Reviewer #1:

“This manuscript presents a novel probabilistic water balance data fusion approach for calibrating multi-scale hydrological datasets. The methodology is innovative, addressing the challenge of reducing uncertainties in datasets by integrating them through water balance constraints. The approach provides a framework for both basin-scale and pixel-scale applications. The application to the Hindon River Basin demonstrates practical utility, with reasonable error estimates and clear improvements in data consistency. The paper is well-written, structured, and accessible, making a substantial contribution to water resource management and hydrological modeling. However, some areas, such as the clarity of methodological details and validation against independent data, could be strengthened to enhance the robustness and reproducibility of the findings. Suggestions are as follows:”

We would like to thank the referee for the time and effort reviewing our manuscript, and for the valuable feedback received. We are pleased that the reviewer found the paper well written and appreciate the recognition of the novelty of the presented methods. In the following, we address the reviewer’s detailed comments.

Reviewer comment 1:
“In Section 2, beginning on line 117, you describe the Hindon Basin and the separation of two irrigation seasons (Kharif and Rabi), yet it is unclear if the rotated crops use the same land or if they are in adjacent regions. It would be helpful to add a sentence or two clarifying this.”

Reply on comment 1:
We have added a sentence (line 121 in the clean revised version) to mention that the distributaries take off from main canals to serve fixed command areas irrigated year-round, with crops rotated between Kharif and Rabi crops.

Reviewer comment 2:
In your results, the validation could be strengthened. Are you able to compare your estimates against any in-situ records? Reported standard errors are useful, but which component dominates the uncertainty (precip, evaporation, storage, discharge, canal imports)? Standard errors are provided but there is no discussion of comparisons with independent ground-truth data or other datasets not used in calibration. Including such validation would enhance confidence in the results.

Reviewer comment 3:
Discussion would benefit from a short explanation on generalization. For example, can this approach work in snow dominated or urban catchments or is it basin specific?

Reply on comments 2 and 3:
Indeed, it will be useful to report the order of the standard errors of the different water balance variables, and it might be more interesting to highlight this at the start of the paper. Therefore, we edited the abstract to reflect the order. We also agree that it’s good to reflect on model generalizability and validation; therefore, we added a separate section 6.4. We have also looked into additional data, such as weather and crop variables, to further validate our results. We added figure C4, comparing the trend in our estimated evaporation with independent crop yield data, and we discussed these results in section 3.5 (lines 431-444 in the clean revised version).

Reviewer comment 4:
“Figures with more than one panel (starting with Figure 2) need tags (a, b, c, etc) and the caption should refer to each panel specifically for clarity (like you did for Figure 4). In Table 2, indicate the meaning of the underlined values. Table 3, Table 4, again indicate the bold and underline importance.”

Reply on comment 4:
Thank you for the suggestions regarding figures and tables. We have added the alphabetic labels for the subplots of Figures 2 and 3 and edited their captions. We have also edited the captions of tables 2, 3, and 4.

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 have improved the flow of equations by adding intermediate steps in the equations that more clearly show how one equation leads to the next (see details below). Also, we have updated the link to the data and software, as follows:

"Code and data availability:
All software used to collect the data is available at https://doi.org/10.5281/zenodo.11148992 . The source code developed for this research is available via https://zenodo.org/records/17348274."

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 added lines 176-181 (in the clean revised version of the paper) to reflect on the challenge of the scale difference between the water balance variables.

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 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. 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 added figure 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 added the information about the size of the evaporation and precipitation correlation matrices. These are now mentioned in the following subsections, which have been edited to show additional intermediate steps when going from one equation to the next. These subsections now read as follows:

3.1.1 Precipitation

For each month *t*, we typically have multiple gridded precipitation products, with unknown bias and random errors. We use the spread across different precipitation products to define a grid-scale precipitation error model for the prior means m_{P_t} (mean vector in Eq. 2) and the prior standard deviations s_{P_t} (vector of the square root of all diagonal entries of the autocovariance matrix V_{P_t} in Eq. 2):

$$m_{P_t} = (1 - w_p)P_t^{obs,min} + w_p P_t^{obs,max} \tag{3}$$

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

Equation (3) models the systematic bias in grid-scale precipitation for each month *t* by describing the grid-scale precipitation prior means m_{P_t} as a weighted average of the minimum ($P_t^{obs,min}$) and maximum ($P_t^{obs,max}$) precipitation for each grid cell. The weight or bias parameter w_P takes on an unknown value between 0 and 1, and is estimated from the data (see section 4.1).
Random errors in grid-scale precipitation are modeled using Eq. (4). This model expresses the grid-scale prior standard deviations s_{P_t} for each month t as a function of the difference between the maximum and minimum precipitation in each grid cell. A noise parameter (r_P), taking a value between 0 and 1, is used to scale the standard deviations. This parameter is also estimated from the data.
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:

$$V_{P_t} = S_{P_t} R_P S_{P_t} \tag{5}$$

where S_{P_t} 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):

$$C_{\mathcal{M}}(d|l_s, \nu) = \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{d}{l_s}\right)^{\nu} K_\nu \left(\frac{d}{l_s}\right) \tag{6}$$

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 $K_\nu(.)$ (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 ν. A value of ν approaching 0 indicates a rough spatial process, while the process is smoother when ν 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., ν , 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 ($\sim 1/2$ 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 $m_{\bar{P}_t}$ and variance $v_{\bar{P}_t}$ in month *t* follow from Eqs. 3, 4, and 5:

$$m_{\bar{P}_t} = \phi_P' m_{P_t} = \phi_P' \left[(1-w_p) P_t^{obs,min} + w_p P_t^{obs,max}\right] \tag{7}$$

$$v_{\bar{P}_t} = \phi_P' V_{P_t} \phi_P = \phi_P' S_{P_t} R_P S_{P_t} \phi_P = \phi_P' (r_P D_{P_t}) R_P (r_P D_{P_t}) \phi_P \tag{8}$$

$$\bar{P}_t \sim \mathcal{N}\ (m_{\bar{P}_t}, v_{\bar{P}_t}) \tag{9}$$

$$\bar{P}_t\ \geq 0 \tag{10}$$

where ϕ_P is the spatial averaging operator used to derive basin-scale mean and variance from grid-scale means and variances (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 $S_{P_t} = r_P D_{P_t}$, 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 field. 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 $\bar{P}_t$ for each month t as a random draw from a truncated normal distribution. The truncation at zero ensures physical consistency (nonnegative precipitation).

3.1.2 Evaporation

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

$$m_{E_t} = \mathfrak{f}_{\boldsymbol{E}}\ [(1-w_E) E_t^{obs,min} + w_E E_t^{obs,max}] \tag{11}$$

$$s_{E_t} = r_E \frac{1}{4} \left(E_t^{obs,max} - E_t^{obs,min}\right) \tag{12}$$

The systematic bias in grid-scale evaporation is modeled with two spatial and time-invariant calibration parameters, namely: w_E and $\mathfrak{f}_{\boldsymbol{E}}$. The parameter w_E is a bias parameter or weight that interpolates between the monthly grid-scale evaporation extrema $E_t^{obs,min}$ and $E_t^{obs,max}$. w_E takes on a value between 0 and 1, and is estimated from the data (see section 4.1). An additional scaling factor ($\mathfrak{f}_{\boldsymbol{E}}$) is incorporated to account for potential bias outside the observed grid-scale evaporation range, and is given a lognormal prior with mode at 1 (no bias) and a coefficient of variation CV of 50%.

On the other hand, Eq. (12) quantifies the random errors in grid-scale evaporation, as the difference between the maximum and minimum evaporation in each grid cell. A noise parameter (r_E), taking a value between 0 and 1, is used to scale the random errors. All parameters are solved as part of the water balance data fusion (Sect. 4).

The basin-scale 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:

$$m_{\bar{E}_t} = \phi_E' m_{E_t} = f_E\, \phi_E' \left[(1-w_E)E_t^{obs,min} + w_E E_t^{obs,max}\right] \quad (13)$$

$$v_{\bar{E}_t} = \phi_E' V_{E_t} \phi_E = \phi_E' S_{E_t} R_E S_{E_t} \phi_E = \phi_E' (r_E D_{E_t}) R_E (r_E D_{E_t}) \phi_E \quad (14)$$

$$\bar{E}_t \sim \mathcal{N}\,(m_{\bar{E}_t}, v_{\bar{E}_t}) \quad (15)$$

$$\bar{E}_t \geq 0 \quad (16)$$

where D_E is a diagonal matrix whose diagonal entries contain 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. (13-14) are precomputed from the evaporation data sets. 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 71059, 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 Gaussian Process kernel (Hensman et al., 2015) with fixed parameters l_s at 50 km and ν at 1.5.
Similar to precipitation, the basin-scale calibrated evaporation $\bar{E}_t$ for month t is treated as a random draw from a truncated normal distribution Eqs. (15-16). The truncation at zero ensures physical consistency (nonnegative evaporation).

3.2 River discharge and canal water import error models

We assume that data on river discharge (Q^{obs}) and canal water imports (C^{obs}) are both unbiased with proportional random errors (Eq. (18)). Both variables are modelled as truncated Gaussian variables (non-negative), with mean m_{x_t}, standard deviation s_{x_t} , and variance v_{x_t} in month t given by:

$$m_{x_t} = x_t^{obs} \quad (17)$$

$$s_{x_t} = a_x x_t^{obs} \quad (18)$$

$$v_{x_t} = s_{x_t}^2 \quad (19)$$

$$x_t \sim \mathcal{N}\,(m_{x_t}, v_{x_t}) \quad (20)$$

$$x_t \geq 0 \quad (21)$$

where x_t^{obs} represents the observed value for each month t (Q_t^{obs} for discharge and C_t^{obs} for canal water imports). s_{x_t}quantifies the random errors in Q^{obs} or C^{obs}, and is assumed proportional to the observed value, with a proportionality constant (or relative error) a_x. We assume a 10% relative error for monthly river discharge, i.e., $a_{\mathrm{Q}} = 0.1$, and a 25% relative error for canal water imports, i.e. $a_{\mathrm{C}} = 0.25$.

3.3 Water Storage Error Model

The water storage error model relates the storage observations from the GRACE satellite (S_t^{obs}) to the modeled storage (S_t). Due to the coarse resolution of GRACE data, the monthly basin-scale total water storage derived from these observations can be polluted by the water storage dynamics occurring outside the basin, that is, “leakage”. To account for the temporal and spatial mismatch between GRACE basin-scale water storage and the modeled storage caused by leakage errors, we employ the following noisy sine wave error model (Schoups and Nasseri, 2021):

$$m_{S_t} = S_t + A\,sin\left(\omega\left(\frac{t}{12} - \delta\right)\right) \quad (22)$$

$$v_{S_t} = \sigma_S^2 \quad (23)$$

$$S_t^{obs} \sim \mathcal{N}(m_{S_t}, v_{S_t}) \quad (24)$$

The first equation models the systematic errors associated with GRACE observations. It captures cyclic patterns and seasonality in the data within the basin via time-invariant error parameters. These parameters describe the magnitude and the timing of the seasonal error: the amplitude A (mm) and phase δ (years), respectively.
The magnitude of random errors as an unknown time-invariant parameter (σ_S), which reflects errors that can arise from (a) any limitations caused by the sine wave models in capturing the calibrated water storage dynamics within the basin and (b) noise in GRACE solutions. In the above equations, ω is fixed at 2π radians per year, yielding a one-year sine wave and thus capturing the annual seasonal cycle. Other parameters are assigned vague prior distributions where A follows a lognormal prior with mode at 30 mm and a CV of 200%, σ_S a lognormal prior with mode equal to 10 mm and a CV of 200%, and δ follows a flat logit-normal prior between 0 and 1 year with a location parameter $\mu = 0$ and scale parameter $\sigma = 1.4$.
While the above prior error model relies on a single storage data set, we assess a posteriori (Sect. 6.1) the sensitivity of the P and E posterior distributions to the use of different storage inputs, including Mascon (JPL) and Spherical Harmonic (CSR) GRACE solutions, which for the basin studied here, exhibit different long-term trends.

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 have updated the caption of Fig. 8 to mention this.

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 have defined CWC in the Figure 1 caption.

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 131 in the revised version of the manuscript), 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, and the newly added section 6.4, reflecting on this.

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.