Comments for hess-2019-565.

Ganggang Zuo on behalf of all the coauthors

May 7, 2020

Dealing with the boundary effects appropriately is very important in improving the generalization ability of practical decomposition-based models. This is because the boundary effects introduce decomposition errors and lead to different error distribution for calibration and validation samples. The two-stage decomposition prediction (TSDP[1]) framework proposed in this work handled the boundary effects with two key steps. One is generating validation samples from decompositions of appended sets. The appended sets obtained by sequentially appending the validation set to the calibration set. The other is mixing and shuffling the calibration and development (half validation) samples. A brief summary of why the boundary effects lead to different error distribution, how the TSDP framework deals with the error distribution, and why the TSDP framework is effective for reducing the boundary effect consequences is provided below.

1 Why the boundary effects lead to different error distribution?

The boundary effects arise at the time-series boundaries where decomposition components are extrapolated. The boundary decompositions are extrapolated because some data points before or after the given time-series, which are served as decomposition parameters, are not available ([Zhang et al., 2015], [Zuo et al., 2020]).

Due to the boundary effects, different time-series boundaries lead to some decomposition algorithms to be shift-variant and sensitive to the addition of new data. These decomposition algorithms including variational mode decomposition (VMD), discrete wavelet transform (DWT), ensemble empirical mode decomposition (EEMD), and singular spectrum analysis (SSA).

Take the VMD for example. Given the monthly runoff data of the Huaxian station from January 1953 to November 2018, i.e., x_0=[q_1, q_2, ..., q_{791}], and a one-step-ahead (shift) copy of x_0, i.e., x_1=[q_2, q_3,...., q_{792}], assume the VMD method is applied to x_0 and x_1. Then, the IMF_1(2:791) for the VMD of x_0 should be maintained by IMF_1(1:790) for the VMD of x_1 since x_0(2:791) is maintained by x_1(1:790). However, as shown in Figure 1 (a) and (b), the boundary decompositions of x_0(2:791) and x_1(1:790) are completely different. See Figure 2 (a) and (b) for DWT, Figure 3 (a) and (b) for EEMD and Figure 4 (a) and (b) for SSA. In Figure 1 - Figure 4, the IMF_1 is the first decomposed signal component of VMD and EEMD, D_1 and S_1 are the first decomposed signal component of DWT and SSA, respectively. The symbol "(2:791)" means the second data point to the 791st data point.

Given the monthly runoff data of Huaxian station from January 1953 to November 2018, i.e., x_{1-791}=[q_1, q_2, ..., q_{791}] and the monthly runoff data from January 1953 to December 2018/12, i.e., x_{1-792}=[q_1, q_2, ..., q_{792}], the IMF_1 for the VMD of x_{1-791} should be maintained by the IMF_1 (1:791)

[1]The code and data are available on `http://dx.doi.org/10.17632/ybfvpgvvsj.3`

for the VMD of x_{1-792}, since x_{1-791} is maintained by x_{1-792}(1:791). However, as shown in Figure 1 (c) and (d) the boundary decompositions of x_{1-791} and x_{1-792}(1:791) are completely different. See Figure 2 (c) and (d) for DWT, Figure 3 (c) and (d) for EEMD, Figure 4 (c) and (d) for SSA. A similar result was obtained for the case in which several data points were appended to a given time series (see Figure 1 (e) and (f) for VMD, Figure 2 (e) and (f) for DWT, Figure 3 (e) and (f) for EEMD, Figure 4 (e) and (f) for SSA).

The calibration set is usually decomposed concurrently ([Zhang et al., 2015, Tan et al., 2018]). The validation set should be decomposition one by one to avoid using future information and decomposing one validation data point will generate the "Not a Number", i.e., NaN. Thus, the validation set is sequentially appended to the calibration set and decomposed. The last decomposition of each signal component is a validation decomposition. However, sequentially append the validation set to the calibration set and decompose the appended set lead to the validation decompositions have large error distribution. This is because every validation decomposition is selected from the boundary decompositions of an appended set. The calibration decomposition errors are very small except for the boundary errors. Therefore, the calibration and validation decomposition have different error distributions (see Figure 5).

2 How does the TSDP framework deal with the introduced decomposition errors?

The different error distribution of calibration and validation decompositions leads to the models calibrated on the calibration samples generalize poorly to the validation samples. The TSDP framework improves the generalization ability of decomposition-based models by generating validation samples from the appended decompositions and mixing and shuffling the calibration samples and development (i.e., half validation) samples. Generating validation samples from appended decompositions maintains the predictors highly correlated to the predicted target. 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. Overall, the TSDP framework deals with the introduced decomposition errors without removing or correcting these errors.

Figure 6 shows the technical details of the TSDP framework and its VMD-SVR realization. The VMD and SVR in the TSDP framework can be replaced with other decomposition and data-driven models. The tuned decomposition parameters (decomposition level (K) for VMD) have to be tuned based on calibration set and remain static for decomposing the appended sets. The default decomposition parameters (the secondary penalty parameter (α), the noise tolerance (τ), and the convergence tolerance (ϵ) of VMD) remain static for decomposing both the calibration and appended sets.

3 Why the TSDP framework is effective for dealing with decomposition errors?

The boundary effects of VMD, DWT, EEMD, and SSA introduce decomposition errors and lead to different error distribution of calibration and validation decompositions. But in our work, we have demonstrated that the TSDP framework can reduce the consequences of boundary effects and provide accurate out-of-sample forecasts.

We directly built models based on the decompositions with introduced errors due to two facts. One fact is that decomposition errors don't affect the reconstructed results of decomposed signal components. Figure 1 (h), Figure 2 (h), Figure 3 (h), and Figure 4 (h) show that the summation of decomposed signal components with decomposition errors can precisely reproduce the original signal. The summation of

VMD signal components cannot completely reproduce the original signal. This is because the noise tolerance (τ) of VMD was set to zero and some noise components were removed. In other words, the decomposition errors don't lead the models calibrated well on these signal components to output poor ensemble forecasts. Another fact is that the calibration and validation samples may have a different distribution in practical forecasting scenarios ([Ng, 2017]). A classical instance is a cat classifier calibrated on the cat pictures with a high revolution (which are collected by a professional camera) generalizes poorly to the cat pictures with a low revolution (which are collected by a phone camera). Training a classifier on the calibration and validation distribution can improve the generalization ability. Therefore, we believe the different error distribution of calibration and validation samples can be handled properly by assessing the validation error distribution during the calibration stages. These two facts can prove that it is ok to build decomposition-based models without removing or correcting decomposition errors. In fact, our view is that the decomposition errors might contain some valuable information to build forecasting models.

References

[Ng, 2017] Ng, A. (2017). Machine learning yearning. *URL: http://www. mlyearning. org/(96)*.

[Tan et al., 2018] Tan, Q.-F., Lei, X.-H., Wang, X., Wang, H., Wen, X., Ji, Y., and Kang, A.-Q. (2018). An adaptive middle and long-term runoff forecast model using eemd-ann hybrid approach. *Journal of hydrology*, 567:767–780.

[Zhang et al., 2015] Zhang, X., Peng, Y., Zhang, C., and Wang, B. (2015). Are hybrid models integrated with data preprocessing techniques suitable for monthly streamflow forecasting? some experiment evidences. *Journal of Hydrology*, 530:137 – 152.

[Zuo et al., 2020] Zuo, G., Luo, J., Wang, N., Lian, Y., and He, X. (2020). Decomposition ensemble model based on variational mode decomposition and long short-term memory for streamflow forecasting. *Journal of Hydrology*, 585:124776.

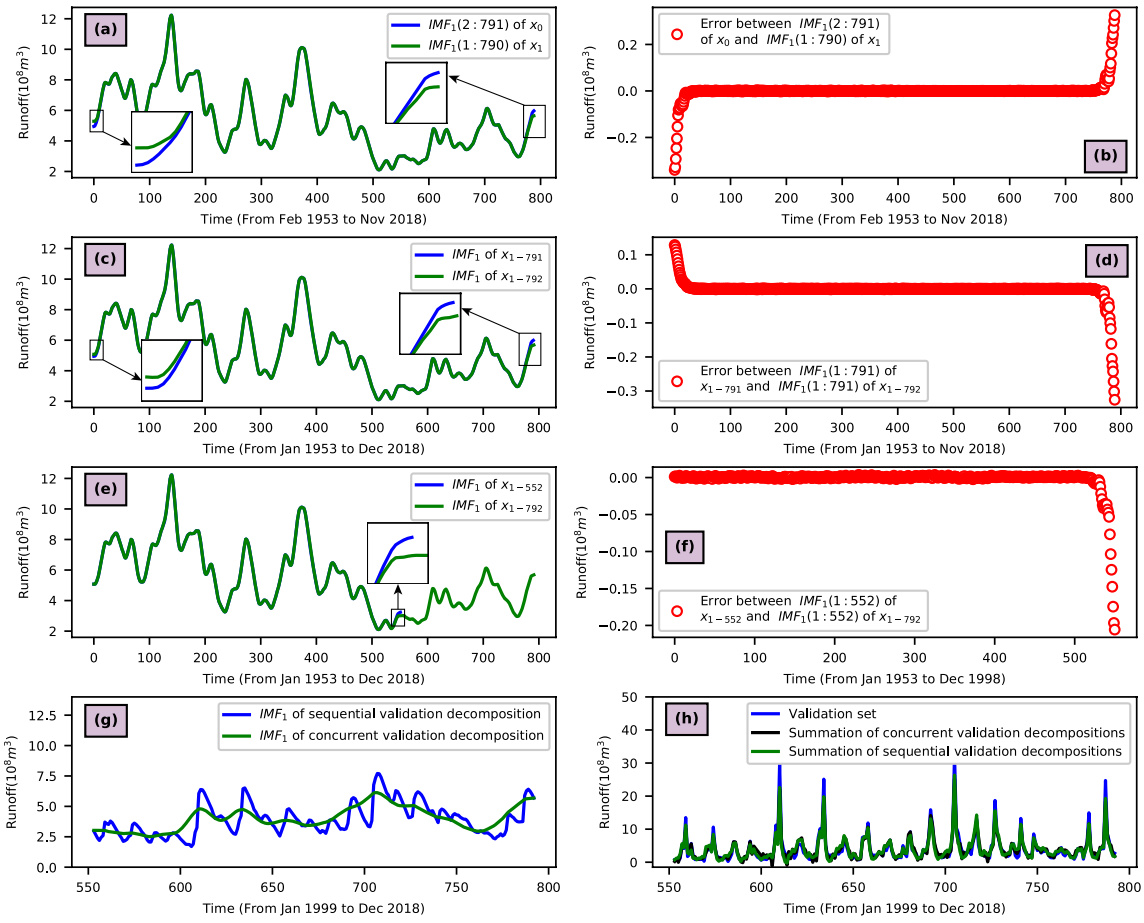

Figure 1: Diagram of boundary effect for illustrating instances of VMD (a and b) shift-variance, (c and d) sensitivity of appending one data point, (e and f) sensitivity of appending several data points, (g) difference between sequential and concurrent validation decompositions and (h) difference between the summation of sequential and concurrent validation decompositions at Huaxian station.

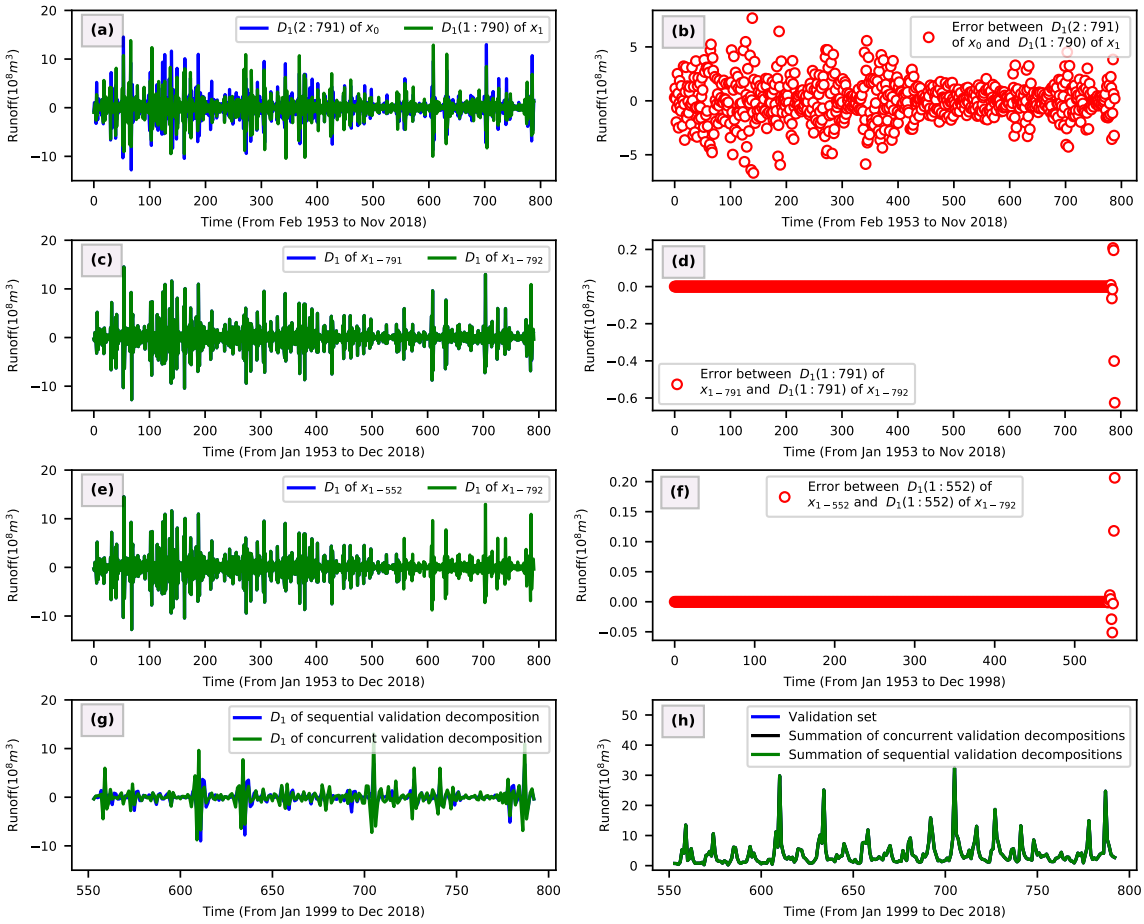

Figure 2: Diagram of boundary effect for illustrating instances of DWT (a and b) shift-variance, (c and d) sensitivity of appending one data point, (e and f) sensitivity of appending several data points, (g) difference between sequential and concurrent validation decompositions and (h) difference between the summation of sequential and concurrent validation decompositions at Huaxian station.

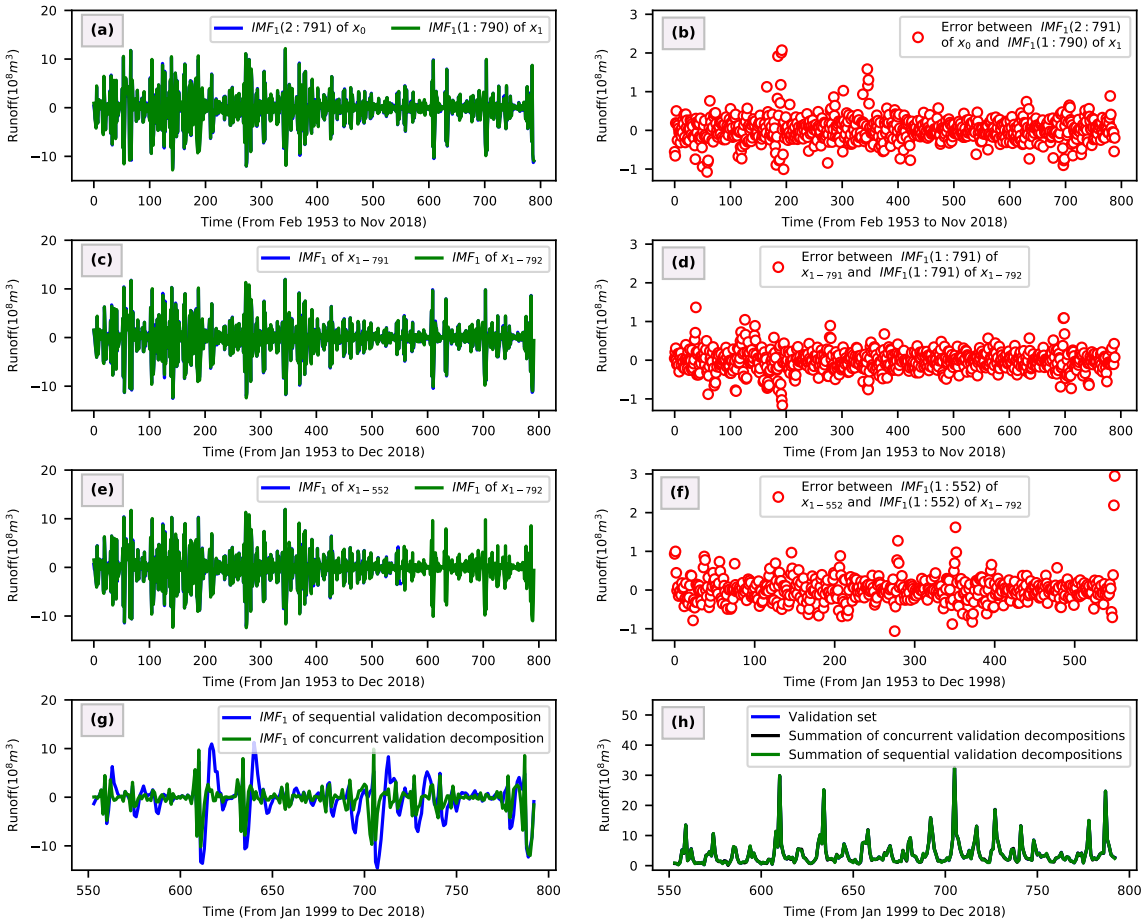

Figure 3: Diagram of boundary effect for illustrating instances of EEMD (a and b) shift-variance, (c and d) sensitivity of appending one data point, (e and f) sensitivity of appending several data points, (g) difference between sequential and concurrent validation decompositions and (h) difference between the summation of sequential and concurrent validation decompositions at Huaxian station.

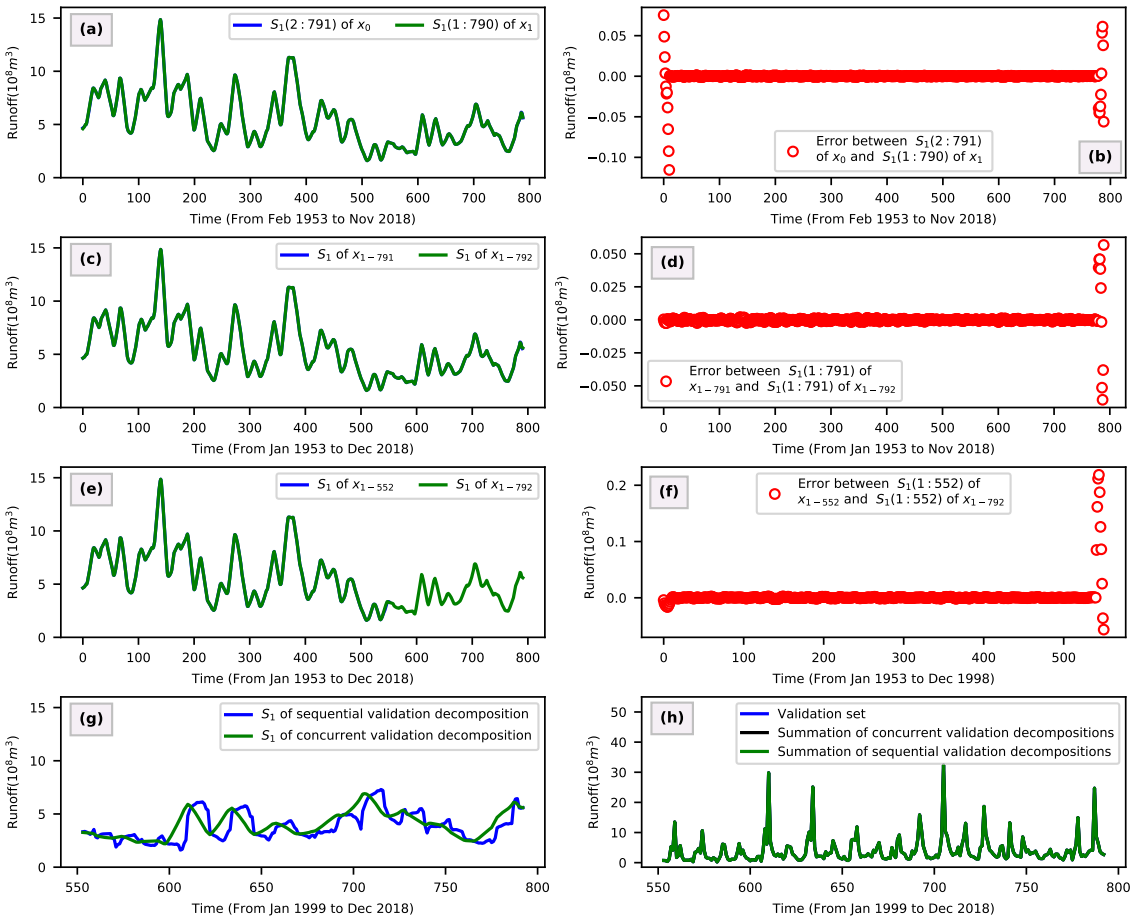

Figure 4: Diagram of boundary effect for illustrating instances of SSA (a and b) shift-variance, (c and d) sensitivity of appending one data point, (e and f) sensitivity of appending several data points, (g) difference between sequential and concurrent validation decompositions and (h) difference between the summation of sequential and concurrent validation decompositions at Huaxian station.

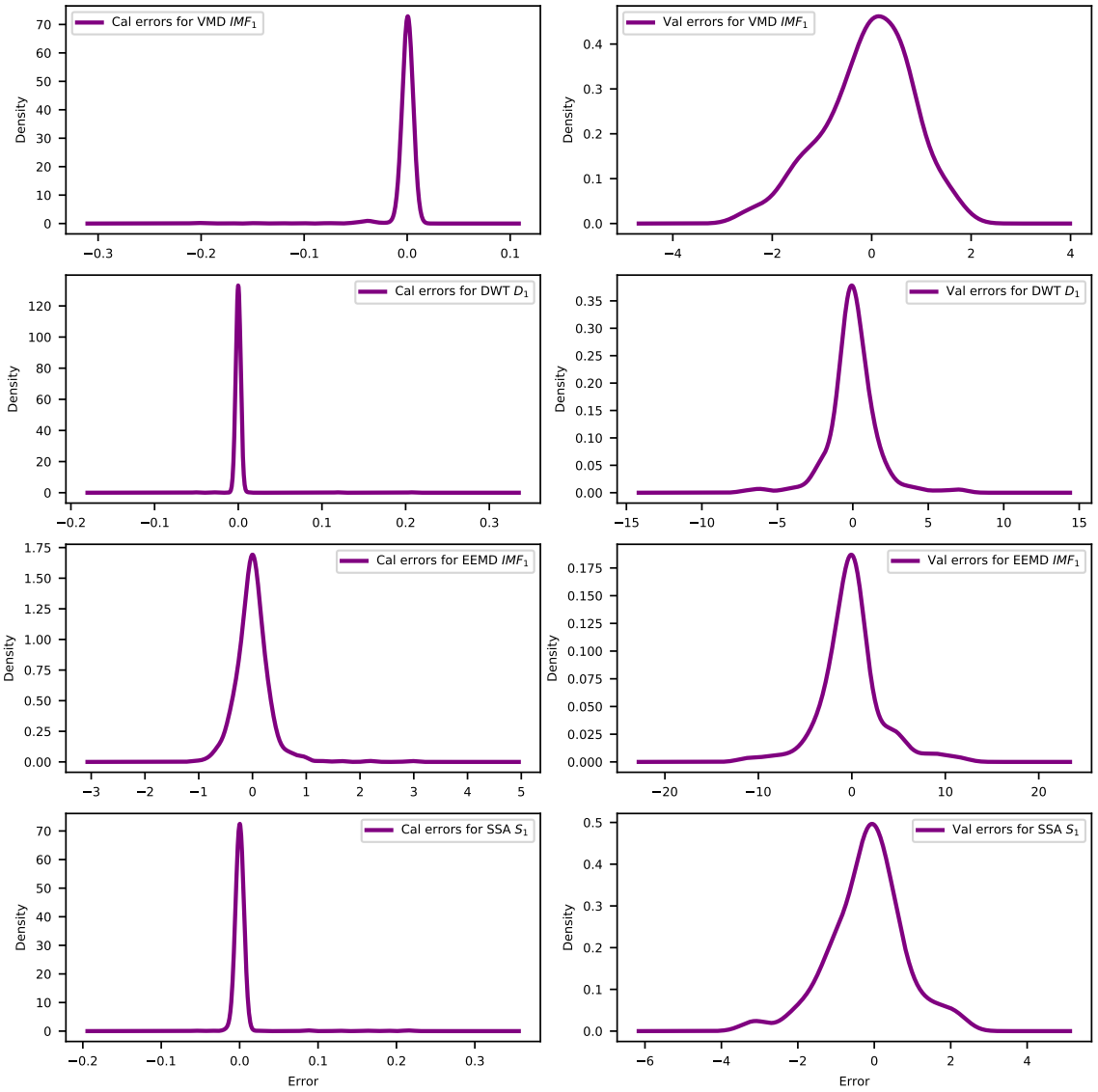

Figure 5: Kernel density estimation of calibration and validation decomposition errors (kernel=Gaussian).

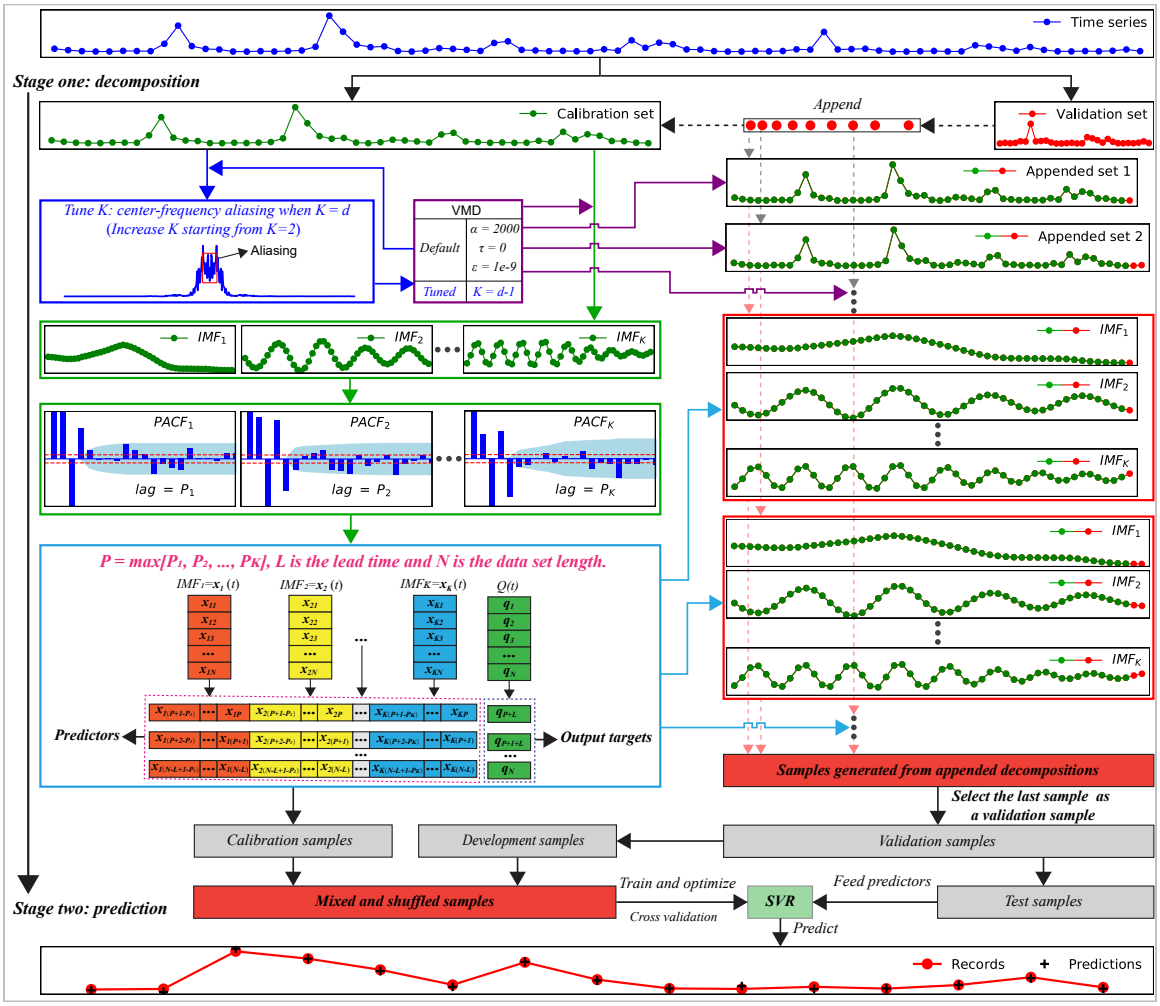

Figure 6: A block diagram of the two-stage decomposition prediction (TSDP) framework with the VMD-SVR realization.