldsymbol{o}_{\text{obs}}^{t:t+\Delta t}\right), \quad (1)$$

where c_1 and c_2 are balancing constants ~~usually~~normally set so that $c_1 = c_2$. Such a multi-criteria evaluation is crucial both to guarantee a good level of fit with the observations (second term) and to avoid the optimization algorithm to produce an initial state that is inconsistent with previous states (~~thus~~first term)—which could potentially ~~resulting~~result in overfitting problems rooted in disproportionate violations of mass and energy conservation laws~~).~~ (e.g., in hydrologic applications a sharp, unrealistic rise in the initial soil moisture could reduce $J_{\text{observations}}$ but would increase $J_{\text{background}}$). In Bayesian methods, since the consistency with the history is maintained by sampling only from the prior/background distribution $\boldsymbol{S}^t$, single term functions are used instead—which typically measure the probability density or likelihood of the modelled values given a distribution of the observations.

In OPTIMISTS~~,~~ any such fitness metric could be used and, most importantly, the algorithm allows defining several of them. Moreover, users can determine whether if each function is to be minimized (e.g., costs or errors) or maximized (e.g., likelihoods). We expect these features to be helpful if one wishes to separate errors when multiple types of observations are available (Montzka et al., 2012)~~(Montzka et al., 2012)~~ and as a more natural way to consider different fitness criteria (lumping them together in a single function as in Eq. (1) can lead to balancing and "apples and oranges" complications). Moreover, it might prove beneficial to take into account the consistency with the state history both by explicitly defining such an objective here and by allowing states to be sampled from the previous distribution (and thus compounding the individual mechanisms

of Bayesian and variational methods). Functions to measure this consistency are proposed in subsection 2.2. With the set of ~~optimization~~objective functions defined by the user, the algorithm computes the vector of fitness metrics $\boldsymbol{f}_i$ for each particle during the evaluation step.

2.1.5 Optimization step

The optimization step is optional and is used to generate additional particles by exploiting the knowledge encoded in the fitness values of the current particle ensemble. In a twist to the signature characteristic of variational data assimilation, OPTIMISTS incorporates evolutionary multi-objective optimization algorithms (Deb, 2014)~~(Deb, 2014)~~ instead of the established gradient-based, single-objective methods. Evolutionary optimizers compensate their slower convergence speed with the capability of efficiently navigating non-convex solution spaces (i.e., the models and the fitness functions do not need to be linear with respect to the observations and the states). This feature effectively opens the door for variational methods to be used in disciplines where the linearization of the driving dynamics is either impractical, inconvenient, or undesirable. Whereas any traditional multi-objective global optimization method would work, our implementation of OPTIMISTS features a state-of-the-art adaptive ensemble algorithm similar to AMALGAM (Vrugt and Robinson, 2007)~~(Vrugt and Robinson, 2007)~~ that allows model simulations to be run in parallel (Crainic and Toulouse, 2010)~~(Crainic and Toulouse, 2010).~~. The optimizer ensemble includes a genetic algorithm (Deb et al., 2002)~~(Deb et al., 2002)~~ and a hybrid approach that combines ant colony optimization (Socha and Dorigo, 2008)~~(Socha and Dorigo, 2008)~~ and Metropolis-Hastings sampling (Haario et al., 2001)~~(Haario et al., 2001).~~.

During the optimization step, the group of optimizers is used to generate n_{evo} new sample states $\boldsymbol{s}_i^t$ based on those in the current ensemble. For example, the genetic algorithm selects pairs of base samples with high performance scores $\boldsymbol{f}_i$ and then proceeds to combine their individual values using standard crossover and mutation operators. The simulation and evaluation steps are repeated for these new samples, and then this iterative process is repeated until the particle ensemble has a size of n. Note that w_{root} and p_{samp} thus determine what percentage of the particles is generated in which way. For example, for relatively small values of w_{root} and a p_{samp} of 0.2, 80% of the particles will be generated by the optimization algorithms. In this way, OPTIMISTS offers its users the flexibility to behave anywhere in the range between "fully Bayesian" ($p_{\text{samp}} = 1$) and "fully variational" ($p_{\text{samp}} = 0$) in terms of particle generation. In the latter case, in which no root and random samples are available, the initial "population~~""~~/ensemble of states $\boldsymbol{s}_i^t$ is sampled uniformly from the viable range of each state variable.

2.1.6 Ranking step

A fundamental aspect of OPTIMISTS is the way in which it provides a probabilistic interpretation to the results of the multi-objective evaluation, thus bridging the gap between Bayesian and variational assimilation. Such method has been used before (Dumedah et al., 2011)~~(Dumedah et al., 2011)~~ and is based on the employment of non-dominated sorting (Deb, 2014)~~(Deb, 2014),~~ another technique from the multi-objective optimization literature, which is used to balance the potential tensions

between various objectives. This sorting approach is centred on the concept of "dominance", instead of organizing all particles from the "best" to the "worst." A particle dominates another if it outperforms it according to at least one of the criteria/objectives while simultaneously is not outperformed according to any of the others. Following this principle, in the ranking step particles are grouped in "fronts" comprised of members which are mutually non-dominated; that is, none of them is dominated by any of the rest. Particles in a front, therefore, represent the effective trade-offs between the competing criteria. Figure 1 illustrates the result of non-dominated sorting applied to nine particles being analysed under two objectives: minimum deviation from observations and maximum likelihood given the background state distribution $\boldsymbol{S}^t$. Note that if a single objective function is used, the sorting method assigns ranks from "best" to "worst" according to that function, and two particles would only share ranks if their fitness value coincides. In our implementation we use the fast non-dominated sorting algorithm to define the fronts and assign the corresponding ranks r_i (Deb et al., 2002). More efficient non-dominated sorting alternatives are available if performance becomes an issue (Zhang et al., 2015).

2.1.7 Weighting step

In this final step, OPTIMISTS assigns weights w_i to each particle according to its rank r_i as shown in Eqs. (2) and (3). This Gaussian weighting depends on the ensemble size n and the greed parameter g, and is similar to the one proposed by (Socha and Dorigo, 2008). When g is equal to zero, particles in all fronts are weighted uniformly; when g is equal to one, only particles in the Pareto/first front are assigned non-zero weights. With this, the final estimated probability distribution of state variables for the next time step $\boldsymbol{S}^{t+\Delta t}$ can be established using multivariate weighted kernel density estimation, (details in the next subsection), as demonstrated in Fig. 1.be., by taking all target states $\boldsymbol{s}_i^{t+\Delta t}$ (circles) as the centroids of the kernels. The obtained distribution $\boldsymbol{S}^{t+\Delta t}$ can then be used as the initial distribution for a new assimilation time step or, if the end of the assimilation window has been reached, it can be used to perform (ensemble) forecast simulations.

$$w_i = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(r_i-1)^2}{2\sigma^2}} \tag{2}$$

$$\sigma = n \cdot [0.1 + 9.9 \cdot (1-g)^5] \tag{3}$$

2.2 Model state probability distributions

As mentioned before, OPTIMISTS uses kernel density probability distributions (West, 1993)OPTIMISTS to model the stochastic estimates of the state variable vectors. The algorithm requires two computations related to the state-variable probability distribution $\boldsymbol{S}^t$: obtaining the probability density p or likelihood $\mathcal{L}$ of a sample and generating random samples. The first computation can be used in the evaluation step as an objective function to preserve the consistency of particles with the state history. (e.g., to penalize aggressive departures from the prior conditions). It should be noted that other suchseveral metrics that try to approximate this consistency exist, from very simple (Dumedah et al., 2011) to quite complex (Ning et al., 2014). For example, it is common in variational data assimilation to utilize the background error term

$$J_{\text{background}} = (\boldsymbol{s} - \boldsymbol{s}_b)^{\mathrm{T}}\mathbf{C}^{-1}(\boldsymbol{s} - \boldsymbol{s}_b), \quad (4)$$

where $\boldsymbol{s}_b$ and $\mathbf{C}$ are the mean and the covariance of the "background" state distribution ($\boldsymbol{S}^t$ in our case) which is assumed to be Gaussian (Fisher, 2003). The term $J_{\text{background}}$ is plugged into the cost function shown in Eq. (1)~~(Ning et al., 2014). The~~. For OPTIMISTS, we propose that the probability density of ~~a~~the weighted state kernel density distribution $\boldsymbol{S}^t$ at a given point ($\boldsymbol{p}$) be used as a stand-alone objective. The density is given by Eq. (5) (Wand and Jones, 1994)~~ can be computed using Eq. (4) (Wand and Jones, 1994).~~ If Gaussian kernels are selected, the kernel function K, parameterized by the bandwidth matrix $\mathbf{B}$, is evaluated using Eq. (6)~~(5).~~.

$$p(\boldsymbol{s}|\boldsymbol{S}) = \frac{1}{\sum w_i}\sum_{i=1}^{n}[w_i \cdot K_{\mathbf{B}}(\boldsymbol{s} - \boldsymbol{s}_i)] \quad (5)$$

$$K_{\mathbf{B}}^{\text{Gauss}}(\boldsymbol{z}) = \frac{1}{\sqrt{(2\pi)^n \cdot |\mathbf{B}|}}\exp\left(-\frac{1}{2}\boldsymbol{z}^{\mathrm{T}}\mathbf{B}^{-1}\boldsymbol{z}\right) \quad (6)$$

Matrix $\mathbf{B}$ is the covariance matrix of the kernels, and thus determines their spread and orientation in the state space. $\mathbf{B}$ is of size $d \times d$, where d is the dimensionality of the state distribution (i.e., the number of variables), and can be thought of as a scaled-down version of the "background error covariance" matrix $\mathbf{C}$ from the variational literature. In this sense matrix $\mathbf{B}$, together with the spread of the ensemble of samples $\boldsymbol{s}_i$, effectively encode the uncertainty of the state variables. Several optimization-based methods exist to compute $\mathbf{B}$ by attempting to minimize the asymptotic mean integrated squared error (AMISE) (Duong and Hazelton, 2005; Sheather and Jones, 1991)~~(Duong and Hazelton, 2005; Sheather and Jones, 1991).~~ However, here we opt to use a simplified approach for the sake of computational efficiency: we determine $\mathbf{B}$ by scaling down the sample covariance matrix $\mathbf{C}$ using Silverman's rule of thumb, which takes into account the number of samples n and the dimensionality of the distribution d, as shown in Eq. (7) (Silverman, 1986)~~(6) (Silverman, 1986). Figure 1.b.~~ Figure 1 shows the density of ~~a~~two two-dimensional example ~~distribution~~distributions using this method~~.~~ (a and e). If computational constraints are not a concern, using AMISE-based methods or kernels with variable bandwidth (Hazelton, 2003; Terrell and Scott, 1992)~~(Hazelton, 2003; Terrell and Scott, 1992)~~ could result in higher accuracy.

$$\mathbf{B}^{\text{Silverman}} = \left(\frac{4}{d+2}\right)^{\frac{2}{d+4}} \cdot n^{-\frac{2}{d+4}} \cdot \mathbf{C} \quad (7)$$

Secondly, ~~we can generate~~OPTIMISTS' sampling step requires generating random samples from a multivariate weighted kernel density distribution. This is achieved by dividing the problem into two: we first select the root sample and then generate a random sample from the kernel associated with that base sample. The first step corresponds to randomly sampling a multinomial distribution with n categories and assigning the normalized weights of the particles as the probability of each category. Once a root sample $\boldsymbol{s}_{\text{root}}$ is selected, a random sample $\boldsymbol{s}_{\text{random}}$ can be generated from a vector $\boldsymbol{v}$ of independent standard normal random values of size d~~,~~ and a matrix $\mathbf{A}$ as shown in Eq. (8)~~(7).~~. $\mathbf{A}$ can be computed from a Cholesky decomposition (Krishnamoorthy and Menon, 2011)~~(Krishnamoorthy and Menon, 2011)~~ such that $\mathbf{A}\mathbf{A}^{\mathrm{T}} = \mathbf{B}$. Alternatively, an eigendecomposition can be used to obtain $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathrm{T}} = \mathbf{B}$ to then set $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}^{½}$.

$$\boldsymbol{s}_{\text{random}} = \boldsymbol{s}_{\text{root}} + \mathbf{A}\boldsymbol{v} \tag{8}$$

Both computations (density/likelihood and sampling) require **B** to be invertible and, therefore, that none of the variables have zero variance or are perfectly linearly~~–~~dependent on each other. Zero-variance variables must therefore be isolated and **B** marginalized before attempting to use Eq. (6)~~(5)~~ or to compute **A**. Similarly, linear dependencies must also be identified beforehand. If we include variables one by one in the construction of **C**, we can determine if a newly added one is linearly dependent if the determinant of the extended sample covariance matrix **C** is zero. Once identified, the regression coefficients for the dependent variable can be efficiently computed from **C** following the method described by (Friedman et al., 2008)~~(Friedman et al., 2008).~~. The constant coefficient of the regression must also be calculated for future reference. What this process effectively does is to determine a linear model for each dependent variable that is represented by a set of regression coefficients. Dependent variables are not included in **C,** but they need to be taken into account afterwards (e.g., by determining their values for the random samples by solving the linear model with the values obtained for the variables in **C**).

2.3 High-dimensional state vectors

When the state vector of the model becomes large (i.e., d increases), as is the case for distributed high-resolution numerical models, difficulties start to arise when dealing with the computations involving the probability distribution. At first, the probability density, as computed with Eqs. (5)~~(4)~~ and (6)~~(5),~~ tends to diverge either towards zero or towards infinity. This phenomenon is related to the normalization of the density—so that it can integrate to one—and to its fast exponential decay as a function of the sample's distance from the kernel's centres. In these cases we propose replacing the density computation with an approximated likelihood formulation that is proportional to the inverse square Mahalanobis distance (Mahalanobis, 1936)~~(Mahalanobis, 1936)~~ to the root samples, thus skipping the exponentiation and normalization operations of the Gaussian density. This simplification, which corresponds to the inverse square difference between the sample value and the kernel's mean in the univariate case, is shown in Eq. (9)~~(8).~~. The resulting distortion of the Gaussian bell-curve shape does not affect the results significantly, given that OPTIMISTS uses the fitness functions only to check for domination between particles—so only the sign of the differences between likelihood values are important and not their actual magnitudes.

$$\mathcal{L}^{\text{Mahalanobis}}(\boldsymbol{s}|\boldsymbol{S}) = \frac{1}{\sum w_i}\sum_{i=1}^{n}\frac{w_i}{|(\boldsymbol{s}-\boldsymbol{s}_i)^{\mathrm{T}}\mathbf{B}^{-1}(\boldsymbol{s}-\boldsymbol{s}_i)|} \tag{9}$$

However, computational constraints might also make this simplified approach unfeasible both due to the $O(d^2)$ space requirements for storing the bandwidth matrix **B** and the $O(d^3)$ time complexity of the decomposition algorithms, which rapidly become huge burdens for the memory and the processors. Therefore, we can chose to sacrifice some accuracy by using a diagonal bandwidth matrix **B** which does not include any covariance term—only the variance terms in the diagonal are computed and stored. This ~~means~~implies that, even though the multiplicity of root samples would help in maintaining a large portion of the covariance, another portion is lost by preventing the kernels from reflecting the existing correlations. In other words, variables would not be rendered completely independent, but rather conditionally independent because the kernels are

still centred on the set of root samples. Kernels using diagonal bandwidth matrices are referred to as "D-class" while those using the full covariance matrix are referred to as "F-class.". The $k_{\mathrm{F-class}}$ parameter controls which caseversion is used.

With only the diagonal terms of matrix **B** available (b_{jj}), we opt to roughly approximate the likelihood by computing the average of the standardized marginal likelihood value for each variable j, as shown in Eq. (10)(9)::

$$\mathcal{L}^{\text{independent}}(\boldsymbol{s}|\boldsymbol{S}) = \frac{1}{d\sqrt{2\pi}\sum w_i}\sum_{j=1}^{d}\sum_{i=1}^{n}\left\{w_i \cdot \exp\left[-\frac{(s_j - s_{i,j})^2}{2b_{jj}}\right]\right\}, \quad (10)$$

where s_j represents the j^{th} element of state vector $\boldsymbol{s}$, and $s_{i,j}$ represents the j^{th} element of the i^{th} sample of probability distribution $\boldsymbol{S}$. Independent/marginal random sampling of each variable can also be applied to replace Eq. (8)(7) by adding random Gaussian residuals to the elements of the selected root sample $\boldsymbol{s}_{\text{root}}$. Sparse bandwidth matrices (Friedman et al., 2008; Ghil and Malanotte-Rizzoli, 1991)(Friedman et al., 2008; Ghil and Malanotte-Rizzoli, 1991) or low-rank approximations (Bannister, 2008; Ghorbanidehno et al., 2015; Li et al., 2015)(Bannister, 2008; Ghorbanidehno et al., 2015; Li et al., 2015) could be worthwhile intermediate alternatives to our proposed quasi-independent approach to be explored in the future.

3 Experimental setup

In this section we prepare the elements to investigate whether if OPTIMISTS can help improve the forecasting skill of hydrologic models. More specifically, the experiments seek to answer the following questions: Which characteristics of Bayesian and variational methods are the most advantageous? How can OPTIMISTS be configured to take advantage of these characteristics? How does the algorithm compare to established data assimilation methods? And how does it perform with high-dimensional applications? To help answer these questions, this section first introduces two case studies and then it presents two assimilation algorithmscharacterizes a traditional PF that werewas used for comparison purposes.

3.1 Case studies

We coupled a Java implementation of OPTIMISTS with two popular open-source distributed hydrologic modelling frameworksengines: Variable Infiltration Capacity (VIC) (Liang et al., 1994, 1996a, 1996b, Liang and Xie, 2001, 2003)(Liang et al., 1994, 1996a, 1996b, Liang and Xie, 2001, 2003) and the Distributed Hydrology Soil and Vegetation Model (DHSVM) (Wigmosta et al., 2002, 1994)(Wigmosta et al., 2002, 1994).. VIC is targeted at large watersheds by focusing on vertical subsurface dynamics, and also enabling intra-cell precipitation, soil, and vegetation heterogeneity. The DHSVM, on the other hand, was conceived for high-resolution representations of the Earth's surface, allowing for saturated and unsaturated subsurface flow routing and 1D/2D surface routing (Zhang et al., 2018).. Both engines needed several modifications so that they could be executed in a non-continuous fashion as required for sequential assimilation. Given the non-Markovian nature of surface routing schemes coupled with VIC that are based either on multiscale routing frameworks (Guo et al., 2004; Wen et al., 2012)approaches (Guo et al., 2004; Wen et al., 2012) or on the unit hydrograph concept (Lohmann et al., 1998)(Lohmann et al., 1998),, a simplified routing routine was developed that treats the model cells as channels—albeit with longer retention

times. In the simplified method, direct runoff and baseflow produced by each model cell is partly routed through an assumed "equivalent" channel (slow component) and partly poured directly to the channel network (fast component). Both the channel network and the equivalent channels representing overland flow hydraulics are modelled using the Muskingum method. On the other hand, several important bugs in version 3.2.1 of the DHSVM, mostly related to the initialization of state variables but also pertaining to routing data and physics, were fixed.

We selected two ~~model applications~~watersheds to perform streamflow forecasting tests using OPTIMISTS: one with the VIC model running at a 1/8th degree resolution for the Blue River watershed in Oklahoma, and the other with the DHSVM running at a 100 m resolution for the Indiantown Run watershed in Pennsylvania. Table 3~~Table 3~~ lists the main characteristics of the two test watersheds and the information of their associated model configurations. Figure 2~~Figure 2~~ shows the land cover map together with the layout of the modelling cells for the two watersheds. The multi-objective ensemble optimization algorithm included in OPTIMISTS was employed to calibrate the ~~two models'~~ parameters of the two models with the streamflow measurements from the corresponding USGS stations. For the Blue River, the traditional ℓ_2-norm Nash-Sutcliffe Efficiency (NSE_{ℓ_2}) (which focuses mostly on the peaks of hydrographs), an ℓ_1-norm version of the Nash-Sutcliffe Efficiency coefficient (NSE_{ℓ_1}) (Krause et al., 2005)~~(Krause et al., 2005),~~ and the mean absolute relative error MARE (which focuses mostly on the inter-peak periods) were used as optimization criteria. From 85,600 candidate parameterizations tried, one was chosen from the resulting Pareto front with NSE_{ℓ_2} = 0.69, NSE_{ℓ_1} = 0.56, and MARE = 44.71%. For the Indiantown Run, the NSE_{ℓ_2}, MARE, and absolute bias were optimized, resulting in a parameterization, out of 2,575, with NSE_{ℓ_2} = 0.81, MARE = 37.85%, and an absolute bias of 11.83 l/s.

These "optimal" parameter sets, together with additional sets produced in the optimization process were used to run the models and determine a set of time-lagged state variable vectors $\boldsymbol{s}$ to construct the state probability distribution $\boldsymbol{S}^0$ at the beginning of each of a set of data assimilation ~~scenario~~scenarios. The state variables include liquid and solid interception; ponding, water equivalent and temperature of the snow packs; and moisture and temperature of each of the soil layers. While we do not expect all of these variables to be identifiable and sensitive within the assimilation problem, we decided to be thorough in their inclusion—a decision that also increases the challenge for the algorithm in terms of the potential for overfitting. The Blue River model application has 20 cells, with a maximum of seven intra-cell soil/vegetation partitions. After adding the stream network variables, the model has a total of d = 812 state variables. The Indiantown Run model application has a total of 1,472 cells and d = 33,455 state variables.

Three diverse scenarios were selected for the Blue River, each of them comprised of a two-week assimilation period (when streamflow observations are assimilated), and a two-week forecasting period (when the model is run in an open loop using the states obtained at the end of the assimilation period): Scenario 1, starting on October 15th, 1996, is rainy through the entire four weeks. Scenario 2, which starts on January 15th, 1997, has a dry assimilation period and a mildly rainy forecast period. Scenario 3, starting on June 1st, 1997, has a relatively rainy assimilation period and a mostly-dry forecast period. Two scenarios, also spanning four weeks, were selected for the Indiantown Run, one starting on July 26th, 2009 and the other on August 26th, 2009.

We used factorial experiments (Montgomery, 2012)~~(Montgomery, 2012)~~ to test different configurations of OPTIMISTS on each of these scenarios, by first assimilating the streamflow and then measuring the forecasting skill. ~~This~~In this type of experimental designs a set of assignments is established for each parameter and then all possible assignment combinations are tried. The design allows to establish the statistical significance of altering several parameters simultaneously, providing an adequate framework for determining, for example, whether if using a short or a long assimilation time step Δt is preferable, or if utilizing the optional optimization step within the algorithm is worthwhile. Table 4~~Table 4~~ shows the setup of each of the three full factorial experiments we conducted, together with the selected set of assignments for OPTIMISTS' parameters. The forecasts were produced in an ensemble fashion, by running the models using each of the samples $\boldsymbol{s}_i$ from the state distribution $\boldsymbol{S}$ at the end of the assimilation time period, and then using the samples' weights w_i to produce an average forecast. Deterministic model parameters (those from the calibrated models) and forcings were used in all simulations.

~~The~~Observation errors are usually taken into account in traditional assimilation algorithms by assuming a probability distribution for the observations at each time step, and then performing a probabilistic evaluation of the predicted value of each particle/member against that distribution. As mentioned in section 2, such a fitness metric like the likelihood utilized in PFs to weight candidate particles, is perfectly compatible with OPTIMISTS. However, since it is difficult to estimate the magnitude of the observation error in general, and fitness metrics $\boldsymbol{f}_i$ here are only used to determine (non-)dominance between particles, we opted to use the mean absolute error (MAE) with respect to the streamflow observations ~~was used as an objective in~~ in all cases.

For the Blue River scenarios, ~~the~~a secondary likelihood objective ~~(when~~/metric was used~~)~~ in some cases to select for particles with higher consistency with the history. It was computed using either Eq. (10)~~(9)~~ if $k_{\mathrm{F-class}}$ was set to false, or Eq. (9)~~(8)~~ if it was set to true. Equation (10)~~(9)~~ was used for all Indiantown Run scenarios given the large number of dimensions. The assimilation period was of two weeks for most configurations, except for those in Experiment 3 which have Δt = 4 weeks. During both the assimilation and the forecasting periods we used unaltered streamflow data from the USGS and forcing data from NLDAS-2 (Cosgrove et al., 2003)~~(Cosgrove et al., 2003)~~ even though a forecasted forcing would be used instead in an operational setting (e.g., from systems like NAM (Rogers et al., 2009)~~(Rogers et al., 2009)~~ or ECMWF (Molteni et al., 1996)~~(Molteni et al., 1996)).~~. While adopting ~~actual~~perfect forcings for the forecast period leads to an overestimation of their accuracy, any comparisons with control runs or between methods are still valid as they all share the same benefit. Also, removing the uncertainty in the meteorological forcings allows the analysis to focus on the uncertainty ~~related~~corresponding to the land surface.

3.2 Data assimilation method comparison

Comparing the performance of different configurations of OPTIMISTS can shed light into the adequacy of individual strategies utilized by traditional Bayesian and variational methods. For example, producing all particles with the optimization algorithms (p_{samp} = 0~~) and~~), setting long values for Δt, and utilizing a traditional two-term cost function as that in Eq. (1), makes the

method behave somewhat as a hard-constrained 4D-Var approach; while sampling all particles from the source state distribution ($p_{\text{samp}} = 1$), ~~and~~ setting Δt equal to the model time step, and using a single likelihood objective involving the observation error, would resemble a ~~traditional~~ PF. Herein we also compare OPTIMISTS with ~~two other data assimilation algorithms~~a traditional PF on the Blue River VIC model application~~: a PF, and an evolutionary 4D variational (Evo4DVar) method.~~

. Since the forcing is assumed to be deterministic, the implemented PF uses Gaussian "regularization"/perturbation of resampled particles to avoid degeneration (Pham, 2001)~~(Pham, 2001).~~. Resampling is executed such that the probability of duplicating a particle is proportional to their weight (Moradkhani et al., 2012)~~(Moradkhani et al., 2012). Evo4DVar is very similar to traditional 4DVar with the difference that an evolutionary optimization algorithm is used to navigate the non-convex solution space imposed by the non-linear nature of VIC. A single-objective version of the ensemble optimizer in OPTIMISTS is used, which is similar to AMALGAM-SO (Vrugt et al., 2009). The objective cost function J to be minimized takes into account departures from the observations and from the mean background state as is the orthodoxy in variational methods. Assuming the background state distribution to be normally distributed, J is given by Eq. (10):~~.

Additionally, the comparison is performed using a continuous forecasting experiment setup instead of a scenario-based one. In this continuous test, forecasts are performed daily (the same as the model time step) and compiled in series for different forecast lead times that span a whole year, from November, 1996 to November, 1997. Forecast lead times are of 1, 3, 6, 12, and 24 days. Before each daily forecast, both OPTIMISTS and the PF assimilate streamflow observations for the assimilation time step of each algorithm (daily for the PF). The assimilation is performed cumulatively, meaning that the initial state distribution $\boldsymbol{S}^t$ was produced by assimilating all the records available since the beginning of the experiment on October, 1996 until time t. The forecasted streamflow series are then compared to the actual measurements to evaluate their quality using deterministic metrics (NSE_{ℓ_2}, NSE_{ℓ_1}, and MARE) and a probabilistic one: the ensemble-based continuous ranked probability score (CRPS) (Bröcker, 2012)~~where $\boldsymbol{s}_B$ and $\boldsymbol{C}$ are the mean and the covariance of the background state distribution, n_t is the number of observations, q^t and q^t_{obs} are the modelled and observed streamflow values at time t, and σ_q is the standard deviation of the observations (due to measurement errors). While the PF produces an ensemble of states (probabilistic) at the end of the assimilation period, Evo4DVar produces a single state vector (deterministic) which results from the state that minimizes J.~~

, which is computed for each time step and then averaged for the entire duration of the forecast.

4 Results and discussion

~~The performance of each forecast produced by OPTIMISTS was analysed as follows.~~This section summarizes the forecasting results obtained from the three scenario-based experiments on the Blue River and the Indiantown Run model applications, and the continuous forecasting experiment performed on the Blue River application. The scenario-based experiments were performed to explore the effects of multiple parameterizations of OPTIMISTS, and the performance was analysed as follows.

The model was run for the duration of the forecast period (two weeks) using the state configuration encoded in each root state s_i of the distribution $\boldsymbol{S}$ obtained at the end of the assimilation period for each configuration of OPTIMISTS and each scenario. We then computed the mean streamflow time series for each case by averaging the model results for each particle $\boldsymbol{P}_i$ (the average was weighted based on the corresponding weights w_i). With this averaged streamflow series, we compute the three performance metrics—the NSE_{ℓ_2}, the NSE_{ℓ_1}, and the MARE—based on the observations from the corresponding stream gauge. The values for each experiment, scenario, and configuration are listed in tables in the supplementary material. With these, we compute the change in the forecast performance between each configuration and a control open-loop model run (one without the benefit of assimilating the observations).

4.1 Blue River – low resolution application

~~Figure 3~~The supplementary material includes the performance metrics for all of the tested configurations on all scenarios and for all experiments. Figure 3 summarizes the results for Experiment 1 with the VIC model application for the Blue River watershed, in which the distributions of the changes in MARE after marginalizing the results for each scenario and each of the parameter assignments are shown. That is, each box (and pair of whiskers) represents the distribution of change in MARE of all cases in the specified scenario or for which the specified parameter assignment was used. Negative values in the vertical axis indicate that OPTIMISTS decreased the error, while positive values indicate it increased the error. It can be seen that, on average, OPTIMISTS improves the precision of the forecast in most cases, except for several of the configurations in Scenario 1 (for this scenario the control already produces a good forecast) and when using an assimilation step Δt of one day. We performed an analysis of variance (ANOVA) to determine the statistical significance of the difference found for each of the factors~~.~~ indicated in the horizontal axis. While Figure 3~~Figure 3~~ shows the p-values for the main effects, the full ANOVA table for all experiments can be found in the supplementary material. From the values in ~~the figure~~Figure 3, we can conclude that the assimilation time step, the number of objectives, and the use of optimization algorithms are all statistically significant. On the other hand, the number of particles and the use of F-class kernels are not.

A Δt of five days produced the best results overall for the tested case, suggesting that there exists a sweet spot that balances the amount of information being assimilated (larger for a long Δt), and the number of state variables to be modified (larger for a small Δt). ~~In~~Based on such ~~case~~results, it ~~would be logical~~is reasonable to assume that ~~this~~the sweet spot ~~depends on the specific~~may depend on the time series of precipitation, the characteristics of the watershed, and the temporal and spatial resolutions of its model application. From ~~that~~this perspective, the poor results for a step of one day could be explained in terms of overfitting, where there are many degrees of freedom and only one value being assimilated per step. Evaluating particles in the light of two objectives, one minimizing departures from the observations and the other maximizing the likelihood of the source state, resulted in statistically-significant improvements compared to using the first objective alone. Additionally, the data suggests that not executing the optional optimization step of the algorithm ("optimization = false"), but instead relying only on particles sampled from the prior/source distribution, is also beneficial. These two results reinforce the

idea that maintaining consistency with the history to some extent is of paramount importance, perhaps to the point where the strategies used in Bayesian filters and variational methods are insufficient in isolation. Indeed, the best performance was observed only when both sampling was limited to generate particles from the prior state distribution and the particles were evaluated for their consistency with that distribution.

On the other hand, we found it counterintuitive that neither using a larger particle ensemble nor taking into account state-variable dependencies through the use of F-class kernels lead to improved results. In the first case it could be hypothesized that using too many particles ~~leads~~could lead to overfitting, since there ~~are~~would be more chances of particles being generated that happen to match the observations better but for the "wrong reasons." In the second case, the non-parametric nature of kernel density estimation could be sufficient for encoding the raw dependencies between variables, especially in low-resolution cases like this one, in which significant correlations between variables in adjacent cells are not expected to be too ~~large~~high. Both results deserve further investigation, especially concerning the impact of D- vs. F-class kernels in high-dimensional models.

Interestingly, the ANOVA also yielded small p-values for several high-order interactions (see the ANOVA table in the supplementary material). This means that, unlike the general case for factorial experiments as characterized by the sparsity-of-effects principle (Montgomery et al., 2009)~~(Montgomery et al., 2009),~~ specific combinations of multiple parameters have a large effect on the forecasting skill of the model. Significant interactions (with p smaller than 0.05) are between the objectives and Δt (p = 0.001); n and $k_{\mathrm{F-class}}$ (p = 0.039); Δt and the use of optimization (p = 0.000); the use of optimization and $k_{\mathrm{F-class}}$ (p = 0.029); the objectives, Δt, and the use of optimization (p = 0.043); n, Δt, and $k_{\mathrm{F-class}}$ (p = 0.020); n, the use of optimization, and $k_{\mathrm{F-class}}$ (p = 0.013); and n, Δt, the use of optimizers, and $k_{\mathrm{F-class}}$ (p = 0.006). These interactions show that, for example, using a single objective is especially inadequate when the time step is of one day or when optimization is used. ~~Figure 4 shows the forecast results for specific parameter combinations to help understand other such interactions. For example, the first column shows the distribution of the six cases with an assimilation time step of one day, 100 particles, no optimization algorithms, and D-class kernels. The boxplots demonstrate that, for instance~~Also, employing optimization is only significantly detrimental when Δt is of one day—probably because of intensified overfitting, and that choosing F-class kernels leads to higher errors when Δt is small, n large, and the optimizers are being used. ~~While other patterns can be pointed to, the supporting p-values (shown in the figure) are inconclusive.~~

Based on these results, we recommend the use of both objectives and no optimization as the preferred configuration of OPTIMISTS for the Blue River application. A time step of five days appears to be adequate for this specific model application. Also, without strong evidence for their advantages, we recommend using more particles or kernels of class F only if there is no pressure for computational frugality. ~~Figure 5 compares the forecasting accuracy of the model after assimilating the streamflow observations with the PF, Evo4DVar, and OPTIMISTS with a configuration of Δt = 5 days, two objectives, n = 100, p_{samp} = 1.0, and $k_{\mathrm{F-class}}$ = true for each of the three Blue River scenarios. A number of 100 particles was used for the PF, and Evo4DVar was allowed to try 100 configurations with c_1 = 0.01 and c_2 = 1.0. This combination of the balancing~~

~~constants c_1 and c_2 was found to yield good results for the Evo4DVar method given how the state variables outnumber the observations 812 to 14. Additionally, both the PF and Evo4DVar were seeded with the same set of states s_t^0 at the beginning of the assimilation period~~However, the number of particles should not be too small to ensure an appropriate sample size.

Table 5 shows the results of the year-long continuous forecasting experiment on the Blue River using a 50-particle PF and a configuration of OPTIMISTS with a 7-day assimilation time step Δt, both objectives, 50 particles, no optimization, and F-class kernels. Both the OPTIMISTS and PF methods show relatively good performance for all lead times (1, 3, 6, 12, and 24 days) based on both the deterministic and probabilistic performance metrics. However, OPTIMISTS generally outperforms the PF, especially for the longest lead times of 12 and 24 days. The errors with OPTIMISTS are usually smaller for longer lead times than the PF method, indicating that the longer Δt leads to reductions in overfitting. This is probably because particles that perform better over a wider time frame are more likely to be selected. Such a result also suggests that the search for an optimal Δt should consider the range of lead times that are deemed most critical for specific applications.

Figure 4~~Table 5 shows the corresponding performance metrics for these forecasts. It can be seen that the three methods perform similarly on the rainy autumn case (Scenario 1) and on the dry-wet winter case (Scenario 2), with OPTIMISTS displaying a modestly higher accuracy. On the other hand, Evo4DVar and especially the PF considerably overestimate the recession flow on the wet-dry summer case (Scenario 3), while OPTIMISTS is able to maintain a very reasonable performance. It can be hypothesized that the large observed peak during the assimilation period forced the PF and Evo4DVar to overestimate the moisture in the entire watershed, while the hybrid method proved better equipped to maintain the overall consistency of the ensemble and thus avoid a bad case of overfitting. We attribute this to the combined strengths of Bayesian sampling and the multi-objective, history-aware ranking of candidate particles; which again appears to outperform each of the two strategies if they are used in isolation.~~ shows the probabilistic streamflow forecasts for both algorithms for a lead time of 6 days and 24 days. The portrayed evolution of the density evidences the non-Gaussian nature of both estimates. While the behaviour of OPTIMISTS' forecasts of the low flow regime seems less stable in contrast with the PF's, its relative higher performance suggests that the estimates of the PF are overconfident and that OPTIMISTS' display a more sensible understanding of the associated uncertainty. These comparisons thus provide evidence showing that the combined features of Bayesian and variational data assimilation, if configured properly, effectively give OPTIMISTS an edge over traditional approaches.

4.2 Indiantown Run – high resolution application

Figure 5~~Figure 6~~ summarizes the changes in performance when using OPTIMISTS in Experiment 2. In this case, the more uniform forcing and streamflow conditions of the two scenarios allowed to statistically analyse all three performance metrics. For Scenario 1 we can see that OPTIMISTS produces a general increase in the Nash-Sutcliffe coefficients, but a decline in the MARE, evidencing tension between fitting the peaks and the inter-peak periods at the same time. For both scenarios there are configurations that performed very poorly, and we can look at the marginalized results in the boxplots for clues into which parameters might have caused this. Similar to the Blue River case, the use of a 1-hour time step significantly reduced the forecast skill, while the longer step almost always improved it; and the inclusion of the secondary history-consistency objective

("2 objectives") also resulted in improved performance. Not only does it seem that for this watershed the secondary objective mitigated the effects of overfitting, but it was interesting to note some configurations in which using it actually helped ~~achieving~~to achieve a better fit during the assimilation period.

While the ANOVA also provided evidence against the use of optimization algorithms, we are reluctant to instantly rule them out on the grounds that there were statistically significant interactions with other parameters (see the ANOVA table in the supplementary material). The optimizers led to poor results in cases with one-hour time steps or when only the first objective was used. Other statistically significant results point to the benefits of using the root samples more intensively (in opposition to using random samples) and, to a lesser extent, to the benefits of maintaining an ensemble of moderate size.

Figure 6~~Figure 7~~ shows the summarized changes in Experiment 3, where ~~we wanted to explore~~ the effect of the time step Δt is explored in ~~greater~~ detail. Once again, there ~~seems~~appears to be evidence favouring the hypothesis that there exists a sweet spot, and in this case it appears to be close to the two weeks mark: both shorter and longer time steps led to considerably poorer performance. ~~This time~~In this experiment, with all configurations using both optimization objectives, we can see that there are no clear disadvantages of using optimization algorithms (but also no advantages). Experiment 3 also shows that the effect of the greed parameter g is not very significant. That is, selecting some particles from dominated fronts to construct the target state distribution, and not only from the Pareto front, does not seem to affect the results.

With this information, we can select a preferred configuration of OPTIMISTS with a time step of two weeks, two objectives, 100 particles, no optimization, w_{root} = 95%, and g = 0.5. Figure 7~~Figure 8~~ shows the forecast comparisons between this configuration and the control open-loop model for scenarios 1 and 2. In both cases we see the control becoming too dry throughout, possibly because of recessions occurring faster than they should, at least during the period being studied. Assimilating streamflow data with OPTIMISTS leads to improvements in both cases. Also note that the performance metrics in Figure 7 (and in many of the results in all three scenario-based experiments) might make the error seem as being too large in the light of traditional standards, but this is justified given the very short time period of evaluation, and the ever-present effect of structural errors in the model applications—which in any case do not invalidate any of the results as all conditions also apply uniformly to the control runs and the PF method.

4.3 Computational performance

It is worth noting that the longer the assimilation time step, the faster the entire process is. This occurs because, even though the number of hydrological calculations is the same in the end, for every assimilation time step the model files need to be generated accordingly, then accessed, and finally the result files written and accessed. This whole process takes a considerable amount of time. Therefore, everything else being constant, sequential assimilation (like with PFs) automatically imposes additional computational requirements. In our tests we used RAM drive software to accelerate the process of running the models sequentially and, even then, the overhead imposed by OPTIMISTS was consistently below 10% of the total computation time. Most of the computational effort remained with running the model, both for VIC and the DHSVM. In this

sense, model developers may consider allowing their engines to be able to receive input data from main memory, if possible, to facilitate data assimilation and other similar processes.

4.4 Recommendations for configuring OPTIMISTS

Finally, here we summarize the recommended choices for the parameters in OPTIMISTS based on the results of the experiments. In the first place, given their low observed effect, default values can be used for g (around 0.5). A w_{root} higher than 90% was found to be advantageous. The execution of the optimization step ($p_{\text{samp}} < 1$) was, on the other hand, not found to be advantageous and, therefore, we consider it a cleaner approach to simply generate all samples from the initial distribution. Similarly, while not found to be disadvantageous, using diagonal bandwidth (D-class) kernels provide a significant improvement in computational efficiency and are thus recommended for the time being. Future work will be conducted to further explore the effect of the bandwidth configuration in OPTIMISTS.

Even though only two objective functions were tested, one measuring the departures from the observations being assimilated and another measuring the compatibility of initial samples with the initial distribution, the results clearly show that it is beneficial to simultaneously evaluate candidate particles using both criteria. While traditional cost functions like the one in Eq. (1) do indeed consider both aspects, we argue that that using multiple objectives has the added benefit of enriching the diversity of the particle ensemble and, ultimately, the resulting probabilistic estimate of the target states.

Our results demonstrated that the assimilation time step is the most sensitive parameter and, therefore, its selection must be done with the greatest involvement. Taken the results together, we recommend that multiple choices be tried for any new case study looking to strike a balance between the amount of information being assimilated and the number of degrees of freedom. This empirical selection should also be performed with a rough sense of what is the range of forecasting lead-times that is considered the most important. Lastly, more work is required to provide guidelines to select the number of particles n to be used. While the literature suggests that more should increase forecast accuracy, our tests did not back this conclusion. We tentatively recommend trying different ensemble sizes based on the computational resources available and selecting the one that offers the best observed trade-off between accuracy and efficiency.

5 Conclusions and future work

In this article we introduced OPTIMISTS, a flexible, model-independent data assimilation algorithm that effectively combines the signature elements from Bayesian and variational methods: By employing essential features from particle filters, it allows performing probabilistic non-Gaussian estimates of state variables through the filtering of a set of particles drawn from a prior distribution to better match the available observations. -Adding critical features from variational methods, OPTIMISTS grants its users the option of exploring the state space using optimization techniques and evaluating candidate states through a time window of arbitrary length. The algorithm fuses a multi-objective/Pareto analysis of candidate particles with kernel density probability distributions to effectively bridge the gap between the probabilistic and the variational perspectives. Moreover, the

use of evolutionary optimization algorithms enables its efficient application on highly non-linear models as those usually found in most ~~geophysical sciences.~~geosciences. This unique combination of features represent a clear differentiation from the existing hybrid assimilation methods in the literature (Bannister, 2016)~~(Bannister, 2016),~~ which ~~have mostly been developed around ensemble Kalman filters and convex optimization techniques (and therefore~~are limited to Gaussian distributions and linear dynamics~~).~~.

We conducted a set of hydrologic forecasting factorial experiments on two watersheds, the Blue River with 812 state variables and the Indiantown Run with 33,455, at two distinct modelling resolutions using two different modelling engines: VIC and the DHSVM, respectively. Capitalizing on the flexible configurations available for OPTIMISTS, these tests allowed to determine which individual characteristics of traditional algorithms prove to be the most advantageous for forecasting applications. For example, while there is a general consensus in the literature favouring extended time steps (4D) over sequential ones (1D-3D), the results from assimilating streamflow data in our experiments suggest that there is an ideal duration of the assimilation time step that is dependent on the case study under consideration ~~and~~, on the spatiotemporal resolution of the corresponding model application, and on the desired forecast length. Sequential time steps not only required considerably longer computational times but also produced the worst results—perhaps given the overwhelming number of degrees of freedom in contrast with the scarce observations available. Similarly, there was a drop in the performance of the forecast ensemble when the algorithm was set to use overly long time steps.

Procuring the consistency of candidate particles, not only with the observations but also with the history, led to significant gains in predictive skill. OPTIMISTS can be configured to both perform Bayesian sampling and find Pareto-optimal particles that trade-off deviations from the observations and from the prior conditions, a strategy that proved superior to those of traditional algorithms. This Bayesian/multi-objective formulation of the optimization problem was especially beneficial for the high-resolution watershed application, as it allows the model to overcome the risk of overfitting generated by the enlarged effect of equifinality.

On the other hand, our experiments did not produce enough evidence to recommend neither exploring the state space with optimization algorithms instead of doing so with simple probabilistic sampling, the use of a larger number of particles above the established baseline of 100, nor the computationally-intensive utilization of full covariance matrices to encode the dependencies between variables in the kernel-based state distributions. Nevertheless, strong interactions between several of these parameters suggest that some specific combinations could potentially yield strong outcomes. Together with OPTIMISTS' observed high level of sensitivity to the parameters, these results indicate that there could be promise in the implementation of self-adaptive strategies (Karafotias et al., 2014)~~(Karafotias et al., 2014)~~ to assist in their selection in the future. With these experiments, we were able to configure the algorithm to consistently improve the forecasting skill of the models compared to control open-loop runs. Additionally, a comparative test using the Blue River model application showed that OPTIMISTS was able to reliably produce adequate forecasts that were better than or similar ~~or better (with errors as much as 50% smaller) than~~to those resulting from assimilating the observations with a particle filter ~~and an evolutionary 4D variational method~~.

Moreover, in this article we offered several alternatives in the implementation of the components of OPTIMISTS whenever there were tensions between prediction accuracy and computational efficiency. In the future, we will focus on incorporating additional successful ideas from diverse assimilation algorithms and on improving components in such a way that both of these goals are attained with ever-smaller compromises. For instance, the estimation of initial states should not be overburdened with the responsibility of compensating structural and calibration deficiencies in the model. In this sense, we embrace the vision of a unified framework for the joint probabilistic estimation of structures, parameters, and state variables (Liu and Gupta, 2007)~~(Liu and Gupta, 2007), but we remain sceptical of~~, where it is important to address challenges associated with approaches that would increase the indeterminacy of the problem by adding unknowns without providing additional information or additional means of relating existing variables. We expect that with continued efforts OPTIMISTS will be a worthy candidate framework to be deployed in operational settings for hydrologic prediction and beyond.

Data and code availability

All the data utilized to construct the models is publicly available through the internet from their corresponding US government agencies' websites. The ~~executable for~~Java implementation of OPTIMISTS~~,~~ is available by request to the authors. The source code of the particle filter~~, and the evolutionary 4D variational algorithm are~~ is available through GitHub (https://github.com/felherc/~~(https://github.com/felherc/). The repositories~~). These sources include all the information needed to replicate the experiments in this article.

Acknowledgements

The authors are thankful to the two anonymous referees and the Editor for their valuable comments and suggestions. This work was supported in part by the United States Department of Transportation through award #OASRTRS-14-H-PIT to the University of Pittsburgh and by the William Kepler Whiteford Professorship from the University of Pittsburgh.

Previous versions

An earlier version of this article was submitted for peer review to be considered for publication on HESS and is available in the HESSD archive (Hernández and Liang, 2016). While the proposed method itself has seen no changes since, this new version attempts to make its presentation much more approachable and has included the comparative tests with the particle filter.

References

Adams, R. M., Houston, L. L., McCarl, B. A., Tiscareño, M. L., Matus, J. G. and Weiher, R. F.: The benefits to Mexican

agriculture of an El Niño-southern oscillation (ENSO) early warning system, Agric. For. Meteorol., 115(3–4), 183–194, doi:10.1016/S0168-1923(02)00201-0, 2003.

Andreadis, K. M. and Lettenmaier, D. P.: Assimilating remotely sensed snow observations into a macroscale hydrology model, Adv. Water Resour., 29(6), 872–886, doi:10.1016/j.advwatres.2005.08.004, 2006.

Bannister, R. N.: A review of forecast error covariance statistics in atmospheric variational data assimilation. II: Modelling the forecast error covariance statistics, Q. J. R. Meteorol. Soc., 134(637), 1971–1996, doi:10.1002/qj.340, 2008.

Bannister, R. N.: A review of operational methods of variational and ensemble-variational data assimilation, Q. J. R. Meteorol. Soc., 29(January), 1–29, doi:10.1002/QJ.2982, 2016.

Beven, K.: A manifesto for the equifinality thesis, J. Hydrol., 320(1–2), 18–36, doi:10.1016/j.jhydrol.2005.07.007, 2006.

Bröcker, J.: Evaluating raw ensembles with the continuous ranked probability score, Q. J. R. Meteorol. Soc., 138(667), 1611–1617, doi:10.1002/qj.1891, 2012.

Buehner, M., Houtekamer, P. L., Charette, C., Mitchell, H. L. and He, B.: Intercomparison of Variational Data Assimilation and the Ensemble Kalman Filter for Global Deterministic NWP. Part II: One-Month Experiments with Real Observations, Mon. Weather Rev., 138(5), 1567–1586, doi:10.1175/2009MWR3158.1, 2010.

Carpenter, J., Clifford, P. and Fearnhead, P.: Improved particle filter for nonlinear problems, IEE Proc. - Radar, Sonar Navig., 146(1), 2, doi:10.1049/ip-rsn:19990255, 1999.

Clark, M. P., Rupp, D. E., Woods, R. a., Zheng, X., Ibbitt, R. P., Slater, A. G., Schmidt, J. and Uddstrom, M. J.: Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model, Adv. Water Resour., 31(10), 1309–1324, doi:10.1016/j.advwatres.2008.06.005, 2008.

Clark, M. P., Bierkens, M. F. P. P., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., Pauwels, V. R. N. N., Cai, X., Wood, A. W., Peters-Lidard, C. D., Uijenhoet, R., Bennet, K. E., Pauwels, V. R. N. N., Cai, X., Wood, A. W. and Peters-Lidard, C. D.: The evolution of process-based hydrologic models: Historical challenges and the collective quest for physical realism, Hydrol. Earth Syst. Sci. Discuss., 21(January), 1–14, doi:10.5194/hess-2016-693, 2017.

Cosgrove, B. A., Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., Marshall, C., Sheffield, J., Duan, Q., Luo, L., Higgins, R. W., Pinker, R. T., Tarpley, J. D. and Meng, J.: Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project, J. Geophys. Res. Atmos., 108(D22), doi:10.1029/2002JD003118, 2003.

Crainic, T. G. and Toulouse, M.: Parallel Meta-heuristics, in Handbook of Metaheuristics, vol. 146, edited by M. Gendreau and J.-Y. Potvin, pp. 497–541, Springer US., 2010.

Deb, K.: Multi-objective Optimization, in Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques, edited by E. K. Burke and G. Kendall, pp. 403–449, Springer US., 2014.

Deb, K., Pratap, A., Agarwal, S. and Meyarivan, T.: A fast and elitist multiobjective genetic algorithm: NSGA-II, Evol. Comput. IEEE Trans., 6(2), 182–197, doi:10.1109/4235.996017, 2002.

Desroziers, G., Camino, J. T. and Berre, L.: 4DEnVar: Link with 4D state formulation of variational assimilation and different possible implementations, Q. J. R. Meteorol. Soc., 140(684), 2097–2110, doi:10.1002/qj.2325, 2014.

Dumedah, G. and Coulibaly, P.: Evolutionary assimilation of streamflow in distributed hydrologic modeling using in-situ soil moisture data, Adv. Water Resour., 53, 231–241, doi:10.1016/j.advwatres.2012.07.012, 2013.

Dumedah, G., Berg, A. a. and Wineberg, M.: An Integrated Framework for a Joint Assimilation of Brightness Temperature and Soil Moisture Using the Nondominated Sorting Genetic Algorithm II, J. Hydrometeorol., 12(6), 1596–1609, doi:10.1175/JHM-D-10-05029.1, 2011.

Duong, T. and Hazelton, M. L.: Cross-validation bandwidth matrices for multivariate kernel density estimation, Scand. J. Stat., 32, 485–506, doi:10.1111/j.1467-9469.2005.00445.x, 2005.

Efstratiadis, A. and Koutsoyiannis, D.: One decade of multi-objective calibration approaches in hydrological modelling: a

review, Hydrol. Sci. J., 55(February 2015), 58–78, doi:10.1080/02626660903526292, 2010.

Errico, R. M.: What Is an Adjoint Model?, Bull. Am. Meteorol. Soc., 78(11), 2577–2591, doi:10.1175/1520-0477(1997)078<2577:WIAAM>2.0.CO;2, 1997.

Evensen, G.: Data assimilation: the ensemble Kalman filter, Springer Science & Business Media., 2009.

Evensen, G. and van Leeuwen, P. J.: An ensemble Kalman smoother for nonlinear dynamics, Mon. Weather Rev., 128(6), 1852–1867, doi:10.1175/1520-0493(2000)128<1852:AEKSFN>2.0.CO;2, 2000.

Fisher, M.: Background error covariance modelling, Semin. Recent Dev. Data Assim. …, 45–63 [online] Available from: ftp://beryl.cerfacs.fr/pub/globc/exchanges/daget/DOCS/sem2003_fisher.pdf%5Cnpapers2://publication/uuid/265DCC42-4A9C-482A-B2CF-74866FCC2312, 2003.

Friedman, J., Hastie, T. and Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso, Biostatistics, 9(3), 432–441, doi:10.1093/biostatistics/kxm045, 2008.

Gauthier, P., Tanguay, M., Laroche, S., Pellerin, S. and Morneau, J.: Extension of 3DVAR to 4DVAR: Implementation of 4DVAR at the Meteorological Service of Canada, Mon. Weather Rev., 135(6), 2339–2354, doi:10.1175/MWR3394.1, 2007.

Ghil, M. and Malanotte-Rizzoli, P.: Data assimilation in meteorology and oceanography, Adv. Geophys, 33, 141–266, doi:10.1016/S0065-2687(08)60442-2, 1991.

Ghorbanidehno, H., Kokkinaki, A., Li, J. Y., Darve, E. and Kitanidis, P. K.: Real-time data assimilation for large-scale systems: The spectral Kalman filter, Adv. Water Resour., 86, 260–272, doi:10.1016/j.advwatres.2015.07.017, 2015.

Gordon, N. J., Salmond, D. J. and Smith, A. F. M.: Novel approach to nonlinear/non-Gaussian Bayesian state estimation, IEE Proc. F Radar Signal Process., 140(2), 107, doi:10.1049/ip-f-2.1993.0015, 1993.

Guo, J., Liang, X. and Leung, L. R.: A new multiscale flow network generation scheme for land surface models, Geophys. Res. Lett., 31(23), 1–4, doi:10.1029/2004GL021381, 2004.

Haario, H., Saksman, E. and Tamminen, J.: An Adaptive Metropolis Algorithm, Bernoulli, 7(2), 223–242, doi:10.2307/3318737, 2001.

Hawkins, D. M.: The Problem of Overfitting, J. Chem. Inf. Comput. Sci., 44(1), 1–12, doi:10.1021/ci0342472, 2004.

Hazelton, M. L.: Variable kernel density estimation, Aust. N. Z. J. Stat., 45(3), 271–284, doi:10.1111/1467-842X.00283, 2003.

Hernández, F. and Liang, X.: Hybridizing sequential and variational data assimilation for robust high-resolution hydrologic forecasting, Hydrol. Earth Syst. Sci. Discuss., (September), 1–25, doi:10.5194/hess-2016-454, 2016.

Homer, C., Fry, J. and Barnes, C.: The National Land Cover Database, US Geol. Surv. Fact Sheet, 3020(February), 1–4 [online] Available from: http://pubs.usgs.gov/fs/2012/3020/, 2012.

Houser, P. R., Shuttleworth, W. J., Famiglietti, J. S., Gupta, H. V, Syed, K. H. and Goodrich, D. C.: Integration of soil moisture remote sensing and hydrologic modeling using data assimilation, Water Resour. Res., 34(12), 1998.

Karafotias, G., Hoogendoorn, M. and Eiben, ~~a~~A. E.: Parameter Control in Evolutionary Algorithms: Trends and Challenges, IEEE Trans. Evol. Comput., to appear,(2), 167–187, doi:10.1109/TEVC.2014.2308294, 2014.

Koster, R. D., Betts, A. K., Dirmeyer, P. A., Bierkens, M., Bennett, K. E., Déry, S. J., Evans, J. P., Fu, R., Hernandez, F., Leung, L. R., Liang, X., Masood, M., Savenije, H., Wang, G. and Yuan, X.: Hydroclimatic variability and predictability: a survey of recent research, Hydrol. Earth Syst. Sci., 21(7), 3777–3798, doi:10.5194/hess-21-3777-2017, 2017.

Krause, P., Boyle, D. P. and Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci., 5(89), 89–97, doi:10.5194/adgeo-5-89-2005, 2005.

Krishnamoorthy, A. and Menon, D.: Matrix Inversion Using Cholesky Decomposition, CoRR, (3), 10–12 [online] Available from: http://arxiv.org/abs/1111.4144, 2011.

van Leeuwen, P. J.: Particle Filtering in Geophysical Systems, Mon. Weather Rev., 137(12), 4089–4114, doi:10.1175/2009MWR2835.1, 2009.

van Leeuwen, P. J.: Nonlinear Data Assimilation for high-dimensional systems, in Nonlinear Data Assimilation, edited by J. P. Van Leeuwen, Y. Cheng, and S. Reich, pp. 1–73, Springer International Publishing., 2015.

Li, J. Y., Kokkinaki, A., Ghorbanidehno, H., Darve, E. F. and Kitanidis, P. K.: The compressed state Kalman filter for nonlinear state estimation: Application to large-scale reservoir monitoring, Water Resour. Res., 51(12), 9942–9963, doi:10.1002/2015WR017203, 2015.

Liang, X. and Xie, Z.: A new surface runoff parameterization with subgrid-scale soil heterogeneity for land surface models, Adv. Water Resour., 24(9), 1173–1193, 2001.

Liang, X. and Xie, Z.: Important factors in land–atmosphere interactions: surface runoff generations and interactions between surface and groundwater, Glob. Planet. Change, 38(1), 101–114, 2003.

Liang, X., Lettenmaier, D. P., Wood, E. F. and Burges, S. J.: A simple hydrologically based model of land surface water and energy fluxes for general circulation models, J. Geophys. Res., 99(D7), 14415, doi:10.1029/94JD00483, 1994.

Liang, X., Lettenmaier, D. P. and Wood, E. F.: One-dimensional statistical dynamic representation of subgrid spatial variability of precipitation in the two-layer variable infiltration capacity model, J. Geophys. Res. Atmos., 101(D16), 21403–21422, 1996a.

Liang, X., Wood, E. F. and Lettenmaier, D. P.: Surface soil moisture parameterization of the VIC-2L model: Evaluation and modification, Glob. Planet. Change, 13(1), 195–206, 1996b.

Liu, J. S. and Chen, R.: Sequential Monte Carlo Methods for Dynamic Systems, J. Am. Stat. Assoc., 93(443), 1032–1044, doi:10.2307/2669847, 1998.

Liu, Y. and Gupta, H. V.: Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, Water Resour. Res., 43(7), 1–18, doi:10.1029/2006WR005756, 2007.

Lohmann, D., Rashke, E., Nijssen, B. and Lettenmaier, D. P.: Regional scale hydrology: I. Formulation of the VIC-2L model coupled to a routing model, Hydrol. Sci. J., 43(1), 131–141, doi:10.1080/02626669809492107, 1998.

Lorenc, A. C., Bowler, N. E., Clayton, A. M., Pring, S. R. and Fairbairn, D.: Comparison of Hybrid-4DEnVar and Hybrid-4DVar Data Assimilation Methods for Global NWP, Mon. Weather Rev., 143(1), 212–229, doi:10.1175/MWR-D-14-00195.1, 2015.

Mahalanobis, P. C.: On the generalized distance in statistics, Proc. Natl. Inst. Sci., 2, 49–55, 1936.

Miller, D. A. and White, R. A.: A Conterminous United States Multilayer Soil Characteristics Dataset for Regional Climate and Hydrology Modeling, Earth Interact., 2(1), 2–2, doi:10.1175/1087-3562(1998)002<0002:CUSMS>2.0.CO;2, 1998.

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T.: The ECMWF ensemble prediction system: Methodology and validation, Q. J. R. Meteorol. Soc., 122(529), 73–119, doi:10.1002/qj.49712252905, 1996.

Montgomery, D. C.: Design and analysis of experiments, Eigth edit., John Wiley & Sons., 2012.

Montgomery, D. C., Runger, G. C. and Hubele, N. F.: Engineering statistics, John Wiley & Sons., 2009.

Montzka, C., Pauwels, V. R. N., Franssen, H.-J. H., Han, X. and Vereecken, H.: Multivariate and Multiscale Data Assimilation in Terrestrial Systems: A Review, Sensors, 12(12), 16291–16333, doi:10.3390/s121216291, 2012.

Moradkhani, H., DeChant, C. M. and Sorooshian, S.: Evolution of ensemble data assimilation for uncertainty quantification using the particle filter-Markov chain Monte Carlo method, Water Resour. Res., 48(12), doi:10.1029/2012WR012144, 2012.

Ning, L., Carli, F. P., Ebtehaj, A. M., Foufoula-Georgiou, E. and Georgiou, T. T.: Coping with model error in variational data assimilation using optimal mass transport, Water Resour. Res., 50(7), 5817–5830, doi:10.1002/2013WR014966, 2014.

Noh, S. J., Tachikawa, Y., Shiiba, M. and Kim, S.: Applying sequential Monte Carlo methods into a distributed hydrologic

model: Lagged particle filtering approach with regularization, Hydrol. Earth Syst. Sci., 15(10), 3237–3251, doi:10.5194/hess-15-3237-2011, 2011.

Park, S., Hwang, J. P., Kim, E. and Kang, H. J.: A new evolutionary particle filter for the prevention of sample impoverishment, IEEE Trans. Evol. Comput., 13(4), 801–809, doi:10.1109/TEVC.2008.2011729, 2009.

Penning-Rowsell, E. C., Tunstall, S. M., Tapsell, S. M. and Parker, D. J.: The benefits of flood warnings: Real but elusive, and politically significant, J. Chart. Inst. Water Environ. Manag., 14(1), 7–14, doi:10.1111/j.1747-6593.2000.tb00219.x, 2000.

Pham, D. T.: Stochastic methods for sequential data assimilation in strongly nonlinear systems, Mon. Weather Rev., 129(5), 1194–1207, doi:10.1175/1520-0493(2001)129<1194:SMFSDA>2.0.CO;2, 2001.

Rawlins, F., Ballard, S. P., Bovis, K. J., Clayton, A. M., Li, D., Inverarity, G. W., Lorenc, A. C. and Payne, T. J.: The Met Office global four-dimensional variational data assimilation scheme, Q. J. R. Meteorol. Soc., 133(623), 347–362, doi:10.1002/qj.32, 2007.

Reichle, R. H., McLaughlin, D. B. and Entekhabi, D.: Variational data assimilation of microwave radiobrightness observations for land surface hydrology applications, Geosci. Remote Sensing, IEEE Trans., 39(8), 1708–1718, doi:10.1109/36.942549, 2001.

Rodríguez, E., Morris, C. S. and Belz, J. E.: A global assessment of the SRTM performance, Photogramm. Eng. Remote Sens., 72(3), 249–260, 2006.

Rogers, E., DiMego, G., Black, T., Ek, M., Ferrier, B., Gayno, G., Janic, Z., Lin, Y., Pyle, M., Wong, V. and Wu, W.-S.: The NCEP North American Mesoscale Modeling System: Recent Changes and Future Plans, 23rd Conf. Weather Anal. Forecast. Conf. Numer. Weather Predict., (1995) [online] Available from: http://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154114.htm, 2009.

Seaber, P. R., Kapinos, F. P. and Knapp, G. L.: Hydrologic unit maps, US Government Printing Office Washington, DC, USA., 1987.

Sheather, S. J. and Jones, M. C.: A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation, J. R. Stat. Soc. Ser. B, 53(3), 683–690, doi:10.2307/2345597, 1991.

Silverman, B. B. W.: Density estimation for statistics and data analysis, CRC press., 1986.

Smith, A., Doucet, A., de Freitas, N. and Gordon, N.: Sequential Monte Carlo methods in practice, Springer Science & Business Media., 2013.

Snyder, C., Bengtsson, T., Bickel, P. and Anderson, J.: Obstacles to High-Dimensional Particle Filtering, Mon. Weather Rev., 136(12), 4629–4640, doi:10.1175/2008MWR2529.1, 2008.

Socha, K. and Dorigo, M.: Ant colony optimization for continuous domains, Eur. J. Oper. Res., 185(3), 1155–1173, doi:10.1016/j.ejor.2006.06.046, 2008.

Terrell, G. R. and Scott, D. W.: Variable kernel density estimation, Ann. Stat., 1236–1265, doi:10.1214/aos/1176348768, 1992.

Trémolet, Y.: Accounting for an imperfect model in 4D-Var, Q. J. R. Meteorol. Soc., 132(621), 2483–2504, doi:10.1256/qj.05.224, 2006.

Verkade, J. S. and Werner, M. G. F.: Estimating the benefits of single value and probability forecasting for flood warning, Hydrol. Earth Syst. Sci., 15(12), 3751–3765, doi:10.5194/hess-15-3751-2011, 2011.

Vrugt, J. A. and Robinson, B. A.: Improved evolutionary optimization from genetically adaptive multimethod search, Proc. Natl. Acad. Sci., 104(3), 708–711, doi:10.1073/pnas.0610471104, 2007.

~~Vrugt, J. A., Robinson, B. A. and Hyman, J. M.: Self-Adaptive Multimethod Search for Global Optimization in Real-Parameter Spaces, Evol. Comput. IEEE Trans., 13(2), 243–259, doi:10.1109/TEVC.2008.924428, 2009.~~

Wand, M. P. and Jones, M. C.: Kernel smoothing, Crc Press., 1994.

Wen, Z., Liang, X. and Yang, S.: A new multiscale routing framework and its evaluation for land surface modeling

applications, , 48(June), 1–16, doi:10.1029/2011WR011337, 2012.

West, M.: Mixture models, Monte Carlo, Bayesian updating, and dynamic models, Comput. Sci. Stat., 1–11 [online] Available from: http://www.stat.duke.edu/~mw/MWextrapubs/West1993a.pdf, 1993.

Wigmosta, M., Nijssen, B. and Storck, P.: The distributed hydrology soil vegetation model, Math. Model. Small Watershed Hydrol. Appl., 7–42 [online] Available from: http://www.cabdirect.org/abstracts/20033121322.html, 2002.

Wigmosta, M. S., Vail, L. W. and Lettenmaier, D. P.: A distributed hydrology-vegetation model for complex terrain, Water Resour. Res., 30(6), doi:10.1029/94WR00436, 1994.

Yang, S.-C., Corazza, M., Carrassi, A., Kalnay, E. and Miyoshi, T.: Comparison of Local Ensemble Transform Kalman Filter, 3DVAR, and 4DVAR in a Quasigeostrophic Model, Mon. Weather Rev., 137(2), 693–709, doi:10.1175/2008MWR2396.1, 2009.

Zhang, F., Zhang, M. and Hansen, J.: Coupling ensemble Kalman filter with four-dimensional variational data assimilation, Adv. Atmos. Sci., 26(1), 1–8, doi:10.1007/s00376-009-0001-8.1.Introduction, 2009.

Zhang, L., Nan, Z., Liang, X., Xu, Y., Hernández, F. and Li, L.: Application of the MacCormack scheme to overland flow routing for high-spatial resolution distributed hydrological model, J. Hydrol., 558, 421–431, doi:10.1016/j.jhydrol.2018.01.048, 2018.

Zhang, X., Tian, Y., Cheng, R. and Jin, Y.: An Efficient Approach to Nondominated Sorting for Evolutionary Multiobjective Optimization, Evol. Comput. IEEE Trans., 19(2), 201–213, doi:10.1109/TEVC.2014.2308305, 2015.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D. and Mylne, K.: The economic value of ensemble-based weather forecasts, Bull. Am. Meteorol. Soc., 83(1), 73–83, doi:10.1175/1520-0477(2002)083<0073:TEVOEB>2.3.CO;2, 2002.

Ziervogel, G., Bithell, M., Washington, R. and Downing, T.: Agent-based social simulation: A method for assessing the impact of seasonal climate forecast applications among smallholder farmers, Agric. Syst., 83(1), 1–26, doi:10.1016/j.agsy.2004.02.009, 2005.

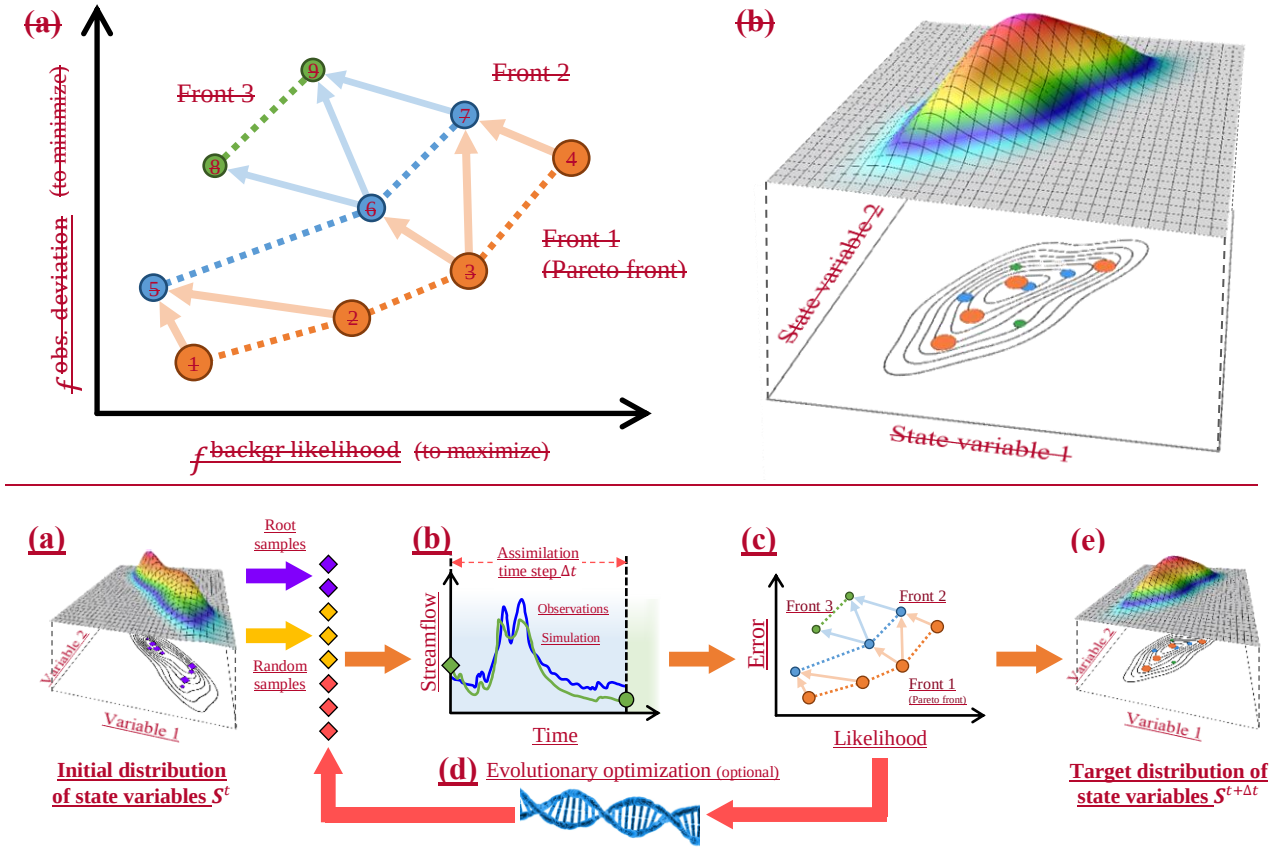
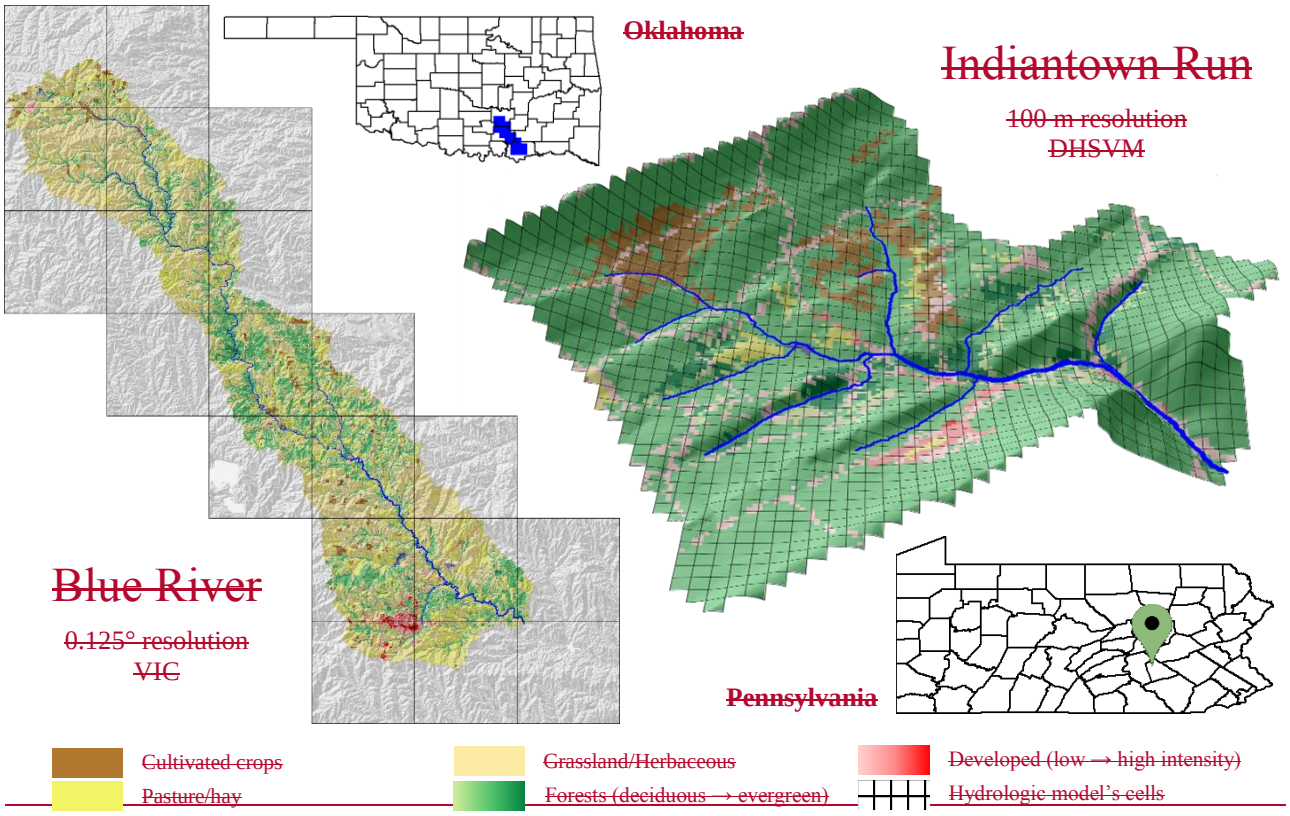


Figure 1~~1. (a) Nine labelled~~. Steps in OPTIMISTS, to be repeated for each assimilation time step Δt. In this example state vectors have two variables, observations are of streamflow, and particles ~~arranged~~are judged using two user-selected objectives: the likelihood given S^t to be maximized and the error given the observations to be minimized. (a) Initial state kernel density distribution S^t from which root samples (purple rhombi) are taken during the drawing step and random samples (yellow rhombi) are taken during the sampling step. (b) Execution of the model (simulation step) for each source sample for a time equal to Δt to compute output variables (for comparison with observations) and target samples (circles). (c) Evaluation of each particle (evaluation step) based on the objectives and organization into ~~three~~ non-domination fronts ~~on a scatterplot of two filtering objectives (minimum deviation from observations and maximum likelihood given background).~~(ranking step). The dashed lines represent the fronts while the arrows denote domination relationships between particles in adjacent fronts. ~~Particles in highly ranked fronts are assigned larger weights w_i (represented by particle size). (b) Probability density (or likelihood) of an example two-dimensional state distribution~~ (d) Optional optimization step which can be executed several times and that uses a population-based evolutionary optimization algorithm to generate additional samples (red rhombi). (e) Target state kernel density distribution $S^{t+\Delta t}$ constructed from the ~~same nine particles using multivariate~~particles' final samples (circles) after being weighted ~~kernel density estimation. Kernels~~according to the rank of their front (weighting step): kernels centred on ~~particles~~samples with higher weight (shown larger) have a higher probability density contribution.


Oklahoma
Indiantown Run
100 m resolution
DHSVM
Blue River
0.125° resolution
VIC
Pennsylvania
Cultivated crops
Pasture/hay
Grassland/Herbaceous
Forests (deciduous → evergreen)
Developed (low → high intensity)
Hydrologic model's cells

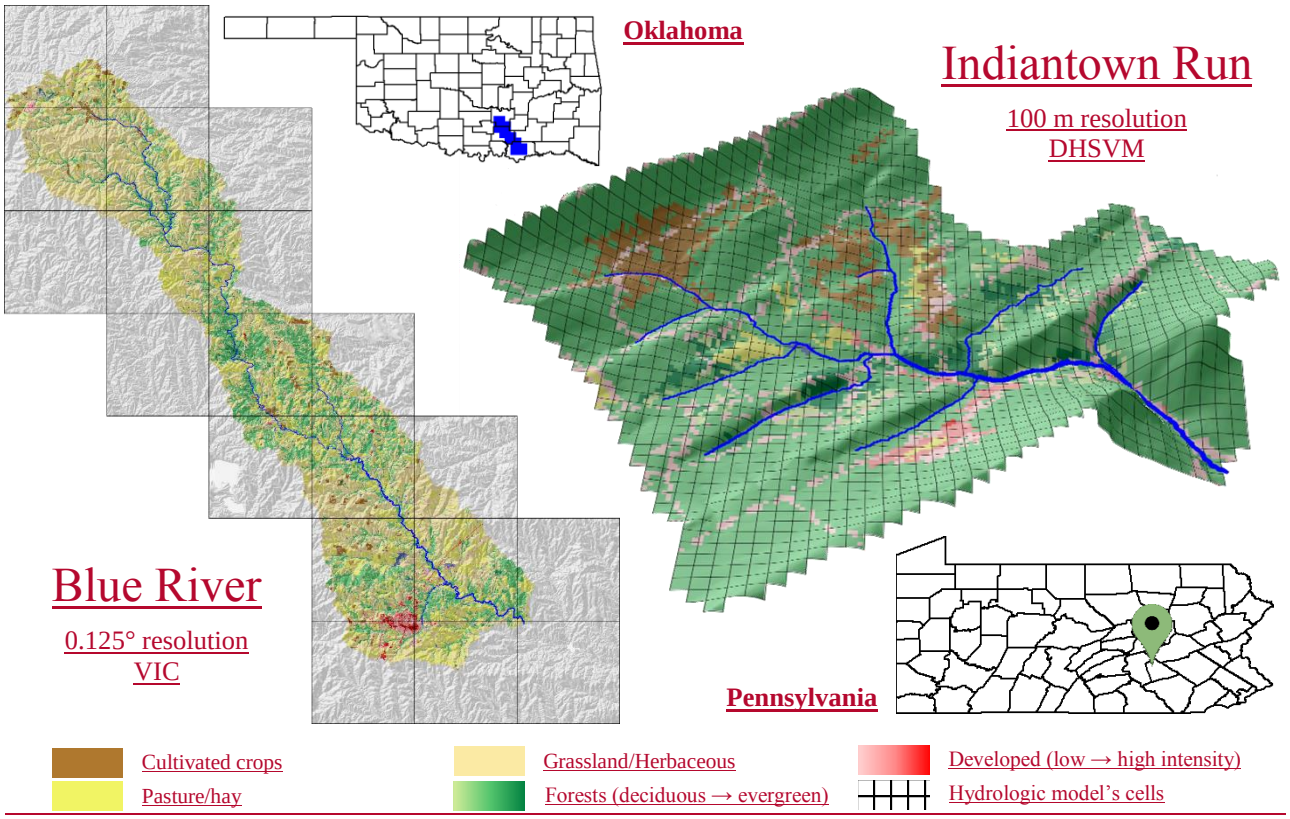
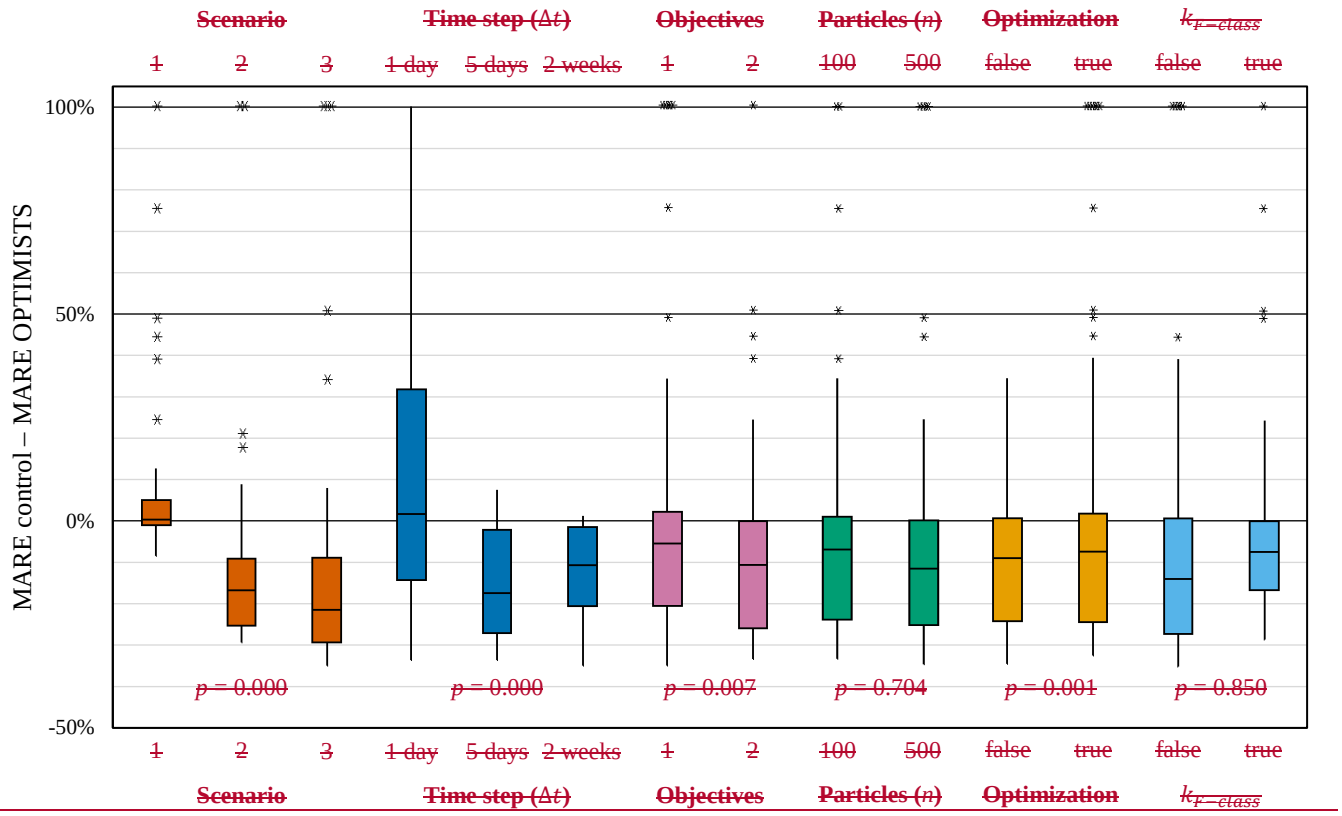


Figure 2~~2~~. Maps of the two test watersheds in the United States displaying the 30 m resolution land cover distribution from the NLCD (Homer et al., 2012)~~(Homer et al., 2012)~~. Left: Oklahoma's Blue River watershed 0.125° resolution VIC model (20 cells). Right: Pennsylvania's Indiantown Run watershed 100 m-resolution DHSVM model (1,472 cells).

5


Scenario
Time step (Δt)
Objectives
Particles (n)
Optimization
$k_{t-class}$
1
2
3
1 day
5 days
2 weeks
1
2
100
500
false
true
false
true
100%
50%
0%
-50%
MARE control – MARE OPTIMISTS
p = 0.000
p = 0.000
p = 0.007
p = 0.704
p = 0.001
p = 0.850

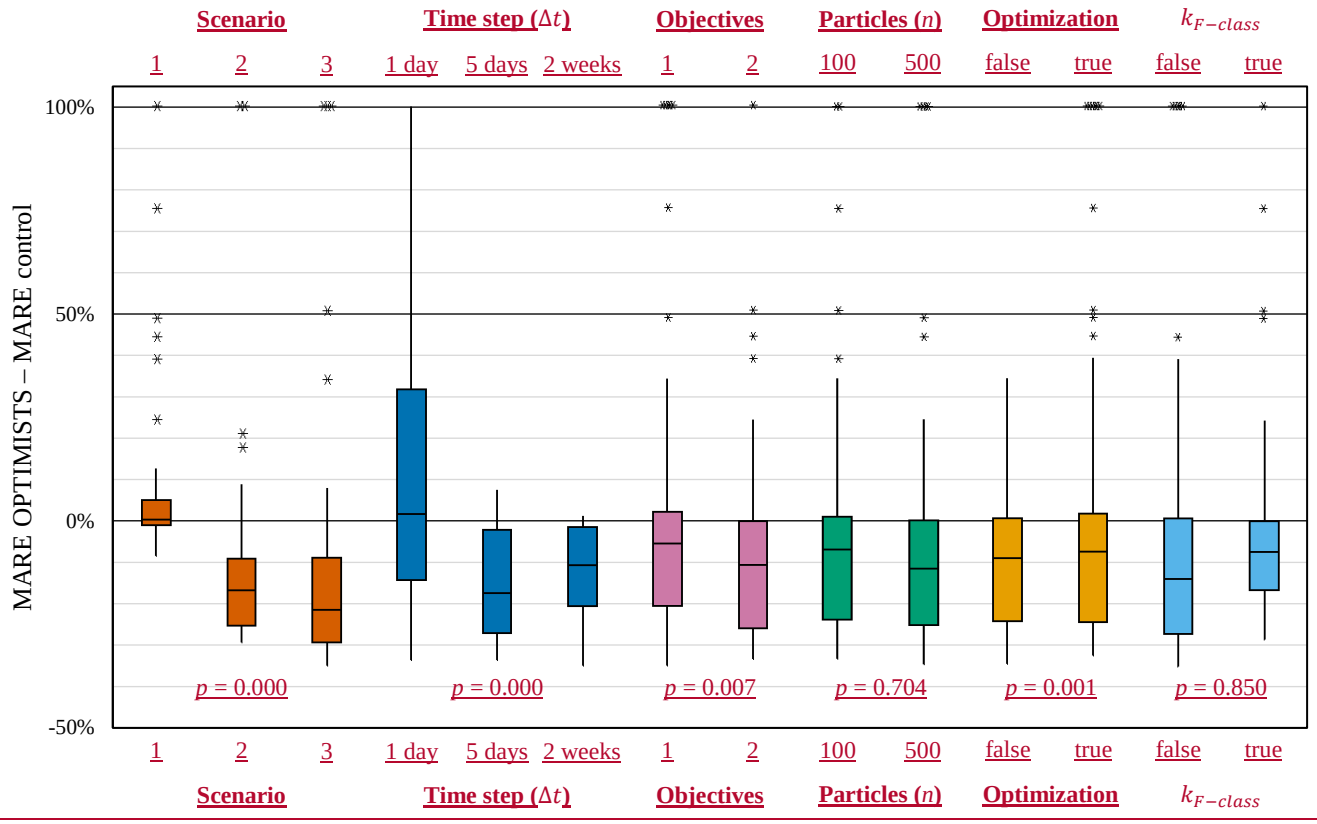
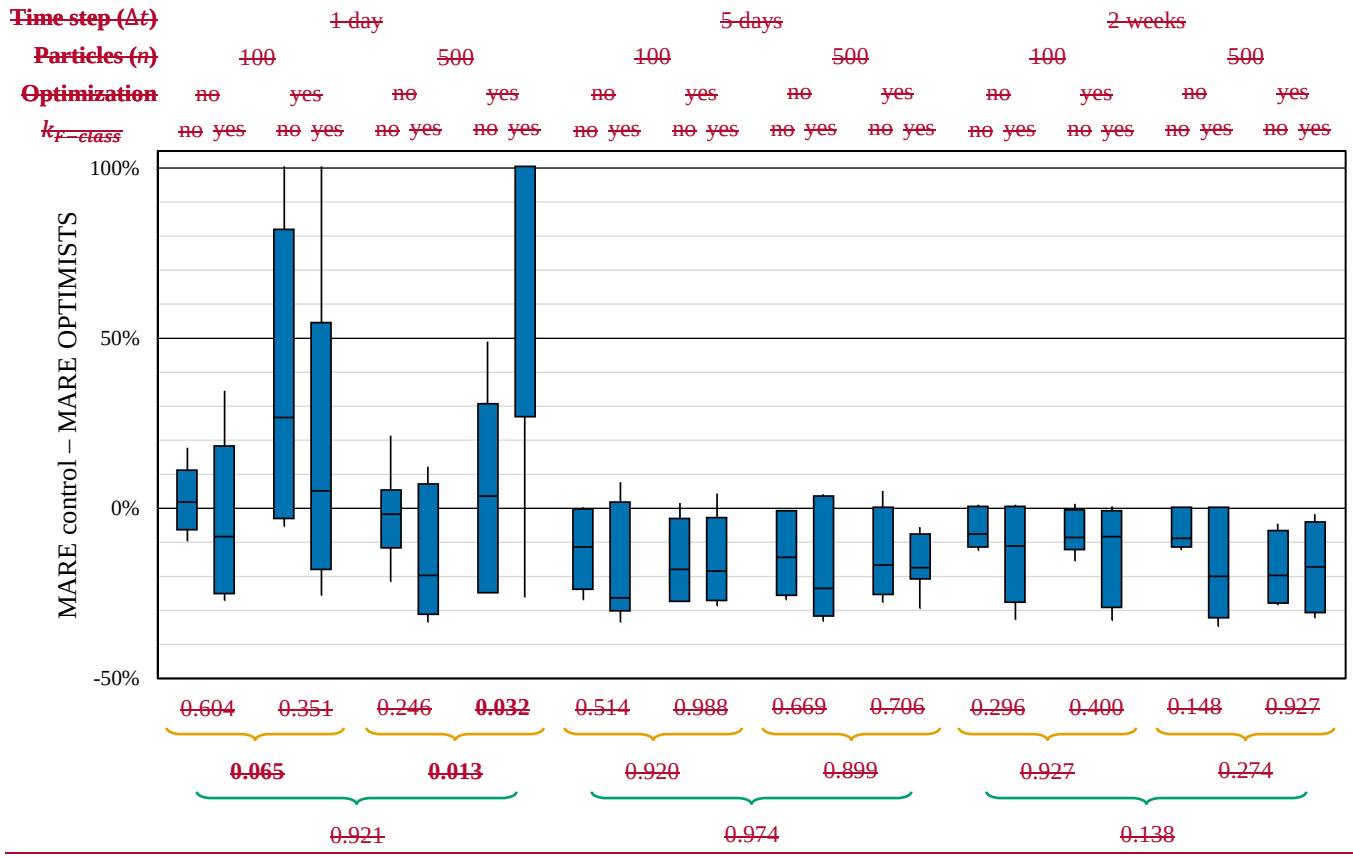


Figure 3~~3.~~. Boxplots of the changes in forecasting error (MARE) achieved while using OPTIMISTS on Experiment 1 (Blue River). Changes are relative to an open-loop control run where no assimilation was performed. Each column corresponds to the distribution of the error changes on the specified scenario or assignment to the indicated parameter. Positive values indicate that OPTIMISTS increased the error, while negative values indicate it decreased the error. Outliers are noted as asterisks and values were limited to 100%. For the one-objective case the particles' MAE was to be minimized; for the two-objective case, the likelihood given the background was to be maximized in addition. No optimization ("false") corresponds to p_{samp} = 1.0 (i.e., all samples are obtained from the prior distribution); "true" corresponds to p_{samp} = 0.25. The p-values were determined using ANOVA (Montgomery, 2012)~~(Montgomery, 2012),~~ and indicate the probability that the differences in means corresponding to boxes of the same colour are produced by chance (e.g., values close to zero indicate certainty that the parameter effectively affects the forecast error).


Time step (Δt)
1 day
5 days
2 weeks
Particles (n)
100
500
100
500
100
500
Optimization
no
yes
no
yes
no
yes
no
yes
no
yes
no
yes
$k_{1-class}$
no yes no yes no yes no yes no yes no yes no yes no yes no yes no yes no yes no yes
100%
50%
0%
-50%
MARE control – MARE OPTIMISTS
0.604 0.351 0.246 0.032 0.514 0.988 0.669 0.706 0.296 0.400 0.148 0.927
0.065 0.013 0.920 0.899 0.927 0.274
0.921 0.974 0.138

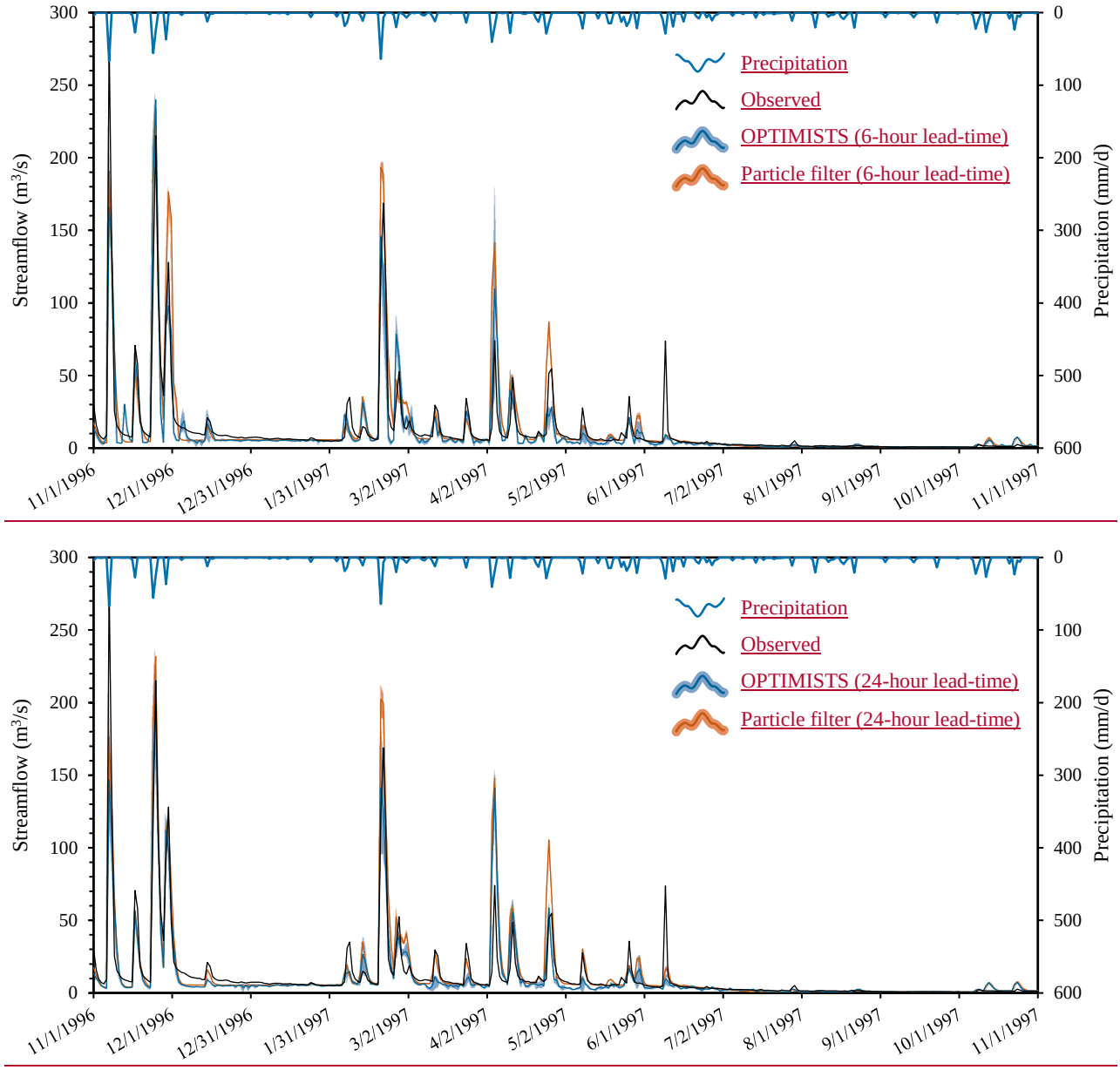
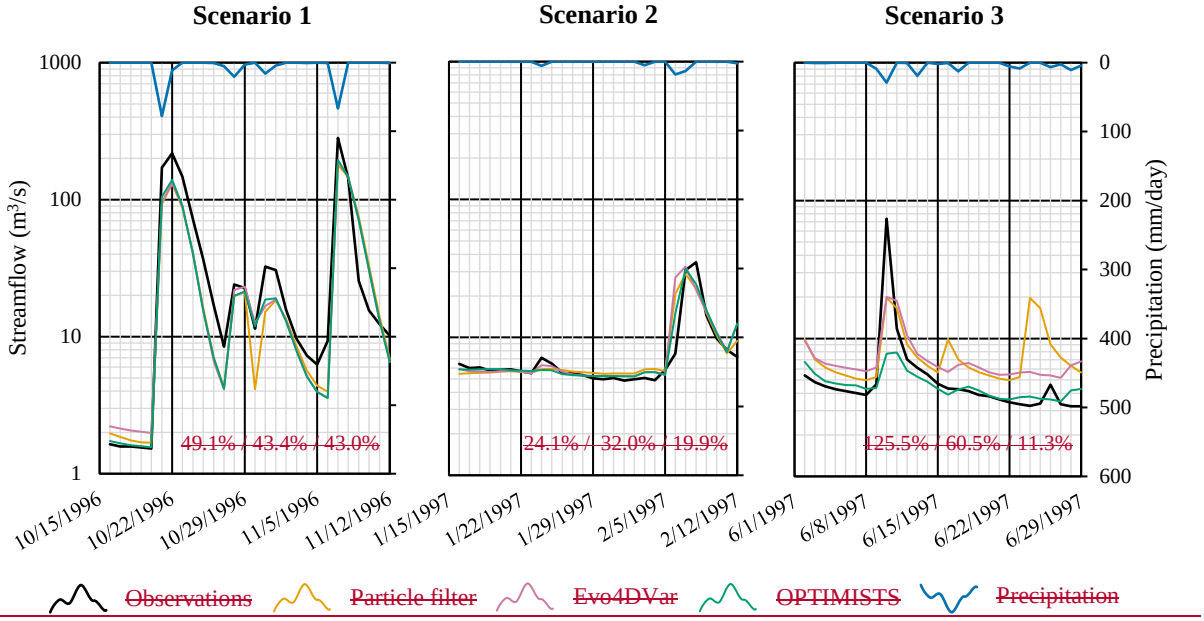
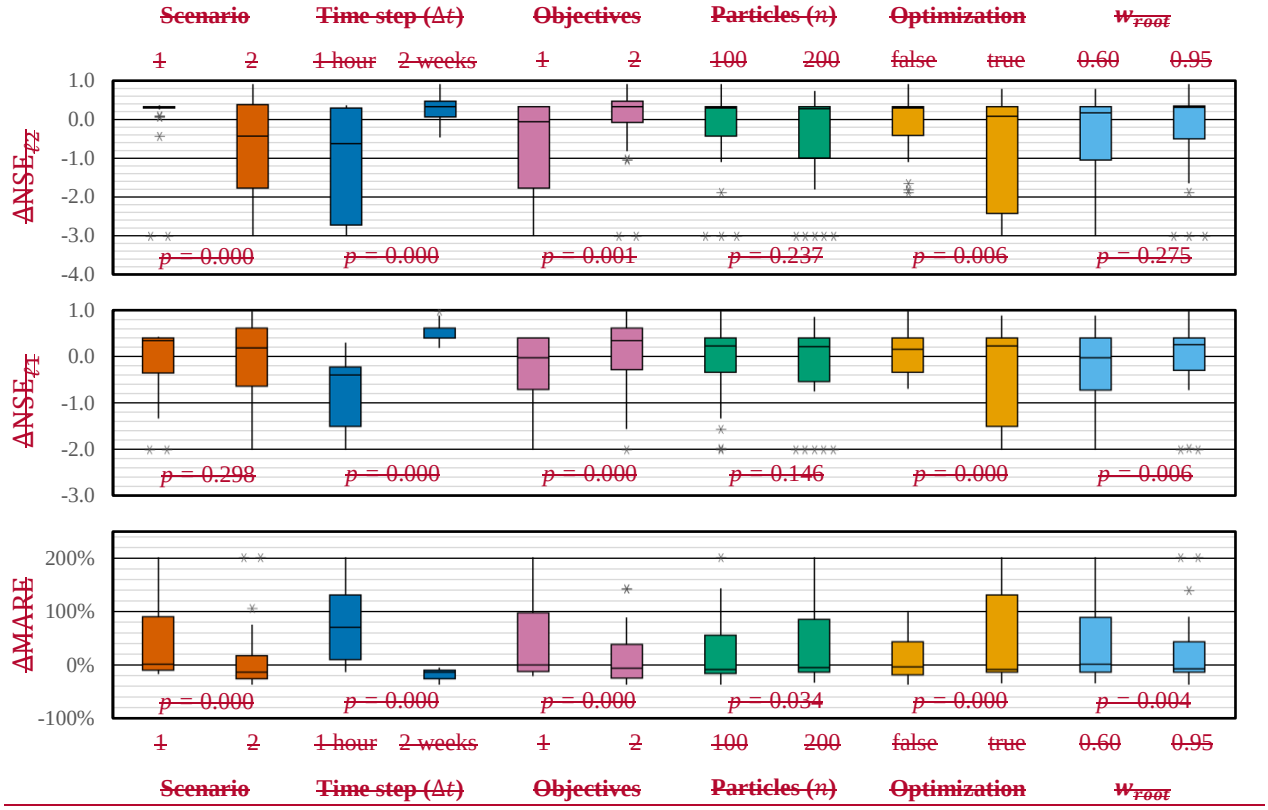


Figure 4~~4. Results of change in forecast MARE in Experiment 1 (Blue River) with different configurations of OPTIMISTS. Changes are relative to an open-loop control run where no assimilation was performed. Each column corresponds to the distribution of the error changes for the specified set of parameter assignments. Each box and pair of whiskers represent six cases (three scenarios and one or two objectives). Positive values indicate that OPTIMISTS increased the error, while negative values indicate it decreased the error. The values beneath the graph represent *p*-values from two-sample *t*-tests between the two indicated columns or group of columns (*p*-values smaller than 0.1 are shown in boldface).~~



~~Figure 5.~~. **Comparison of ~~hydrographs produced by three different data assimilation methods for each of the three Blue River scenarios. The methods compared are the PF, Evo4DVar (c_1 = 0.01 and c_2 = 1.0), and OPTIMISTS (Δt = 5~~6-day (top) and 24-day (bottom) lead time probabilistic forecasts between OPTIMISTS ("OP" – Δt = 7 days, 2 objectives, n = ~~100~~50, no optimization, and F-class kernels~~). Each scenario consists of two weeks where observations are assimilated (100 particles for each method),~~) and a ~~subsequent open-loop forecast of two weeks. For~~ traditional PF (n = 50) for the Blue River. The dark blue and orange lines indicate the mean of OPTIMISTS' and the PF's ensembles respectively, while the light blue and light orange bands illustrate the spread of the forecast by highlighting the areas where the probability density of the estimate is at least 50% of the ~~PF and OPTIMISTS~~density at the ~~hydrographs represent ensemble averages. The percentage values at~~ mode (the ~~bottom of each panel are~~ maximum) at that time step. The green bands indicate areas where the ~~MAREs in the forecast period with respect to the observations for each of the methods (in the order they appear in the legend).~~light blue and light orange bands intersect.


Scenario
Time step (Δt)
Objectives
Particles (n)
Optimization
w_root
1
2
1 hour
2 weeks
1
2
100
200
false
true
0.60
0.95
ΔNSE_zz
1.0
0.0
-1.0
-2.0
-3.0
-4.0
p = 0.000
p = 0.000
p = 0.001
p = 0.237
p = 0.006
p = 0.275
ΔNSE_zt
1.0
0.0
-1.0
-2.0
-3.0
p = 0.298
p = 0.000
p = 0.000
p = 0.146
p = 0.000
p = 0.006
ΔMARE
200%
100%
0%
-100%
p = 0.000
p = 0.000
p = 0.000
p = 0.034
p = 0.000
p = 0.004
1
2
1 hour
2 weeks
1
2
100
200
false
true
0.60
0.95
Scenario
Time step (Δt)
Objectives
Particles (n)
Optimization
w_root

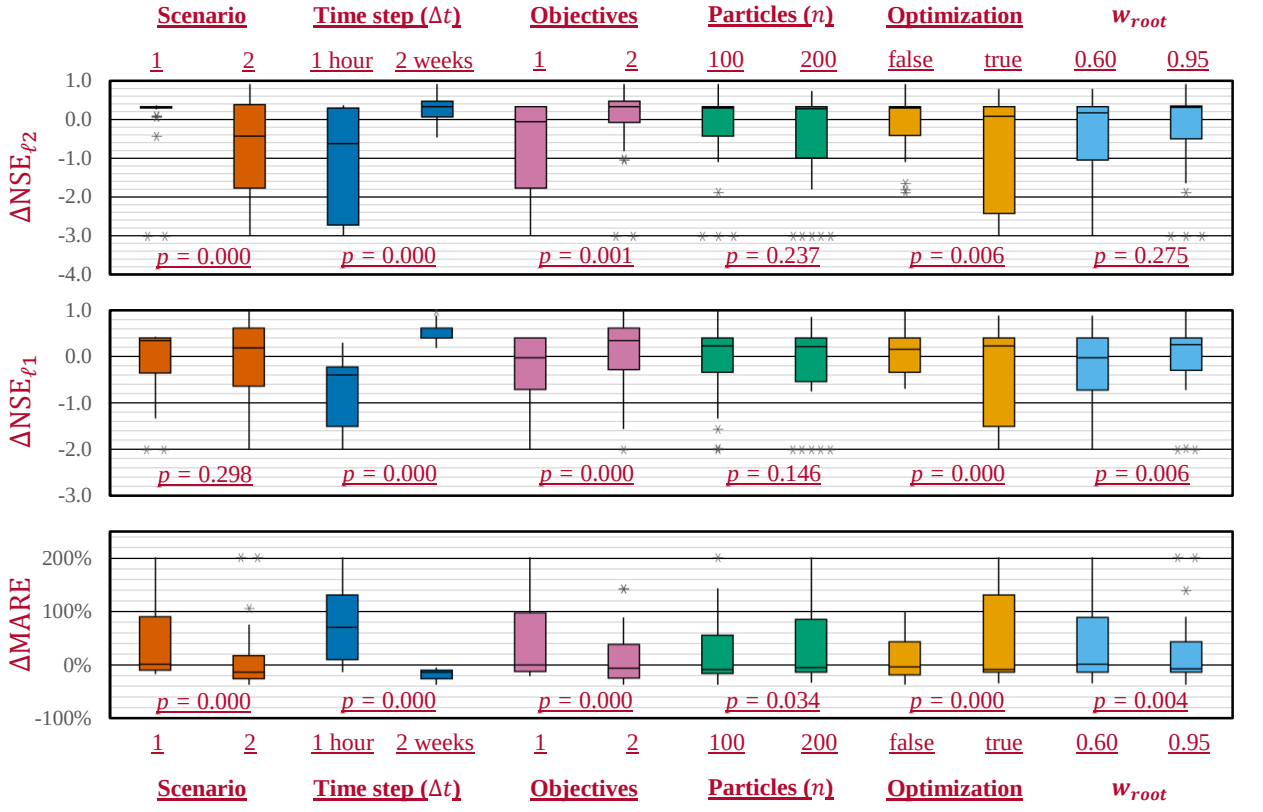
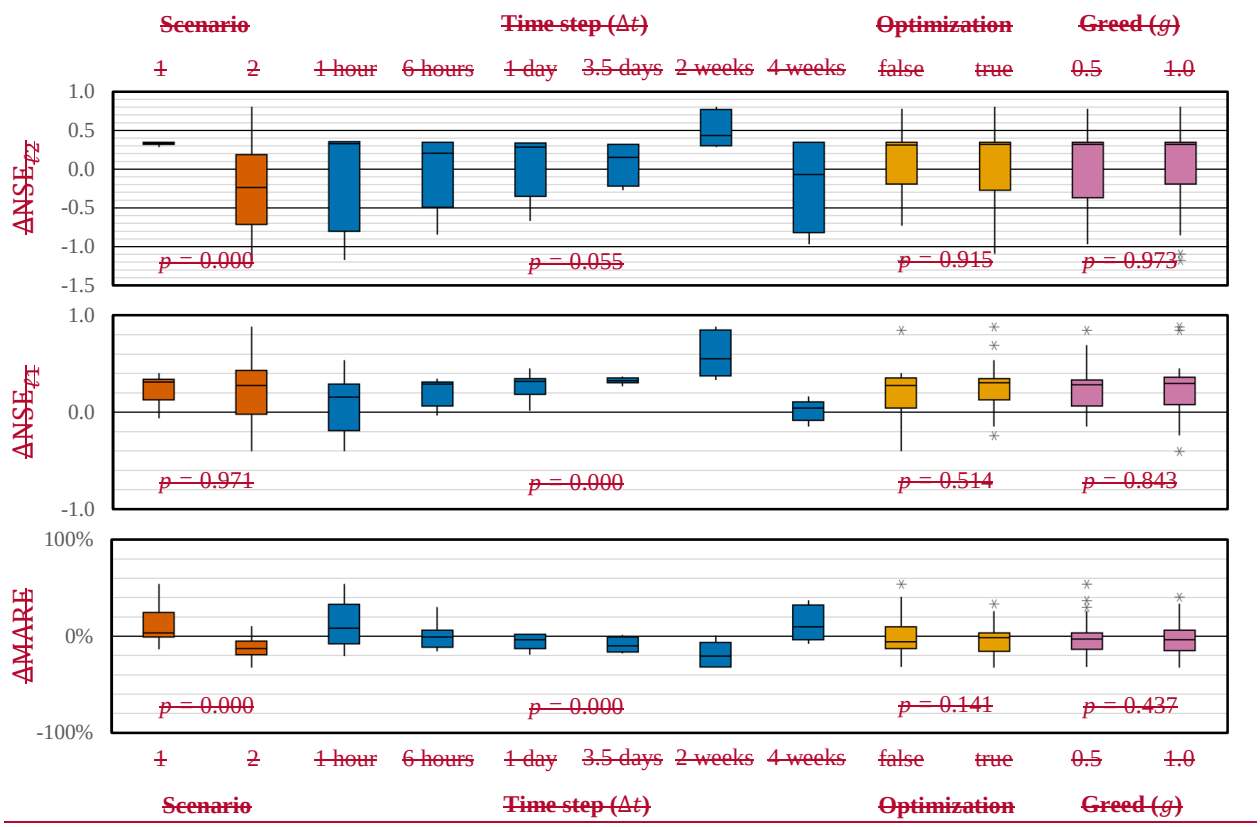


Figure 5~~6.~~ Boxplots of the changes in forecasting performance ($NSE_{\ell 2}$, $NSE_{\ell 1}$, and MARE) achieved while using OPTIMISTS on Experiment 2 (Indiantown Run). Changes are relative to an open-loop control run where no assimilation was performed. Each column corresponds to the distribution of the error metric changes on the specified scenario or assignment to the indicated parameter. Outliers are noted as stars and values were constrained to $NSE_{\ell 2} \geq -3$, $NSE_{\ell 1} \geq -3$, and $MARE \leq 200\%$. Positive values indicate improvements for the $NSE_{\ell 2}$ and the $NSE_{\ell 1}$. The meanings for the MARE and for other symbols are the same as those defined in Fig. 3.


Scenario
Time step (Δt)
Optimization
Greed (g)
1
2
1 hour
6 hours
1 day
3.5 days
2 weeks
4 weeks
false
true
0.5
1.0
ΔNSE_zz
1.0
0.5
0.0
-0.5
-1.0
-1.5
p = 0.000
p = 0.055
p = 0.915
p = 0.973
ΔNSE_tt
1.0
0.0
-1.0
p = 0.971
p = 0.000
p = 0.514
p = 0.843
ΔMARE
100%
0%
-100%
p = 0.000
p = 0.000
p = 0.141
p = 0.437
1
2
1 hour
6 hours
1 day
3.5 days
2 weeks
4 weeks
false
true
0.5
1.0
Scenario
Time step (Δt)
Optimization
Greed (g)

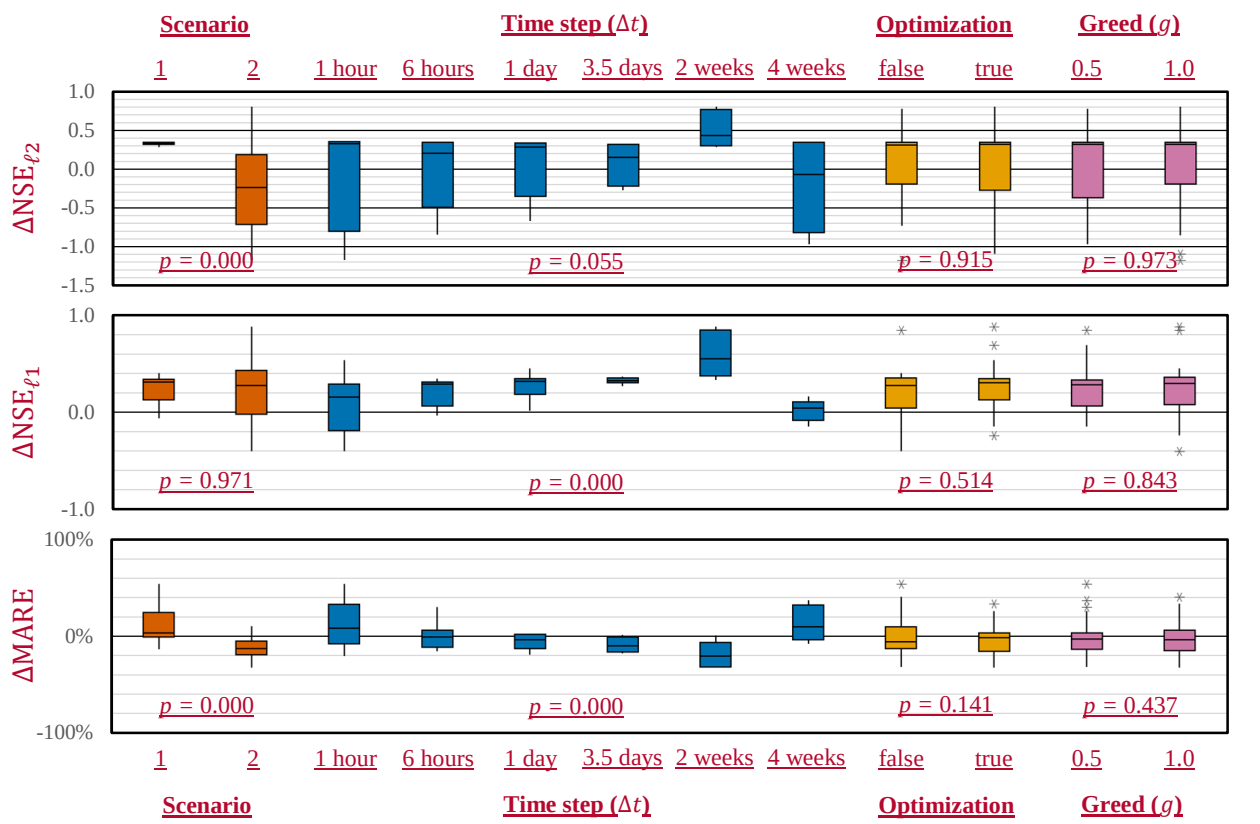
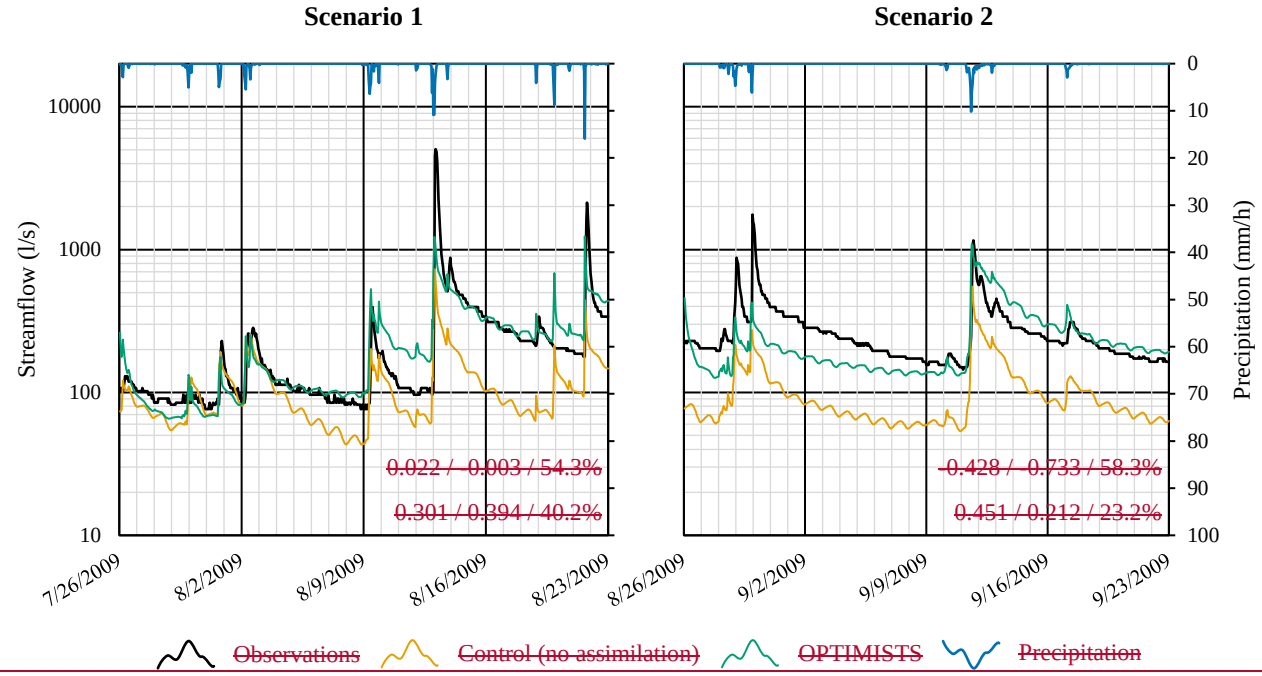


Figure 6~~7~~. Boxplots of the changes in forecasting performance ($NSE_{\ell 2}$, $NSE_{\ell 1}$, and MARE) achieved while using OPTIMISTS on Experiment 3 (Indiantown Run). Changes are relative to an open-loop control run where no assimilation was performed. Each column corresponds to the distribution of the error metric changes on the specified scenario or assignment to the indicated parameter. Positive values indicate improvements for the $NSE_{\ell 2}$ and the $NSE_{\ell 1}$. See the caption of Fig. 3 for more information.



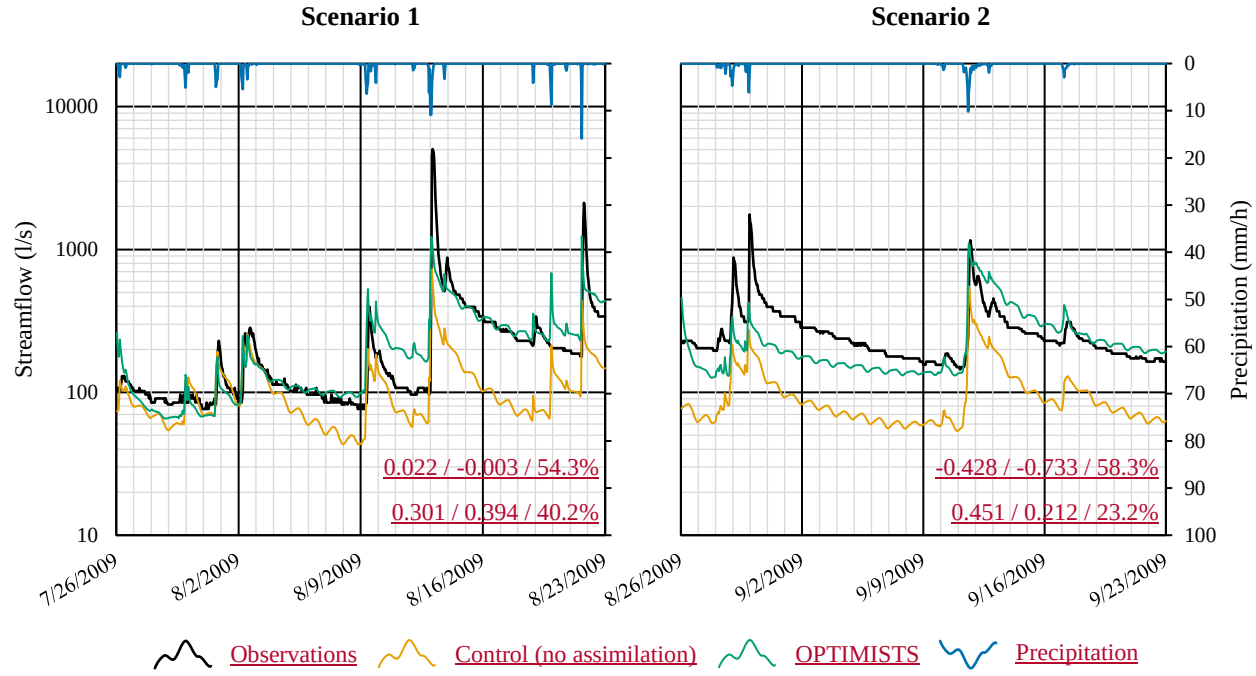


Figure 7~~8~~. Comparison of hydrographs for each of the two Indiantown Run scenarios. The results from the control model (with no data assimilation) and from the ensemble average of an OPTIMISTS configuration with Δt = 2 weeks, 2 objectives, n = 100, no optimization, and D-class kernels are compared with the observations from the stream gauge. The first two weeks correspond to the

assimilation period while the latter correspond to the forecast period. The error metrics corresponding to the forecast period are indicated for the default model (above) and the OPTIMISTS ensemble (below) as follows: $\mathbf{NSE}_{\ell 2}$ / $\mathbf{NSE}_{\ell 1}$ / MARE.

Table 1~~1~~. Comparison between the main features of standard Bayesian data assimilation algorithms (KF: Kalman Filter, EnKF: Ensemble KF, PF: Particle Filter), variational data assimilation (one- to four-dimensional), and OPTIMISTS.

	Bayesian	Variational	OPTIMISTS
Resulting state-variable estimate	Probabilistic: Gaussian (KF, EnKF), Non-Gaussian (PF)	Deterministic (unless adjoint model is used)	Probabilistic (using kernel density estimation)
Solution quality criteria	High likelihood given observations	Minimum cost value (error, departure from history)	Minimum error, maximum consistency with history
Analysis time step	Sequential	Sequential (1D-3D) or entire assimilation window (4D)	Flexible
Search method	Iterative Bayesian belief propagation	Convex optimization	Coupled belief propagation/multi-objective optimization
Model dynamics	Linear (KF), non-linear (EnKF, PF)	Linearized to obtain convex solution space	Non-linear (non-convex solution space)

Table 2~~2~~. List of global parameters in OPTIMISTS

Symbol	Description	Range
Δt	Assimilation time step (particle evaluation time frame)	$\mathbb{R}^+$
n	Total number of root states $\boldsymbol{s}_i$ in the probability distributions	$\mathbb{N} \geq 2$
w_{root}	Total weight of root samples drawn from $\boldsymbol{S}^t$	$\mathbb{R} \in [0, 1]$
p_{samp}	Percentage of n corresponding to drawn and random samples	$\mathbb{R} \in [0, 1]$
$k_{\text{F-class}}$	Whether or not to use F-class kernels. If not: D-class kernels.	true or false
n_{evo}	Samples to be generated by the optimizers per iteration	$\mathbb{N} \geq 2$
g	Level of greed for the assignment of particle weights w_i	$\mathbb{R} \in [-1, 1]$

Table 3~~3.~~ Characteristics of the two test watersheds: Blue River and Indiantown Run. US hydrologic units are defined in (Seaber et al., 1987)~~Seaber et al. (1987).~~ Elevation information was obtained from the Shuttle Radar Topography Mission (Rodríguez et al., 2006)~~(Rodríguez et al., 2006);~~ land cover and impervious percentage from the National Land Cover Database (Homer et al., 2012)~~(Homer et al., 2012);~~ soil type from CONUS-SOIL (Miller and White, 1998)~~(Miller and White, 1998);~~ and precipitation, evapotranspiration, and temperature from NLDAS-2 (Cosgrove et al., 2003)~~(Cosgrove et al., 2003~~. The streamflow and temperature include their range of variation for 90% of the time (5% tails at the high and low end are excluded).

Model characteristic	Blue River	Indiantown Run
USGS station; US hydrologic unit	07332500; 11140102	01572950; 02050305
Area (km^2); impervious	3,031; 8.05%	14.78; 0.83%
Elevation range; average slope	158 m – 403 m; 3.5%	153 m – 412 m; 14.5%
Land cover	43% grassland, 28% forest, 21% pasture/hay	74.6% deciduous forest
Soil type	Clay loam (26.4%), clay (24.8%), sandy loam (20.26%)	Silt loam (51%), sandy loam (49%)
Avg. streamflow (90% range)	9.06 m^3/s (0.59 m^3/s – 44.71 m^3/s)	0.3 m^3/s (0.035 m^3/s – 0.793 m^3/s)
Avg. precipitation; avg. ET	1,086 mm/year; 748 mm/year	1,176 mm/year; 528 mm/year
Avg. temperature (90% range)	17.26°C (2.5°C – 31°C)	10.9°C (-3.5°C – 24°C)
Model cells; stream segments; d	20; 14; 812	1,472; 21; 33,455
Resolution	0.125°; daily	100 m; hourly
Calibration	167 parameters; 85 months; objectives: NSE_{ℓ_2}, NSE_{ℓ_1}, MARE	18 parameters; 20 months; objectives: NSE_{ℓ_2}, MARE, absolute bias

Table 4~~4.~~ Setup of the three factorial experiments, including the watershed, the total number of configurations (conf.), the values assigned to OPTIMISTS' parameters, and which objectives (objs.) were used (one objective: minimize MAE given the streamflow observations; two objectives: minimize MAE and maximize likelihood given source/background state distribution S^t). n_{evo} was set to 25 in all cases. The total number of configurations results from combining all the possible parameter assignments listed for each experiment. Note that for Experiment 3 there are configurations that require a four-week assimilation period (all others have a length of two weeks).

No.	Watershed	Conf.	Δt	n	w_{root}	p_{samp}	$k_{F-class}$	g	objs.
1	Blue River	48	1d, 5d, 2w	100, 500	0.95	0.25, 1	false, true	0.75	1, 2
2	Indiantown Run	32	1h, 2w	100, 200	0.6, 0.95	0.25, 1	false	0.75	1, 2
3	Indiantown Run	24	1h, 6h, 1d, 3.5d, 2w, 4w	100	0.95	0.4, 1	false	0.5, 1	2

Table 5~~5. Forecast~~. Continuous daily streamflow forecast performance metrics for the Blue River application ~~after assimilating two weeks' worth of daily streamflow observations~~ using ~~a PF, Evo4DVar (c_1 = 0.01 and c_2 = 1.0), and~~ OPTIMISTS (Δt = ~~5~~7 days, 2 objectives, n = ~~100~~50, no optimization, and F-class kernels) ~~with 100 particles.~~and a traditional PF (n = 50). The ~~error metrics~~continuous forecast extends from November, 1996 to November, 1997. The $NSE_{\ell 2}$, $NSE_{\ell 1}$, and MARE (deterministic) are computed ~~throughout the two weeks of the forecast period by comparing the streamflow of the deterministic solution (Evo4DVar) or~~using the mean streamflow of the forecast ensembles and contrasting it with the daily observations, while the CRPS (probabilistic) is computed taking into account all the particles in the ensemble ~~(PF and OPTIMISTS) with the observations~~.

Algorithm	Lead time (days)	$NSE_{\ell 2}$	$NSE_{\ell 1}$	MARE	CRPS (m^3/s)
OPTIMISTS	1 day	0.827	0.636	37.20%	3.745
	3 days	0.750	0.555	42.86%	4.899
	6 days	0.812	0.604	38.88%	4.067
	12 days	0.788	0.625	38.69%	3.959
	24 days	0.801	0.609	39.01%	3.861
Particle filter	1 day	0.796	0.638	35.59%	4.175
	3 days	0.776	0.612	37.52%	4.463
	6 days	0.733	0.588	39.44%	4.766
	12 days	0.705	0.578	40.48%	4.857
	24 days	0.743	0.588	41.16%	4.772

Hybridizing Bayesian and variational data assimilation for robust high-resolution hydrologic forecasting

Felipe Hernández, Xu Liang

Civil and Environmental Engineering Department, University of Pittsburgh, Pittsburgh, 15213, United States of America

Correspondence to: Xu Liang (xuliang@pitt.edu)

We would first like to express our appreciation to both Referee #1 and Referee #2 for their careful and thorough review of our manuscript. Their comments will certainly help us improve the quality and clarity of our manuscript. Below we offer our response to Referee #1.

Anonymous Referee #1 on 13 October 2017:

"I really enjoyed reading the paper, which deals with the important issue of improving model updating techniques for better flood predictions. This manuscript proposes a new data assimilation procedure which combines Bayesian and variational approaches. I believe this is an important contribute to DA research field and of notable interest and modernity, especially for the HESS readership. However, I still have some comments regarding method, structure, and readability of the paper. Below you can find some major comments:"

We are very glad that you enjoyed the manuscript and found it of value. We appreciate your generous compliments and hope that our work will be a worthy contribution to the community. We thank you for your careful analysis and suggestions. Below we respond to your comments.

"Results of this research are not well described in the abstract and it is somehow difficult to grasp the main advantage of this approach."

We will modify the last sentences of the abstract to better convey the findings of our experiments: that OPTIMISTS allowed to produce more robust forecasts when compared to a more traditional method, and that its application on a high-dimensional model was successful in that it maintained the levels of performance observed in the case of lower dimensionality without sacrificing computational efficiency.

"From the introduction, it is not really clear the difference between OPTIMIST and other hybrid 4D approaches. Novelty has to be better explained in order to further appreciate the added value of such method."

We will extend the introduction to better convey the differences and our proposed innovations in the revised manuscript. For your information, the main differences were summarized in the conclusions (p16-l28) in the original submission: that OPTIMISTS is inspired in part by the particle filter (instead of being inspired by the EnKF), that it allows for multi-criteria evaluations of candidate particles, and that it utilizes global (evolutionary) optimization methods (instead of using convex ones). Further differences can be extracted from table 1: e.g., that OPTIMISTS allows for non-Gaussian state estimates through kernel density estimation.

"Nowadays, there are many DA methods with varying complexities and accuracies. However, few of these methods are used in early warning system to improve flood predictions. Did the authors investigate the way to easily implement OPTIMISTS by water authorities for flood forecasting in any existing early warning system? Is there any advantage in terms of computational time if compared to PF and 4D-Var?"

It is our vision that, once OPTIMISTS' advantages prove greater than its disadvantages (e.g., its relatively higher complexity) in a significant set of test cases, the method will be considered for integration in operational prediction systems for multiple applications. Applicability in high-dimensional cases has been one of our guiding design principles, as we believe this is key for adequate performance in these large-scale/complex systems. While the method currently can be executed in parallel environments and it offers configurations that work well with very large state variable vectors, we are already working in developing enhancements to take the next step in scalability and efficiency, and we will be happy to continue working alongside government agencies or private partners to carry out our vision.

Regarding the comparisons with the PF and 4D-Var, we believe our experiments indicate that OPTIMISTS can provide gains in computational time given that the main strategy for increased performance in most DA methods is through the use of additional model runs: e.g., enlarging the ensemble size in EnKF and PF or the number of candidate solutions explored in the optimization problem in Var. Therefore, showing a more robust performance indicates not only that OPTIMISTS can produce better forecasts with the same resources, but that the same level of performance can be obtained with fewer resources (in this case particles/ simulations).

"Page 1, lines 14-16: Is this sentence related to the watershed's location or to the use of different models for different case studies? Authors should clarify this point."

Both: two different locations are used and, because the watersheds are of significantly different scale, we used a different modelling engine for each. We made this decision to diversify the test conditions of our experiments. We will change the wording in this sentence to convey the correct meaning more clearly.

"Page 3, lines 8-9: The authors mentioned that "a hybrid data assimilation algorithm that incorporates the most valuable features from both Bayesian and variational methods ". Which ones are these valuable features?
Description of Table 1 should be better included within the paper. Right now it looks quite disconnected from the other part of the introduction."

Our intention was to let Table 1 convey the contrast in features between OPTIMISTS and traditional methods. We will add discussions in the main text so that there is a stronger connection with the table and that these claimed advantages are more easily understood.

"I found very difficult to follow the flow of thoughts of the authors in describing the DA method. I think it will be beneficial for the readability of the paper to include in section 2 a figure representing the structure of OPTIMISTS. In addition, authors tend to use complex terms for non-DA expert. I suggest revising the description of the paper in order to make it "accessible" to everyone and increase its impact on the scientific community."

Thank you for your good suggestions. We will use more common terminology in the revised manuscript to accompany technical/domain-specific terms to improve the accessibility of the section. We will also add a figure that will help in understanding how the algorithm works.

"At this point, results are valid only for the 5 considered flood events and 2 basins obtained. As expected, results largely depend on the features of the flood events and quality of rainfall data. I am afraid that the samll number of events makes results rather random. I suggest to increase the number of flood events to make more general conclusions for this study."

Thanks for the good suggestions. We will change the comparative test design so that we can analyze multiple months' worth of forecasts using OPTIMISTS and the particle filter. We already developed some scripts that allow performing assimilation and forecast continuously for this purpose. This will allow for a more thorough and realistic comparison between the two methods. However, we decided to drop the comparison with Evo4DVar as this is not a standard method found in the literature, and its deterministic nature makes a direct comparison complicated.

"A crucial component in each DA application is the proper definition of model and observational error. While model error is accurately described, I could not see a clear definition of the observations error (standard deviation in Eq.10). The authors have to include more information and references about it."

We will incorporate the effect of uncertainties in the observations within the description of the algorithm to better convey the multiple ways in which it can be addressed through OPTIMISTS and the contrasts with traditional methods.

"Are you using actual meteorological forecasts or are you using the observations as perfect forecasts? Please specify"

In this study, "perfect" meteorological forecasts are used to force the model in all cases (see in page 12, lines 8-13). We also argue that this advantage is applied uniformly to both OPTIMISTS and the other algorithms so that it does not represent an unfair advantage to any. In the revised manuscript, we will make this clearer.

"I suggest the authors to split results and discussions in two different sections, this would make reading the text so much easier."

We understand that having separate results and discussion sections allows having the "cold facts" separated from the "subjective" interpretations and opinions of the authors. However, as many authors do, we prefer having these two sections combined because otherwise the discussion section will often have to reference results and figures from the previous section and have the reader jumping back and forth between sections. This both helps the readability of the article and its compactness. Having both styles be accepted in many communities, we hope this reviewer would agree with us on this point to maintain the two sections combined. We will, however, scan the entire section in search for instances where the distinction between objective results and subjective opinion is not clearly established and revised them accordingly based on this reviewer's good comments.

Hybridizing Bayesian and variational data assimilation for robust high-resolution hydrologic forecasting

Felipe Hernández, Xu Liang

Civil and Environmental Engineering Department, University of Pittsburgh, Pittsburgh, 15213, United States of America

Correspondence to: Xu Liang (xuliang@pitt.edu)

We would first like to express our appreciation to both Referee #1 and Referee #2 for their careful and thorough review of our manuscript. Their comments will certainly help us improve the quality and clarity of our manuscript. Below we offer our response to Referee #2.

Anonymous Referee #2 on 21 December 2017:

"Before beginning review of this manuscript, although not mentioned in the text and reference, this should be considered a re-submission of the previous HESSD manuscript entitled "Hybridizing sequential and variational data assimilation for robust high-resolution hydrologic forecasting (https://doi.org/10.5194/hess-2016-454)" by the same authors, which was rejected in 2016. I suggest the editorial board compare the final revision of the previous HESSD manuscript with the current one if the track record was not screened yet. I cannot examine whether the authors submitted the final revision in the previous submission in 2016 or not. However, if so, improvement and uniqueness of the current manuscript over the rejected final manuscript should be carefully evaluated. In addition, since HESSD is independent publication, the previous manuscript in 2016 should be cited and discussed in this manuscript."

This point had been discussed with the editor, to whom we expressed that we will adhere to the guidelines required by the journal and the editorial board. This manuscript is a revised version of the cited one in 2016, which takes into account the comments made by the referees back then and by the previous editor. A detailed account of the changes made was submitted to the journal. The 2016 manuscript was rejected by the editor because he considered that the required modifications deserved a more careful timeframe than the one available for the special issue it was submitted to. No final version of the 2016 manuscript was submitted.

"This manuscript proposed a hybrid DA method, OPTIMISTS, combining sequential and variational methods, and compared performance of developed methodology over PF and VAR using distributed hydrologic models. The topic is of interest to a wide range of hydrologic modelling community. The strategy of the proposed methodology to leverage different DA approaches, sequential and variational DA, is one of the important trends in recent studies. However, there are major gaps in experimental setup and evaluation, and incomplete reasoning in new methodology which require significant changes before publication. I hope the followings would be helpful to improve the quality of manuscript.

1) Evaluation period and methods In this manuscript, the total evaluation period is 10 weeks (5 cases with a 2-weeks period each): 3 scenarios for 2-weeks forecasts in the Blue River and 2 scenarios for 2-weeks forecasts in the Indiantown Run, not including assimilation period.

The evaluation period for hydrologic modelling and data assimilation is usually longer than at least 6-8 months and up to multiple decades. The total 10-weeks forecasts (2-weeks piecewise each) and associated metrics cannot be accepted as a rigorous evaluation.

Given that the selected events in the Blue River in the 2016 manuscript are different from those in the current one, there seems to a potential to further increase evaluation period. In Table 3, the authors also mentioned calibration periods are 85 and 20 months, respectively.

Considering the availability of observation data, what is the maximum evaluation period for two catchments? Why don't you use the whole or most calibration period for DA evaluation? Was there any reason to use the limited period for evaluation?

For the larger domain, the Blue River catchment, is there just one streamflow observation gage over 3,000 square kilometer area? Why don't you assimilate observations in multiple locations to reduce equifinality and overfitting?

In this study, evaluation metrics were estimated for the whole 2-weeks forecast period. However, it is more common to evaluate metrics for varying forecast lead times because the impact of updating varies and disappears over time.

I highly suggest the evaluation period and method should be reconsidered to qualify a kind of general standard shown in many forecast and DA-related papers: simulating more than several months for each catchment and evaluating metrics for varying forecast lead times."

Thank you for the good suggestions. We will evaluate our method based on the suggestions here in the revised manuscript. We already extended our scripts to allow running an extended-time data assimilation experiment where assimilation is performed with OPTIMISTS continuously to allow producing multiple time series of forecasts with a fixed lead time for the Blue River. We should be able to produce these forecasts for multiple months in order to analyze the performance of the algorithm. We will similarly develop a script to run the particle filter in the same fashion and to be able to compare its forecasts with those of OPTIMISTS.

On the other hand, this new experimental setup will preclude the comparison with the 4D evolutionary variational algorithm for the reason that we were using the same prior "particle" ensemble to seed its population that the one used for the other methods. However, given that 4DVar is inherently a deterministic approach, such ensemble will not be able to be updated to be the seed of continuous assimilation periods. In practice, variational methods compensate the lack of an ensemble by performing a guided search of the initial state solution space until convergence, but in this case we consider that the model simulation quota that we are allowing each of the methods will not suffice to reach an optimal solution. Moreover, evolutionary variational methods are rare in the literature (more so in operational settings) and therefore we now consider that the comparison would not be of enough significance. While ideally we would like to compare OPTIMISTS with proven 4DVar

methods, these require the linearization of the model's dynamics—which is rare in hydrology, would require an enormous amount of additional work, and it is outside the scope of this paper.

Also, while we will implement this new evaluation scheme for the comparison of OPTIMISTS with the particle filter, we will still maintain the scenario-based design for the analysis of the parameters of the assimilator. Even though not "rigorous," we are confident that the variability in the scenarios considered is enough to differentiate relatively adequate configurations of the algorithm. As discussed in the manuscript, we will conduct future tests for cases where there was not enough statistical evidence to conclude one configuration was better than the other. Extended-time evaluations would, in this case, require an extensive computational budget given the large number of parameter combinations selected.

We agree that assimilating data from additional streamflow gages in the Blue River would allow for improved forecasts but, for the purposes of this manuscript, the assimilation of a single measurement provides a balanced challenge that enables the analysis of the strengths and weaknesses of OPTIMISTS in contrast with other methods, and the determination of an adequate set of parameters. That said, we have actually already worked on testing the algorithm using distributed high-resolution observations in a watershed and look forward to include such analyses in a future publication at its due time.

"2) Probabilistic evaluation Although the proposed method is a stochastic approach, probabilistic metrics were not measured and analyzed. At least, basic metrics such as reliability, CRPS, predictive QQ plot and Brier score should be compared over the conventional method such as PF. Without such evaluation, improvements and features of the hybrid ensemble method cannot be understood in terms of stochastic perspectives.
In addition, Figs 5 and 8 (streamflow hydrographs) should include traces or spreads of ensemble for visual inspection."

Indeed, probabilistic evaluation is very important to determine if forecasts are over-confident or under-confident. We will include an evaluation metric that allows comparing the confidence of forecasts between OPTIMISTS and the particle filter. We will also modify the plots to show the temporal evolution of the distributions in the revised manuscript.

"3) uncertainty specification on hydrologic DA In order to apply DA for hydrologic modelling, uncertainties for states and observations should be carefully taken care of. Sometimes, not surprisingly, noise configuration or specification may significantly affect DA performance. However, there is no description on how uncertainties of different state variables and observations such as interception, snow, soil moisture and streamflows were formulated and implemented for hydrologic ensemble modelling, which should impact DA process to generate ensemble, optimize state variables and estimate likelihood or weight. A detailed description is required for reproducibility of this study.

Regarding this issue, for example, how different particles of distributed hydrologic models are generated in "the sampling step" of this DA algorithm? More specifically, how high-dimensional model states are being perturbed to avoid sample impoverishment in this step?"

As explained in subsubsection 2.1.4 in the original manuscript, any numerical objective can be used to judge candidate particles in OPTIMISTS. The likelihood of simulated outputs given the distributions of the corresponding observations is cited as an example. In such a case, the user would require specifying how the likelihood is computed based on how the error of the observations is being modeled. However, the error metric used in our tests (the mean absolute error, page 12, line 4) is a deterministic one. While this constitutes a departure from the Bayesian theoretical framework, the estimation performed in OPTIMISTS retains its probabilistic character due to the way in which samples are generated and, especially, due to the proposed probabilistic interpretation of the resulting Pareto front (page 7).

Uncertainties in the state variables are all captured by the use of kernel density probability distributions, which is the whole focus of subsection 2.2. The details of the implementation are not introduced in subsection 2.1 but saved for subsection 2.2 because OPTIMISTS offers a modular design in which any type of non-parametric (ensemble-based) probabilistic representation could be used. How new samples are generated from the prior distributions (the core mechanism to "perturb" the ensemble) and how the likelihood of samples given these priors is computed is all explained in this part of the manuscript. While this arrangement was announced in page 4, lines 12-15, we will add reminders on subsubsections 2.1.2 and 2.1.4 in the revised manuscript to make the presentation clearer.

"4) Under-simulation or filter degeneracy in assimilation step In the analysis or assimilation step which corresponds the first 2-weeks in Figs 5 and 8, under-simulation or filter degeneracy (scenario 3 in Fig. 5 and scenario 2 in Fig. 8) is found. Usually, whatever filter is used, traces of simulated states (here streamflow) overlap observations in the assimilation step since uncertainty of observation is set smaller than that of state variables. It is common that NSE values of the assimilation step or the first forecast step are higher than 0.9 – 0.95. However, a large gap between simulation and observation exists even in the assimilation step, which should be clearly diagnosed and discussed."

There are several reasons that might explain the relatively low level of agreement seen between the observations and the adjusted ensemble during the assimilation period. In the first place, it must be noted that the models do indeed have considerable errors, probably mainly in their structures, that prevent them from faithfully replicating the observations at every time step precisely. This is especially apparent in Scenario 3 for the Blue River and both scenarios for Indiantown Run, in which there appears to be conflict between fitting the peaks and fitting the drier inter-peak periods. While both models underwent parameter calibration processes, as documented in subsection 3.1 and in Table 3, no attempt was made to optimize the models' structures (e.g., equations, missing phenomena, resolution, connectivity, etc.). The calibration process, similar to

the assimilation, was based on multiple objectives and not only on the maximization of the NSE: we also used the relative error which is more sensitive to errors during dry periods than those during peaks. There is also a telescopic effect of the NSE, according to which, computing it over long periods of time yields higher values than when computed over short ones: for example, if the Indiantown Run model had an overall NSE of 0.81 during the entire calibration period, zooming in on a specific month would result, in average, on a reduced rating. This effect is compounded with the relatively short period of time used for assimilating data and performing forecasts. Finally, a comparably "poor" performance during the assimilation period was also observed for the particle filter and the variational algorithm. With these, we do not find the results to be especially concerning in this regard and, on the other hand, consider that all the provided contrasts are valid given that these conditions were uniform in all cases. In fact, these "defects" reflect the current state-of-the-art challenges in the operational forecasts and it is one of the objectives that we all try to improve from different aspects/angles. We will, however, include a few words in the manuscript regarding these low fitting scores.

"5) Comparison of posteriors of state variables What potential readers want to see in the result section may be not only comparison of NSE at the outlet location. The authors need to address why and how their DA method can improve over the conventional ones in hydrologic forecasting from perspectives of distributed modelling. A comparison of posterior distributions of state variables updated by the new and conventional methods may be useful to show how and why the new DA works for high dimensional applications.
Especially, given that the authors urged OPTIMISTS employed essential features from but outperformed particle filters, a comparison of posteriors between two methods is also required to demonstrate whether non-Gaussian and multi-modal distributions are preserved or not."

We will include probabilistic time series of average soil moisture for forecasts produced both by OPTIMISTS and the particle filter and perform the corresponding analysis. However, we plan to perform detailed analyses of OPTIMISTS' capability of estimating soil moisture, and not only aggregated outputs like streamflow, in a later investigation (when such observations are available). For this study, due to the data limitations at the test watersheds and the length of the manuscript, the distributed comparisons won't be carried.

"6) Evaluation and optimization steps for hydrologic modelling It is not clear how the cost function is formulated for distributed hydrologic models. The authors need to show explicitly how multiple spatially-distributed state variables and associated uncertainties are taken into account to formulate the cost function in evaluation and optimization steps."

As explained in page 12, lines 4-7 in the original manuscript, one or two objective functions were used for our experiments: the mean absolute error given the streamflow observations and the likelihood of the particle given the prior state distribution. These objectives can be seen as analogous to the "cost function" used in variational data assimilation, and their equivalence is

established in subsubsection 2.1.4. The likelihood is computed using either equation 8 or 9 depending on which type of kernels are used for the state variable distribution. These distributions encode the spatial variability and relationships between state variables in all cells of the model, so the likelihood is thus a measure of how well a candidate particle conforms to the values and (spatial) patterns in the prior distribution. Again due to the limitations of spatial data availability, such evaluations are not directly carried out in this study in a spatially distributed fashion, but indirectly evaluated through the integrated quantity of streamflow.

"7) Tuning hyper-parameters There are numerous hyper-parameters such as time step, objectives, no. particles, optimization, Wroot and Kf-class, Psamp and g, related to this DA method which may increase uncertainty and subjectivity of forecasting. However, analysis methods and results on hyper-parameters shown in Figs. 3, 6 and7 are still confusing and do not provide well-organized understandings. A summary or guideline is required for proper range or values of hyper-parameters."

We acknowledge that using factorial experiments is not a common practice when evaluating the hyper-parameters of these kind of methods. We will revise our presentation of the results to attempt to convey their significance in a more understandable and clearer way. This will include the suggested summary of guidelines for potential OPTIMISTS users to parameterize the algorithm to better fit the needs of their specific application. We will possibly remove Figure 4 which introduces a format different from the other boxplots but that does not provide many significant insights.

"8) Terms and units In Table 3, use of two different units for one variable (m3/s and l/s) is not recommended."

We will change the table to use unified units to *m3/s*.

"Throughout the manuscript, the term 'time step' is used to represent 'assimilation time step' or 'assimilation window'. Since the time step usually stands for a temporal increment for numerical schemes, 'assimilation window' or 'analysis window' may be more appropriate to avoid possible confusion."

It is true that the term "assimilation time step" might be confused with the model time step, but it is still necessary to use it because the "assimilation window" or the "analysis window" would refer to the entire period of time in which data is assimilated prior to performing a forecast. As explained in the original manuscript, OPTIMISTS allows dividing this time window into "time steps" of arbitrary length at which the main loop of the algorithm is executed. These time steps can be as short as the model time step ("sequential," like with particle filters) or as long as the assimilation window (like in 4DVar). The assimilation or analysis window would correspond to two weeks in most of our examples, and the assimilation time step varies from one day to two weeks in the Blue River case or from one hour to four weeks in the Indiantown Run case. We will review

the entire manuscript in search for places in which the distinction between model and assimilation time steps could be made more apparent to avoid confusion.