Reply to Anonymous Referee#2(R2)

Dear referee,

We thank you very much for taking the time to review our manuscript, for your valuable comments/suggestions and the points of discussion you raised. We will carefully revise our manuscript to improve its quality. Below please find our responses to your comments and the revised directions of the manuscript that we are going to resubmit.

Comment from R2 — General comments: Streamflow forecasting is important for water management and optimal allocation of water resources. This study aimed to improve the model performance of decomposition based forecasting methods. A two stage decomposition predication framework (TSDP) was proposed by the authors based on VMD and SVR, to avoid the influences of validation information on training. The effectiveness, efficiency and reliability of the TSDP framework and its VMD-SVR realization in terms of the boundary effect reduction, decomposition performance, prediction outcomes, time consumption, overfitting, and forecasting capability for long leading times were investigated. The final results on monthly runoff from three stations at the Wei River showed the superiority of the TSDP framework compared to benchmark models. It is found that the results are interesting for guiding proper use of decomposition-based forecasting methods in streamflow forecasting practice.

Reply: We thank you for the positive evaluation of our work and all the comments/suggestions on our manuscript. We will make improvements to the manuscript following your suggestions.

Comment from R2 — Specific comments: 1) This study only focused on decomposition-based methods and aimed to solve one disadvantage existing in applying decomposition methods. Although this might be interesting for readers who use decomposition based methods, a wider scope including more streamflow forecasting techniques like ARIMA, BP, LSTM etc. can be more interesting. Even if a new technique is proposed (not the case in this manuscript), a companion with different types of techniques is often needed to support the application of the proposed technique.

Reply: We will add non-decomposition ARIMA, BP, LSTM models in our manuscript.

Comment from R2 — 2) Five experiments were designed for the assessment of different performance aspects including the reduction of the boundary effects, decomposition performance, predictability, time consumption, overfitting, and forecasting capabilities for long leading times. This might be interesting for readers. However, it is difficult to understand these experiments, since the complicated five-experiment design and presentation styles stopped the successful understanding and digestion of the results. I suggest the authors rewrite this part and add tables (for a comparison of five experiments) to help readers better understand the six different experiments and their differences.

Reply: We will rewrite this part and add tables for comparisons of these experiments.

Comment from R2 — 3) Lines 66-67: when you mentioned the boundary effect for the first time in the manuscript, I expect an explanation of the ‘boundary effect’.

Reply:

We will explain the boundary effect the first time it is mentioned.

Comment from R2 — 4) VMD and SVR are well-known techniques. The authors can shorten the descriptions of these two techniques and focus on the new things the authors proposed.

Reply: We will shorten the descriptions of VMD and SVR in our revised manuscript.

Comment from R2 — 5) Line 81: change 'usage' to 'use'

Reply: We will correct this misuse.

Comment from R2 — 6) Line 259ïijŽ what is BOGP? Do you mean 'Bayesian optimization based on Gaussian processes'? How is BOGP used to optimize EEMD, SSA, DWT and SVR? Add some details.

Reply: Yes, the BOGP means Bayesian optimization based Gaussian processes. The BOGP was only used to optimize SVR-based models. The BOGP was not used to optimize the parameters of EEMD, SSA, DWT and VMD because (1) it is hard to define an objective function for the decomposition processes (the object function for SVR-based models is mean square error) and (2) we can manually control the decomposition performance by setting specific parameters (e.g., set the noise tolerance of VMD to zero to obtain sub-signals with low noise level). We will add more details about how to use the BOGP to optimize the SVR models in the revised manuscript. Below we give a brief explanation of using BOGP to optimize SVR models.

In this study, the BOGP was used to obtain the optimized hyperparameters of SVR, i.e., the weight penalty(C), the error tolerance (ϵ), and the width control coefficient (σ). As shown in Figure 1, the BOGP algorithm can be wrapped up as follows:

Step 1. Input a set of mixed and shuffled samples, the object function(f),i.e., the loss function(the mean square error was used in this study), the convergence error (e.g., E=1e-6) and the number of iterations (e.g., N_c=100), and the hyperparameters search space (e.g., C=[0.1,200],$\epsilon = [1e-6, 1]$,$\sigma = [1e-6, 1]$).

Step 2. Randomly sample a candidate (e.g., x_0=[C=25,ϵ=0.0001,σ=0.26]) based on the given search space and set the iteration index as i=1.

Step 3. Given the previous candidate, update the posterior expectation of f using the Gaussian process model.

Step 4. Track the new candidate (x_i) that maximize the expected improvement (EI) function, i.e., $x_i = argmaxEI(x)$.

Step 5. Compute $f(x_i)$ based on the mixed and shuffled samples (including predictors and predicted targets) and set the iteration index to $i = i + 1$.

Step 6. Repeat steps 3-5 until the convergence is achieved or the number of iterations is reached.

Step 7. Output the last candidate as the optimal hyperparameters of the SVR model.

Comment from R2 — 7) Add a table for a clear comparison of five experiments

Reply: We will add a table to compare the results of five experiments.

Comment from R2 — 8) 4 'Experimental Results and Analysis' should be 'Experimental results'

Reply: We will change 'Experimental Results and Analysis' to 'Experimental Results'.

Comment from R2 — 9) Line 354: Why 3,5,7,9?

Reply: We aim to evaluate the performance gap of the TSDP models for long lead times and the workload for evaluating both the odd- and even-numbered lead times is huge. Therefore, we think only evaluate the odd-numbered (or even-numbered) lead times is enough to tell the difference of models. We can also evaluate 1-, 2-, 3-, 4-month ahead forecasting models, however, the 1-, 3-, 5-, 7- and 9-month ahead forecasting models can tell the predication performance of much longer lead times. In fact, we will remove the 9-month ahead monthly runoff forecasting in our revised manuscript for the convenience of comparison and results presentation.

Comment from R2 — 10) Line 356: What does that mean by 'the 20-month lag'? Does that make sense for monthly forecast?

Reply: The '20-month lag' is the upper limit of lags for computing the Pearson correlation coefficient (PCC). The 20-month lags (i.e., 1-month lag, 2-month lag, ...) were used to compute the PCC, and the lags with higher absolute PCC were finally selected as input predictors. We set a 20-month lag as the upper limit is due to the maximum lags of Partial autocorrelation function(PACF) was also set to 20. Therefore, we can compare the prediction performance of models established using the input predictors determined by PCC and PACF.

Comment from R2 — 11) Figure 2: if possible, put a map of China

Reply: We will add a map of China to illustrate the location of the Wei River.

Comment from R2 — 12) I didn't really get how 'the mixing and shuffling step' works. If possible, please clarify.

Reply:

The mixing-and-shuffling step first mixes (concatenates) the calibration samples and development samples (i.e., half validation samples) as a single set of samples and then randomly shuffles the mixed sample rows (e.g., original row indexes are 1,2,3,4,5,6,7,8,9,..., and the shuffled row indexes might be 6,3,9,1,5,2,7,4,8,...).

In fact, two crucial steps help to reduce the influences of the boundary effect. One is generating validation samples from appended decompositions and the other is mixing and shuffling the training and development samples. In the current manuscript, we only have discussed the later one. In the revised manuscript, we will add discussion about the former one and clarify how these steps deal with the boundary effect.

The different error distribution of calibration and validation decompositions, which is caused by the boundary effect, leads to the models calibrated on the calibration samples generalize poorly to the validation samples (see *AC2: 'Potential ideas supporting the two-stage decomposition prediction (TSDP) framework'*). The aforementioned two steps are worked for reducing the influence of boundary effect because : **(1) The relationship between input predictors selected from appended decompositions and output target is maintained by the decomposition algorithms. In other words, the predictors can be reconstructed to original monthly runoff values by the decomposition algorithm. However, due to the decomposition errors come from different sets of appended decompositions, the predictors selected from the validation decompositions cannot be reconstructed to original monthly runoff values. This leads to the absolute Pearson correlation coefficients (PCCs) between predictors and predicted targets of validation samples generated from appended decompositions are larger than that of validation samples generated from validation decompositions (see Figure 2); (2) Mixing and shuffling the calibration and development samples, and training the models based on the shuffled samples enable the models to assess the validation error distribution during the calibration stage. The TSDP models were established based on the mixed and shuffled samples using cross-validation (CV) strategy (e.g.,10-fold CV means the mixed and shuffled samples are divided into 10 sub-samples, of which each one will be used to train and validate the TSDP models). Shuffling the mixed samples enables the validation samples to be randomly distributed throughout the mixed samples, which means the sub-samples extracted**

from the mixed and shuffled samples can be used to train and validate the validation error distribution. Therefore, the mixing-and-shuffling step can improve the generalization ability.

The results (see Figure 3) indicate that (1) generating validation samples from appended decompositions, (2) mixing and shuffling the calibration samples and development samples improve the prediction performance compared with the scheme without these two steps.

List of Figures

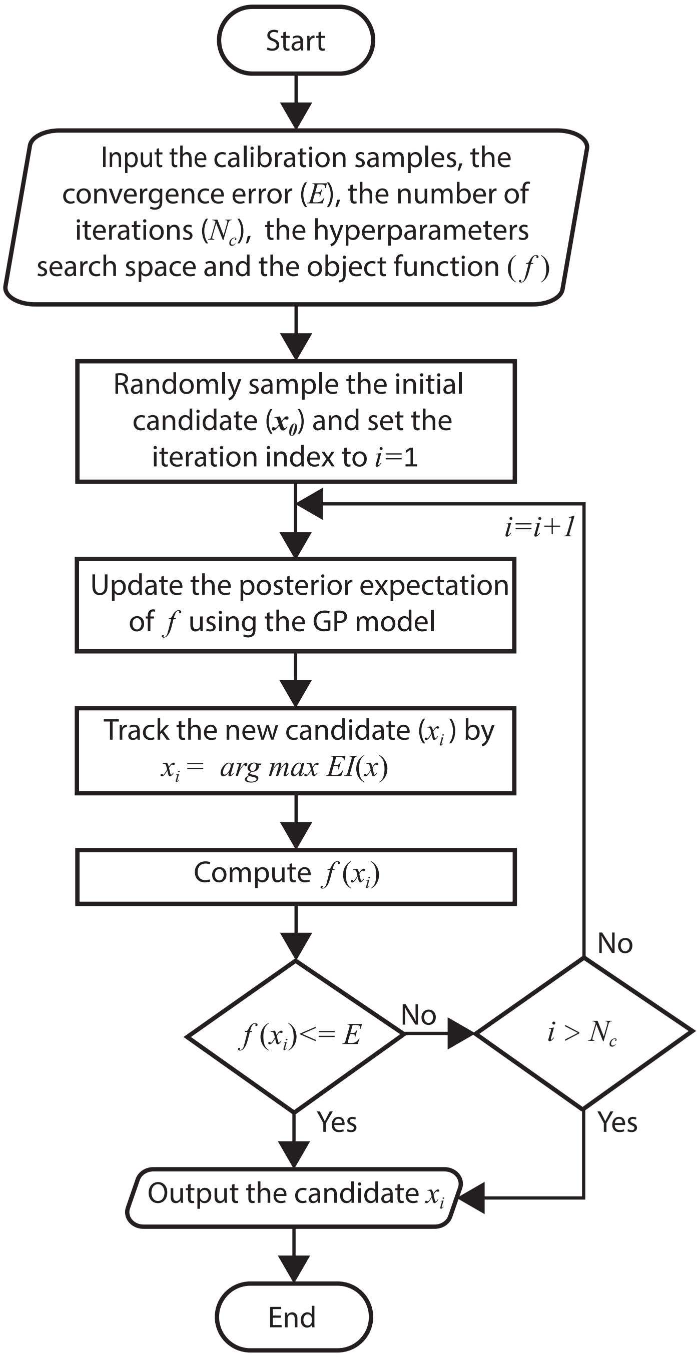
Start
Input the calibration samples, the convergence error (E), the number of iterations (N_c), the hyperparameters search space and the object function (f)
Randomly sample the initial candidate ($\boldsymbol{x_0}$) and set the iteration index to i=1
$i=i+1$
Update the posterior expectation of f using the GP model
Track the new candidate (x_i) by $x_i =$ $arg\ max\ EI(x)$
Compute $f(x_i)$
$f(x_i) <= E$
No
$i > N_c$
No
Yes
Yes
Output the candidate x_i
End

Figure 1: Flowchart of the Bayesian optimization.

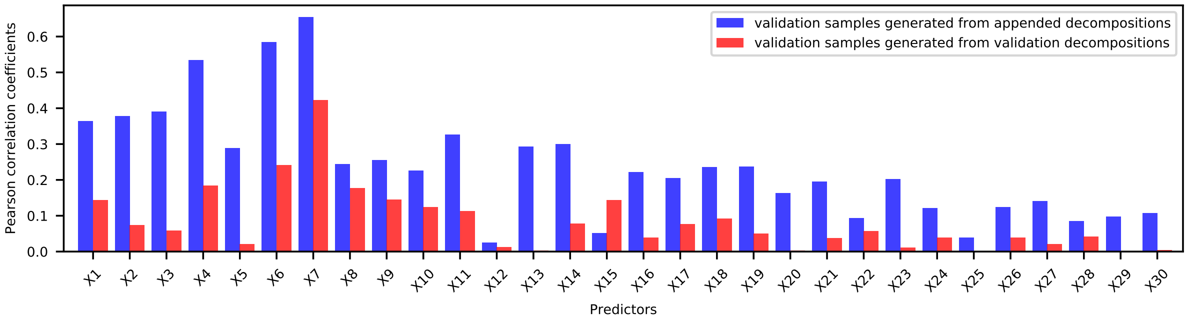

Figure 2: Absolute Pearson correlation coefficients (PCCs) between predictors and predicted targets of validation samples generated from the VMD appended decompositions and validation decompositions. The samples were collected at the Huaxian station.

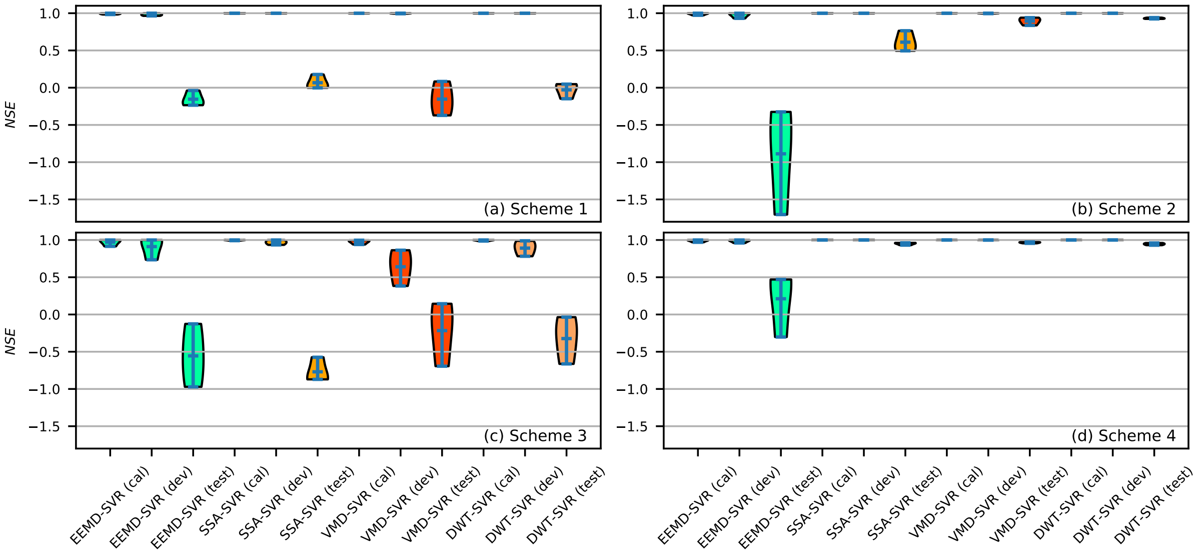

Figure 3: Violin plots of the NSE criterion for TSDP models and one-month-ahead runoff forecasting. (a) Samples generated from calibration-and-development and test decompositions without the mixing-and-shuffling step. (b) Samples generated from the calibration-and-development and appended decompositions without the mixing-and-shuffling step. (c) Samples generated from the calibration and validation decomposition with the mixing-and-shuffling step. (d) Samples generated from the calibration and appended decompositions with the mixing-and-shuffling step (the proposed TSDP framework). The entire monthly runoff series contains 792 data points, the calibration-and-development set contains 672 data points, the testing set contains 120 data points, the calibration set contains 552 points, and the validation set contains 240 data points.