Supplementary Information

1 Limitations of efficiency statistics

5 Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta Efficiency (KGE) have often been used in the context of inundation extent data (e.g. Yamazaki et al. (2011)), but these measures are predominantly designed for the analysis of stream gauge data where (i) uncertainty in both observed and simulated flows are at least an order of magnitude less than the overall mean of observed flow and simulated flow, respectively, and (ii) there is no possibility of spatial displacement between observed and simulated inundation. Why might (i) be a problem? If two sets of observational data are used with very similar values but one reported

10 inundation extent to higher decimal accuracy (e.g. fractions 0.20, 0.21, 0.19, ...) than the other (e.g. fractions 0.20, 0.20, 0.20, ...), then the much smaller σ_{obs} values in the less precise dataset will produce spuriously large and negative KGE values. This is an issue especially in low inundation areas where inundation extent is consistently close to zero, but these efficiency statistics do not return a perfect match score even if observations match predictions very closely: they instead usually return negative efficiency values, which are then interpreted erroneously as a very poor matches (NSE and KGE penalise simulated and

15 observed uncertainty unequally, as may simply be demonstrated using artificial data, Fig. S1). Why might (ii) be a problem? If a temporary inundation is predicted correctly in terms of magnitude, but spatially displaced by one gridcell, then very poor NSE and KGE values will result (q.v. the standard thresholds, Table 2).

In most analyses of observed *versus* simulated data streams, the variability of the simulated data can differ in magnitude from the variability of the observed data, and spatial displacement is a problem that increases at finer spatial scale.

20 . In this study, we are considering data on water cycle quantities that are highly variable, but our use of these efficiency statistics is justified by noting that the uncertainties at all our study sites are within the region where NSE may legitimately be calculated (Fig. S1).

The use of NSE or KGE as validation measures for hydrological quantities is well-established (Knoben et al., 2019), however these statistics must be used with caution with inundation data where observed estimates of inundation often have

25 zero variability. There do exist versions of these statistics that have been modified for use with low flow cases (e.g. Thirel et al. (2015)), however using these can introduce higher uncertainty for high flow cases. In order to avoid these issues, in our analysis we have simply excluded all gridcells where the flow regime did not fit the assumptions of the NSE and KGE statistics.
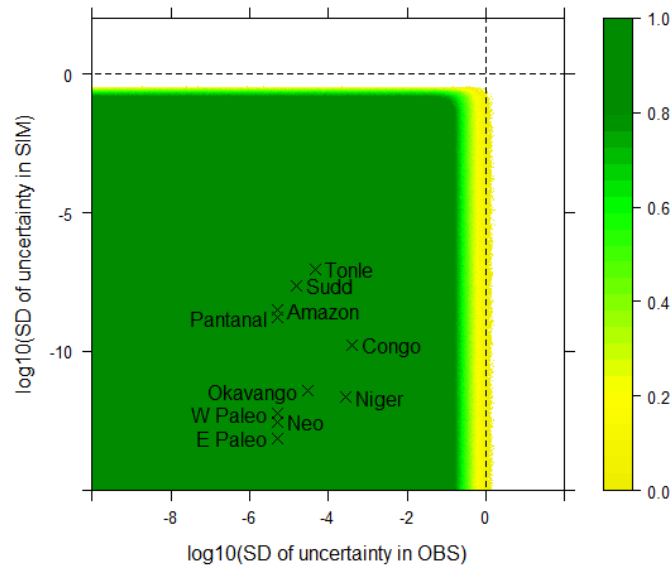
Figure S1: How Nash-Sutcliffe efficiency (NSE; Table 2) penalises uncertain data. This plot was generated using artificial data: (i) we assumed perfect predictive power in our model, (i.e. simulated mean always precisely equalled observed mean, so any divergence from a perfect efficiency score is attributable only to differences in variance), (ii) observational data was assumed to have Gaussian uncertainty with mean 0.0 and nonzero SD and (iii) simulated results were also assumed to have Gaussian uncertainty with mean 0.0 and nonzero SD. Note that despite perfect predictive power (for which we would in theory expect NSE values close to 1.0), the NSE values actually drop below zero as soon as observational uncertainty is equivalent to the size of $mean(Q_{obs})$ and/or simulation uncertainty is equivalent to 0.3* the size of $mean(Q_{obs})$. Points show the averaged uncertainties of all our study sites, showing that they all occur in the region where NSE may legitimately be calculated (i.e. observed and simulated uncertainties are <0.1* the mean of the observed or simulated data, respectively).

References

Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores, Hydrology and Earth System Sciences, 23, 4323-4331, https://doi.org/10.5194/hess-23-4323-2019, 2019.

Thirel, G., Andreassian, V., Perrin, C., Audouy, J. N., Berthet, L., Edwards, P., Folton, N., Furusho, C., Kuentz, A., Lerat, J., Lindstrom, G., Martin, E., Mathevet, T., Merz, R., Parajka, J., Ruelland, D., and Vaze, J.: Hydrology under change: an evaluation protocol to investigate how hydrological models deal with changing catchments, Hydrolog Sci J, 60, 1184-1199, 2015.

Yamazaki, D., Kanae, S., Kim, H., and Oki, T.: A physically based description of floodplain inundation dynamics in a global river routing model, Water Resources Research, 47, https://doi.org/10.1029/2010wr009726, 2011.