Reviewer #2:

This work proposes a two-step method, i.e., from basin scale to grid scale, to produce a water budget closed datasets by introducing the Bayesian model with predefined prior distribution and posterior parameter estimation, considering the covariance between water components and the entire time series, under a specific case study in the severely irrigated Hindon River Basin, India. This work tried to introduce an innovative theoretical basis and apply it to a basin with abundant discharge records. Although the logical structure of the work is clear, the theoretical introduction and the equations are hard to follow since there is no deductive process of the equations provided and the code and data are not accessible with the provided links. This makes it hard for the readers to follow the work and estimate the robustness of the work.

We would like to thank the referee for the time and effort reviewing our manuscript, and for the valuable feedback received. We understand that the theoretical introduction of the methods and equations might be hard to follow. We will improve the flow of equations by adding intermediate steps in the equations that more clearly show how one equation leads to the next. Also, we will update the link to the data and software, as follows:

"Code and data availability:
The software used to collect the data is available at https://doi.org/10.5281/zenodo.11148992, and the source code developed for this research is available at http://doi.org/10.5281/zenodo.4116451."

Reviewer major comment 1:
"the spatial scale problem is not thoroughly discussed. How were the different scales between water components in the budget closure equation handled? Is the resolution 50 km really feasible in such a small basin that is only one pixel wide and two pixel height?"

Reply on major comment 1:

We agree this point deserves more attention, and we will add the following paragraph before Table 1 to reflect on the challenge of the scale difference between the water balance variables:

"As can be seen in Table 1, the native resolution of the different individual data sets and between the water balance variables varies widely, posing a challenge in their merging process. In principle, the water balance data fusion can be performed at any spatial resolution; however, for this study design, we try to preserve the native information as much as possible by choosing to resample all precipitation data sets to a common resolution of 0.05° and all evaporation data sets to a common resolution of 250m. An alternative design would be to resample all water balance variables to the same resolution (e.g., 250m), but this might introduce artefacts, for example, when up-sampling from coarse to high resolution."

As for the 50 km, this refers to the spatial correlation length (l_s) that we specify a priori to generate the covariance matrices of precipitation (Eqs. 5-6) and evaporation. The basic assumption here is that the values of the grid cells in the spatial domains of precipitation and evaporation are spatially correlated, i.e., the correlation between the grid values depends on the distance between the grid locations. Grid cells that are close to each other tend to have high correlation. As the spatial separation between grid cells increases, the correlation decreases. Often, information on the correlation length parameter is not available. In such a case, and given the differences in scales between the data sets, we rely on specifying a mid-range value of

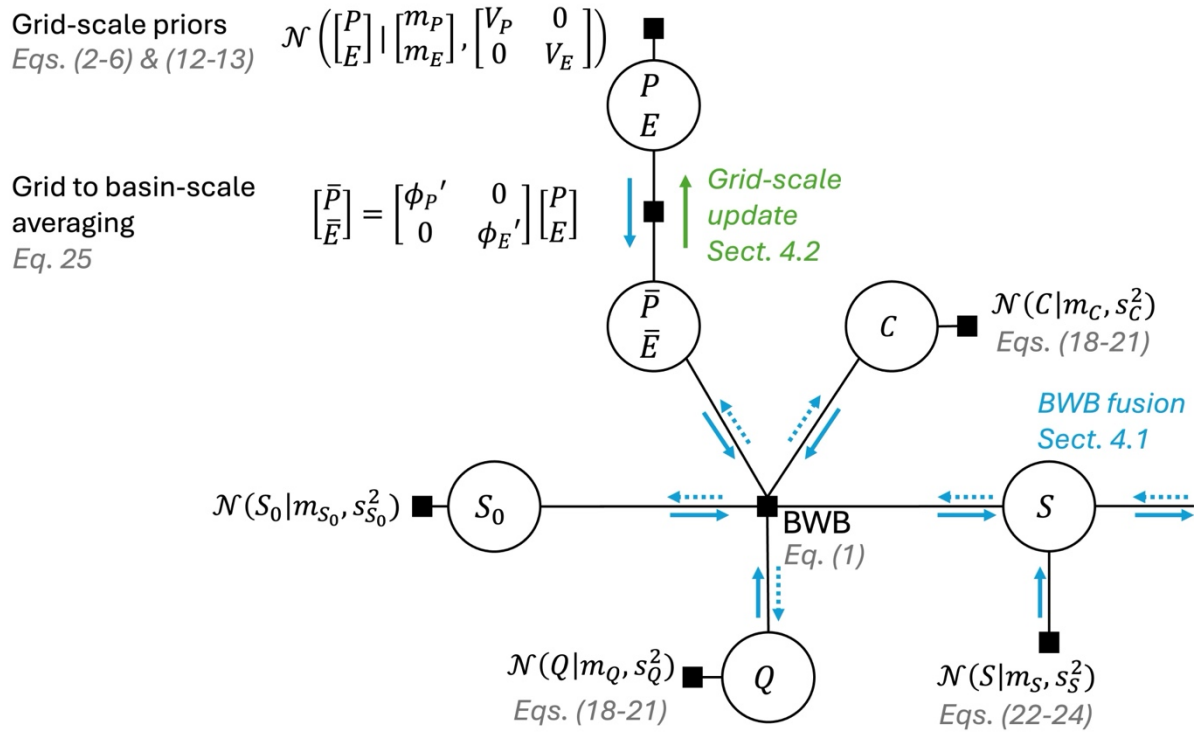
the prior correlation length, which is about half the basin's length from North to South (line 238). Finally, we assess the sensitivity of the results to the chosen correlation length in sect 6.2.

Reviewer major comment 2:
"L173/L884-885. About the gap-filling method, the authors made the assumption that the canal operations do not vary widely between years in which the conditions are similar. Is it possible to use the existing data to validate the assumption? I mean compare the data in the years with similar conditions to check whether that assumption is tenable."

Reply on major comment 2:
Unfortunately, we are constrained by the amount of data available, making cross-validation difficult. Specifically, we at most have five years of data, and these years can be distributed between wet, dry, and normal years, leaving us with no out-of-sample data for cross-validation. E.g., if among these 5 years, 2 years are dry, we simply took the average of these two years to fill all other missing dry years. In cases where we know that a distributary might significantly contribute to the overall imports to the basin, e.g., the Deoband branch, we follow a different approach that avoids any average-based related biases and relies on using the design discharge and operation time to estimate the canal water imports for the missing years. In this vein, using the canal design capacities for gap-filling results in conservative (upper bound) estimates of the irrigation water imports, which can also be seen as acceptable approximate initial values for the Bayesian methods used in the study. Additionally, in sect 6, we include scenario 2, where, instead of filling the missing years with the data average of years with similar conditions, we use the upper bound estimates (design capacities multiplied by the operation time). The results show insensitivity to the canal water estimates used, as these are much smaller than the surface exchanges (e.g., precipitation and evaporation). To account for potential uncertainties around all assumptions we make to generate the corresponding full timeseries, we attach a large prior uncertainty (25%) to these gap-filled estimates, i.e., a wide confidence interval around their values.

Reviewer major comment 3:
"It will be much clearer and easier to follow if a framework diagram is provided in the method section."

Reply on major comment 3:
Good suggestion, a flow diagram can aid the reader in grasping the overall flow of the methods section. We will add the following diagram:



"Figure caption: a factor graph showing the variables in circles and constraints as squares for a single month. Three constraints are incorporated in the water balance data fusion method, including: a prior Gaussian distribution assigned to each water balance variable (squares attached to each variable), the basin-scale water balance (BWB) constraint that links all water balance variables together, and the spatial averaging constraint that links grid-scale and basin-scale variables for precipitation and evaporation. Computation of posteriors of individual variables proceeds in two steps. The first step is basin-scale water balance data fusion (BWB fusion, see section 4.1). This step involves multiple forward (blue arrows along the edges) and backward (dotted blue arrows) passes over the data using the entire timeseries to compute the posteriors of all water balance variables that jointly close the water balance (Schoups and Nasseri, 2021). The second step computes grid-scale posteriors for precipitation and evaporation from their basin-scale posteriors using a Kalman smoothing algorithm (green arrow, see sect. 4.2)."

Reviewer major comment 4:
"About the matrix variables mentioned in all equations, for example, in Eq. 5, it is better to provide the size parameters of each matrix."

Reviewer major comment 5:
"For me, the relationships between equations are quite independent and the connections are weak. For example, Eq. 3-8, the input and output of each equation are vague thus hard to understand the method itself as a whole. The same issue exists for the entire theoretical part."

Reply on major comments 4 and 5:
We will add the information about the size of the evaporation and precipitation correlation matrices. These are now mentioned in the following modified sections, which have been modified to show additional intermediate steps when going from one equation to the next.

Sect. 3.1.1

The independence structure between P_t and E_t shown in Eq. (2) allows us to represent the errors in each variable separately. For each month *t*, we typically have multiple gridded precipitation products, with unknown bias and random errors. To characterize these errors, the grid-scale precipitation bias and random error models for all elements in the mean vector m_{P_t} and the square root of all diagonal entries of the autocovariance matrix V_{P_t} are defined using Eqs. (3-4), respectively:

$$m_{P_t} = P_t^{obs,min} + w_p\left(P_t^{obs,max} - P_t^{obs,min}\right) \quad (3)$$

$$s_{P_t} = r_P \frac{1}{4}\left(P_t^{obs,max} - P_t^{obs,min}\right) \quad (4)$$

Equation (3) models the systematic bias in the gridded precipitation for each month *t* by describing the precipitation prior mean m_{P_t} as a function of two parts. The first part represents the baseline precipitation estimated as the minimum across all the products ($P_t^{obs,min}$), while the second consists of a weighted deviation from this base value. Specifically, the latter denotes the full range of the grid-scale observed precipitation ($P_t^{obs,max} - P_t^{obs,min}$) with a weight parameter w_p. The bias parameter w_P takes on an unknown value between 0 and 1, with a logit-normal prior distribution of parameter $\mu = 0$ and scale parameter $\sigma = 1.4$ to reflect prior uncertainty about the bias. In other words, w_P controls the relative position of the m_{P_t} within the observed precipitation space (range). As it approaches 0, more weight is given to the minimum across all products ($P_t^{obs,min}$), whereas approaching a value of 1 gives more weight to the maximum across all products ($P_t^{obs,max}$).

Random errors in the precipitation are modeled using Eq. (4). This model expresses the prior gridded standard deviation s_{P_t} for each month *t* as a function of the maximum potential random error, defined as a quarter of the range: $\frac{1}{4}\left(P_t^{obs,max} - P_t^{obs,min}\right)$. A noise parameter ($r_P$) is used to scale this conservative quantity and is given a quasi-uniform prior distribution between 0 and 1.

To account for the effect of spatial correlation of the random error component (Eq. (4)), we write the precipitation prior covariance matrix V_{P_t} in terms of the grid-scale standard deviations and a grid-scale auto-correlation matrix:

$$V_{P_t} = S_P R_P S_P \quad (5)$$

where S_P is a diagonal matrix containing the grid-scale s_{P_t} values for all locations of the spatial field (Eq. (4)), and R_P is the correlation matrix that captures the spatial dependence structure. $R_P \in \mathbb{R}^{n_P \times n_P}$, where $n_P \times n_P$ is the matrix dimension, and n_P equals 176, representing the total number of grid cell locations of the precipitation spatial domain. R_P is jointly estimated from all precipitation data using an isotropic parametric correlation function with the following form (Handcock and Stein, 1993):

$$C_{\mathcal{M}}(d|l_s, v) = \frac{1}{2^{v-1}\Gamma(v)}\left(\frac{d}{l_s}\right)^v K_v\left(\frac{d}{l_s}\right) \quad (6)$$

where $C_{\mathcal{M}}$ is the Matérn correlation function for variables separated by distance d. This correlation model is flexible and widely used, with two functions: gamma function $\Gamma(.)$ and the modified Bessel function $K_v(.)$ (Abramowitz and Stegun, 1968). $C_{\mathcal{M}}$ also consists of two unknown nonnegative parameters, namely the spatial correlation length scale l_s and a spatial smoothness parameter v. A value of v approaching 0 indicates a rough spatial process, while the process is smoother when v approaches infinity. Since the smoothness parameter is usually small in many applications (Chen et al., 2022), while it increases as the aggregation time increases (Sun et al., 2015), we choose a balanced value between a rough and smooth random field, i.e., v , fixed at 1.5. On the other hand, the correlation length scale (l_s) defines an average length scale on which grid cells are correlated with each other. In principle, l_s ranges from 0 (the case of uncorrelated grid cells) and extends to a scale larger than the spatial domain length

(the case of maximally correlated pixels). With no prior information on the l_s parameter, we fix it at 50 km (~$1/2$ the basin's length from North to South). The sensitivity of the results to the fixed l_s will be evaluated in section 6.2.

Since water balance data fusion (Schoups and Nasseri, 2021) uses basin-scale error models, we derive these from the above-described grid-scale models by spatial averaging. Specifically, the basin-scale prior mean $m_{\bar{P}_t}$, variance $v_{\bar{P}_t}$, and standard deviation $s_{\bar{P}_t}$ in month *t* follow from Eqs. 3, 4 and 5:

$$m_{\bar{P}_t} = \phi_P' m_{P_t} = \phi_P' \left[P_t^{obs,min} + w_p \left(P_t^{obs,max} - P_t^{obs,min}\right)\right] \quad (7)$$

$$v_{\bar{P}_t} = \phi_P' V_P \phi_P = \phi_P' S_P R_P S_P \phi_P = \phi_P' (r_P D_P) R_P (r_P D_P) \phi_P \quad (8)$$

$$s_{\bar{P}_t} = \sqrt{v_{\bar{P}_t}} = r_P \sqrt{\phi_P' D_P R_P D_P \phi_P} \quad (9)$$

$$\overline{P}_t \sim \mathcal{N}\left(m_{\bar{P}_t}, s_{\bar{P}_t}^2\right) \quad (10)$$

$$\overline{P}_t \geq 0 \quad (11)$$

where ϕ_P is the spatial averaging operator used to derive basin-scale moments from grid-scale moments (i.e., $n_P \times 1$ vector with each element equal to $1/n_P$, where n_P is the number of spatial locations in the precipitation spatial field). ϕ_P' is the transpose of ϕ_P. We also used $S_P = r_P D_P$, where D_P is a diagonal matrix containing the $\frac{1}{4}\left(P_t^{obs,max} - P_t^{obs,min}\right)$ values (from Eq. 4) for all grid cells within the precipitation spatial domain. All basin-averaged input quantities to Eqs. (7-8) are precomputed from the precipitation data sets, and the constant but unknown parameters w_p and r_P are estimated as part of the water balance data fusion (see Sect. 4.1). Finally, the last two equations in the precipitation error model treat the basin-scale calibrated precipitation $\overline{P}_t$ for each month *t* as a random draw from a truncated normal distribution. The truncation at zero ensures physical consistency (nonnegative precipitation).

Sect. 3.1.2

As with precipitation, an evaporation error model with the following range-based form is adopted:

$$m_{E_t} = \mathfrak{f}_{\boldsymbol{E}} \left[E_t^{obs,min} + w_E \left(E_t^{obs,max} - E_t^{obs,min}\right)\right] \quad (12)$$

$$s_{E_t} = r_E \frac{1}{4}\left(E_t^{obs,max} - E_t^{obs,min}\right) \quad (13)$$

The bias in evaporation is modeled with two spatial and time-invariant calibration parameters, namely: w_E and $\mathfrak{f}_{\boldsymbol{E}}$. The parameter w_E is the weight that interpolates between the monthly gridded evaporation extrema $E_t^{obs,min}$ and $E_t^{obs,max}$ in each month *t*. An additional scaling factor ($\mathfrak{f}_{\boldsymbol{E}}$) is incorporated to account for potential bias outside the observed range. On the other hand, Eq. (14) quantifies evaporation prior uncertainty ($\frac{1}{4}\left(E_t^{obs,max} - E_t^{obs,min}\right)$) scaled via the spatially and temporally constant noise parameter (r_E). All parameters are treated as random variables with prior distributions. The w_E and r_E parameters are specified with a vague prior distribution bounded between 0 and 1, specifically, flat logit-normal with location parameter $\mu = 0$ and scale parameter $\sigma = 1.4$. Whereas, $\mathfrak{f}_{\boldsymbol{E}}$ is given a lognormal prior with mode at 1 (no bias) and a coefficient of variation CV of 50%.

The basin-scale evaporation error models are derived from the grid-based models defined above, using the same spatial averaging process applied to precipitation (Eqs. (7-8)). The averaging formulas are then obtained as follows:

$$m_{\bar{E}_t} = \phi_E' m_{E_t} = \mathfrak{f}_{\boldsymbol{E}}\, \phi_E' \left[E_t^{obs,min} + w_E \left(E_t^{obs,max} - E_t^{obs,min}\right)\right] \quad (14)$$

$$s_{\bar{E}_t} = r_E \sqrt{\phi_E' D_E R_E D_E \phi_E} \quad (15)$$

$$\overline{E_t} \sim \mathcal{N}\left(m_{\bar{E}_t}, s^2_{\bar{E}_t}\right) \quad (16)$$

$$\overline{E_t} \geq 0 \quad (17)$$

where D_E is a diagonal matrix whose diagonal entries containing the $\frac{1}{4}\left(E_t^{obs,max} - E_t^{obs,min}\right)$ values for all grid cells within the evaporation spatial domain. All inputs of Eqs. (15-16) are precomputed from the evaporation data sets while the unknown parameters $\mathbf{f_E}$ and r_E are solved as part of the water balance data fusion (Sect. 4). ϕ_E is the spatial averaging operator, and the R_E term stands for the correlation matrix, which captures the spatial dependencies between the evaporation grid cells. $R_E \in \mathbb{R}^{n_E \times n_E}$, where $n_E \times n_E$ is the matrix dimension, and n_E equals 71235, representing the total number of grid cell locations of the evaporation spatial domain. For the large-sized evaporation data sets considered here, we parameterize the evaporation correlation matrix using a Matérn kernel implemented within a stochastic variational Gaussian Process (Hensman et al., 2015) with fixed parameter l_s at 50 km and ν at 1.5.
Similar to precipitation, the basin-scale calibrated evaporation $\overline{E_t}$ for month *t* is treated as a random draw from a truncated normal distribution Eqs. (17-18). The truncation at zero ensures physical consistency (nonnegative evaporation).

Reviewer major comment 6:
"L545. The labels and tick marks of x and y missed in Fig. 8".

Reply on major comment 6:
Since the river discharge data are classified (can't be made publicly available), the labels and tick marks of x and y are not shown in Fig. 8. We will update the caption of Fig. 8 as follows:

Figure 8: Distribution of Q in comparison to the observed Q_{obs} value (July 2009), for the following cases: (i) ignoring and accounting for posterior correlations between all water balance variables, (ii) ignoring only posterior correlation between *P* and *E*, and (iii) Q_{wb} obtained from uncalibrated water balance data. The labels and the tick marks are not shown due to data sharing restrictions.

Reviewer minor comment 1:
"L125 the abbreviation of Central Water Commission (CWC) should be explained near the figure instead later in L168".

Reply on minor comment 1:
That's true; we will define CWC in the Figure 1 caption as follows:

"Figure 1: Location of the Hindon basin in the Uttar Pradesh state of India (inset map); with a detailed view of the basin featuring its boundaries, topographic profile, and the location of the Galeta outlet, where a river discharge station that belongs to the Central Water Commission (CWC) network of India is located. The main map shows the irrigation scheme with reservoirs and canal system. Topographic basemap sources: Esri, USGS, FAO, NPS, GIS user community, and others."

Reviewer minor comment 2:
"L241. the symbols mpt and vpt with Eq.7-8 are different."

Reply on minor comment 2:
We have removed these notations for conciseness and to enhance clarity.

Reviewer minor comment 3:
"L255 I think "Evapotranspiration" is better than "Evaporation" throughout the paper."

Reply on minor comment 3:
At the outset of introducing the probabilistic water balance model (line 130), we define evapotranspiration as evaporation (including transpiration). Later, we use the term "evaporation" throughout the paper for conciseness.

Reviewer minor comment 4:
"L581. There is no ground-water level data? It is weird that the discharge of the canals have been paid great attention while no ground-water data available in such a heavily ground-water-based irrigated basin."

Reply on minor comment 4:
The reviewer is right, the inclusion of groundwater level as an independent evaluation would be valuable; however, in this study, we only considered the total water storage. In a follow-up study, we plan to extend the methodology presented here to incorporate detailed information, such as groundwater pumping, soil moisture, and groundwater level data, where we also focus on separating the rootzone from the groundwater contribution. See also our response to Reviewer 1's comment on validation.

Abramowitz, M. and Stegun, I. A.: Handbook of mathematical functions with formulas, graphs, and mathematical tables, US Government printing office1968.
Chen, H., Ding, L., and Tuo, R.: Kernel packet: An exact and scalable algorithm for Gaussian process regression with Matérn correlations, Journal of machine learning research, 23, 1-32, 2022.
Handcock, M. S. and Stein, M. L.: A Bayesian analysis of kriging, Technometrics, 35, 403-410, 1993.
Hensman, J., Matthews, A., and Ghahramani, Z.: Scalable variational Gaussian process classification, Artificial intelligence and statistics, 351-360,
Schoups, G. and Nasseri, M.: GRACEfully closing the water balance: A data-driven probabilistic approach applied to river basins in Iran, Water Resources Research, 57, e2020WR029071, 2021.
Sun, Y., Bowman, K. P., Genton, M. G., and Tokay, A.: A Matérn model of the spatial covariance structure of point rain rates, Stochastic Environmental Research and Risk Assessment, 29, 411-416, 2015.