

### S1. Determination of ARX model orders $p$ and $k$

To determine the autoregressive order  $p$  of the ARX model, the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are employed for preliminary screening. The calculation formulas are as follows:

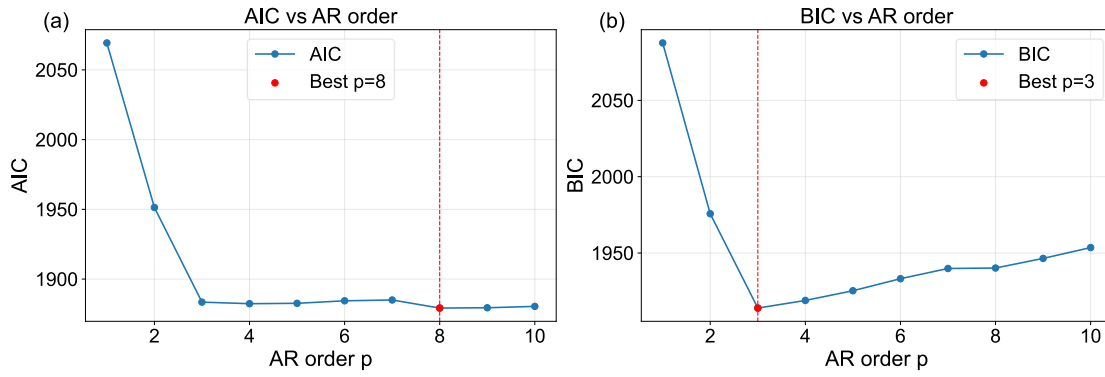
$$AIC = -2 \ln(L) + 2k \quad (S1)$$

$$BIC = -2 \ln(L) + k \ln(n) \quad (S2)$$

where  $L$  denotes the maximum likelihood function,  $k$  is the number of parameters, and  $n$  represents the sample size. Both criteria aim to balance fitting performance against model complexity. Notably, BIC incorporates a sample-size penalty, imposing stricter constraints on high-order models when the sample size is large, thereby helping to prevent overfitting.

The AIC and BIC values are calculated for candidate orders  $p$  from 1 to 10 to identify the theoretically optimal order, as shown in Fig. S1. The AIC suggests an optimal order of  $p = 8$ , while the BIC suggests  $p = 3$ . Subsequently, to determine the lag order  $k$  for simulated streamflow, the candidate orders derived from the preliminary step ( $p = 3$  and  $p = 8$ ) are combined with varying  $k$  values to identify the optimal configuration based on model performance. Considering that  $k$  physically represents the dependence of error on historical simulated streamflow—a relationship that typically decays rapidly over time—a grid search is conducted within the physically reasonable range of  $k \in [1,5]$ . The evaluation results for each model combination are presented in Table S1.

Table S1 indicates that the model with  $k = 0$  performs significantly worse than those with  $k > 0$ , confirming the necessity of incorporating simulated streamflow into the error post-processing. For models with  $k > 0$ , the performance differences are marginal. However, since models with higher complexity are more prone to overfitting and instability, the combination of  $p = 3$  and  $k = 3$  is selected as the final configuration to strike a balance between model performance and parameter efficiency.



30

**Figure S1.** AIC and BIC values for candidate autoregressive orders  $p$  ranging from 1 to 10.

**Table S1.** Performance evaluation of ARX models with varying autoregressive  $p$  and simulation lag  $k$  orders ( $n$  represents the number of model parameters).

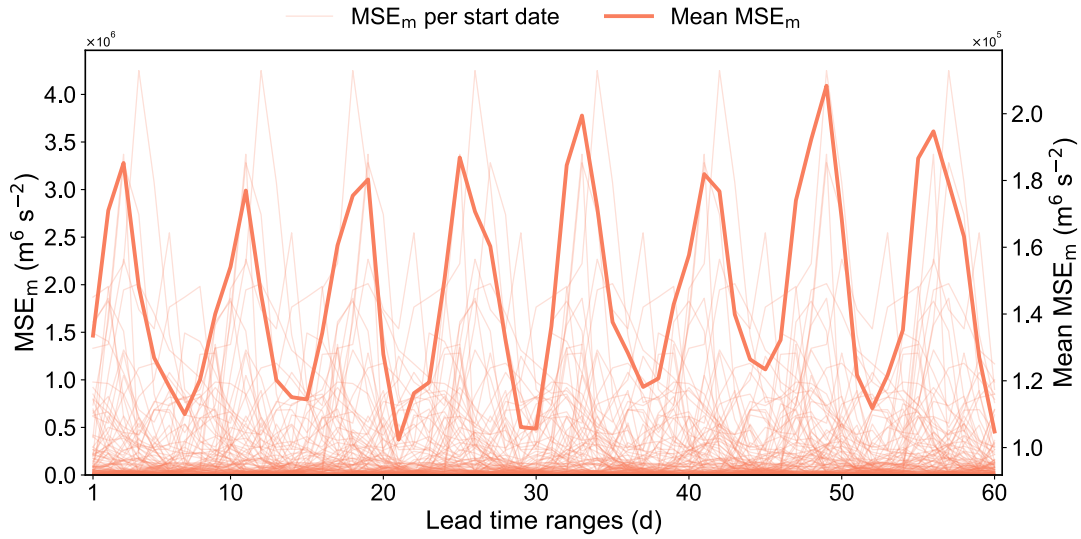
$p$	$k$	$n$	AIC	BIC	RMSE	RE	NSE
8	3	13	802.33	881.57	264.75	12.382	0.94949
8	4	14	804.03	889.37	264.90	12.389	0.94943
8	5	15	804.49	895.92	265.04	12.378	0.94938
3	5	10	802.86	863.83	265.29	12.404	0.94929
3	4	9	803.83	858.70	266.77	12.444	0.94872
<b>3</b>	<b>3</b>	<b>8</b>	<b>801.20</b>	<b>849.98</b>	<b>267.26</b>	<b>12.453</b>	<b>0.94853</b>
8	2	12	814.38	887.53	268.56	12.463	0.94803
8	1	11	813.44	880.49	270.36	12.492	0.94733
3	2	7	813.15	855.83	274.73	12.577	0.94561
3	1	6	812.06	848.64	277.10	12.614	0.94467
3	0	5	1883.47	1913.95	336.32	13.838	0.91849
8	0	10	1879.22	1940.17	354.43	14.231	0.90948

35

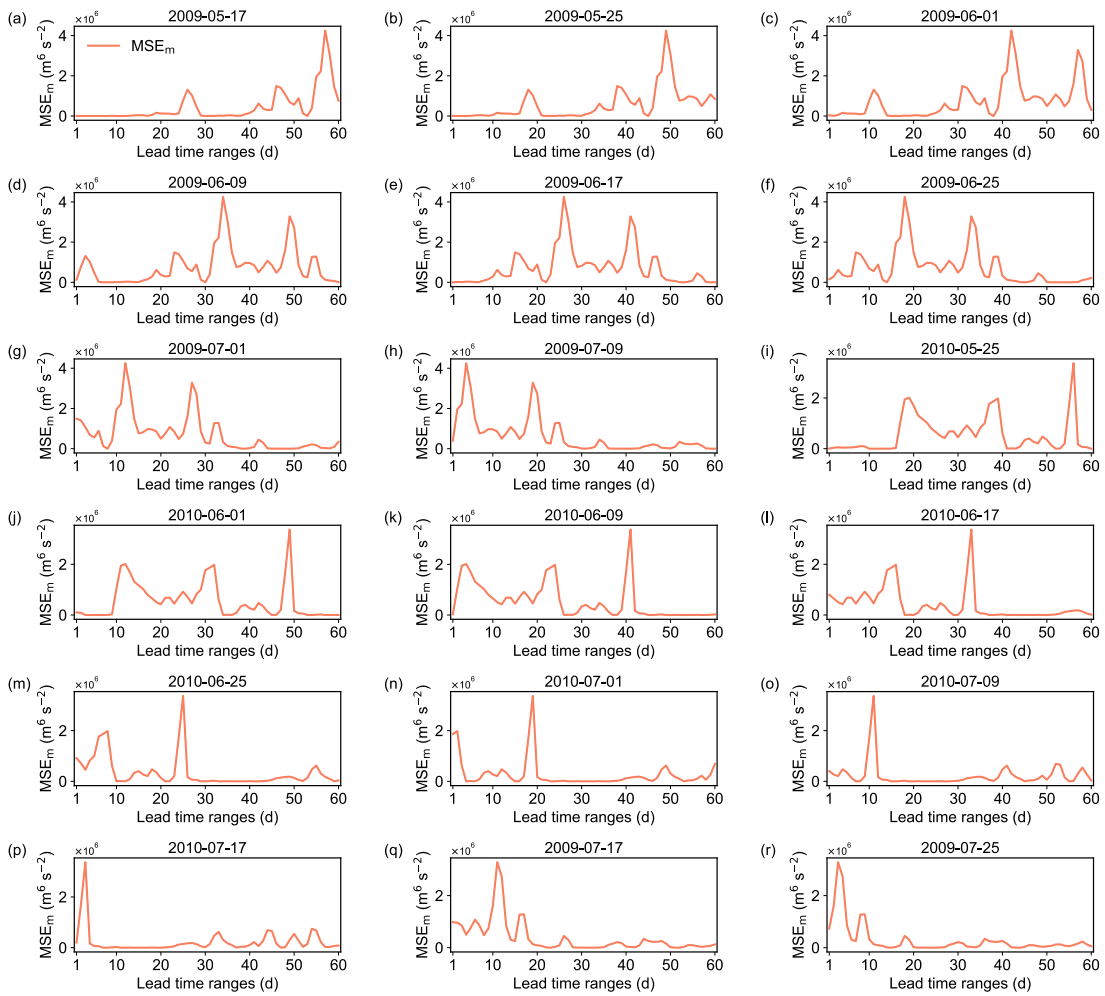
## S2. Interpretation of periodic fluctuations in MSE

The significant 8-day periodic oscillation observed in the MSE along the forecast lead time is primarily attributed to the interaction between the 8-day initialization interval adopted in this study and the 60-day forecast horizon. This discontinuous initialization strategy (e.g., forecasts initiated on the 1st, 9th, 17th, and 25th of most months) induces a systematic overlap of forecast sequences along the verification timeline. Specifically, the forecast for lead time  $L$  from a current initialization date targets the exact same calendar date as the forecast for lead time  $L + 8$  from the previous initialization (8 days prior). This 8-day temporal setup implies that lead times separated by an 8-day interval resample the same set of hydrological events through different time windows. Consequently, errors stemming from specific events are repeatedly mapped onto these correlated lead times.

Within this 8-day oscillation, peak MSE values are significantly concentrated at specific lead time phases (e.g., leads 3, 11, 19...). This phenomenon arises from the interplay between the temporal heterogeneity of hydrological errors and the discrete sampling strategy. Given that MSE is highly sensitive to extreme values, the overall error magnitude is often dominated by forecast deviations from a few extreme hydrological events (Fig. S2). Under the 8-day sampling grid of this study, once a specific target date yielding a large error is fixed, its temporal distance (i.e., lead time) relative to different initialization dates strictly follows an arithmetic progression with a common difference of 8 ( $L, L + 8, L + 16 \dots$ ). For instance, if an extreme event occurs on the 3rd day following an initialization date, it will necessarily correspond to the 11th day of the previous forecast (initiated 8 days earlier) and the 19th day of the one before that (Fig. S3). This systematic alignment ensures that the substantial error contributions from sparse extreme events are not uniformly distributed but are concentrated on these lead times separated by 8-day intervals, thereby causing distinct peaks in the average MSE curve at these specific nodes.



65 **Figure S2.** MSE series over 60 forecast lead days for individual initialization dates and the average of all MSE series.



**Figure S3.** Selected MSE series for individual initialization dates with the largest maximum MSE values.

70