Contents of this file

1. Sections S1 to S4

2. Figures S1 to S5

Introduction This supplementary file contains a hypothetical example of how model simulated changes in runoff and evapotranspiration can be erroneous due to climatology biases, followed by an analysis of the sensitivity of our conclusions to the choice of precipitation dataset (Fig. S1), and a thorough discussion on the impact of all other factors besides aridity change (integrated by ω) on our conclusions (Figs. S2-5). Finally, we list the 34 CMIP5 models used in this study.

Section S1

To illustrate the potential disadvantage of using direct GCM output for water availability projections (or a GHM, if forced with biased input parameters), we present two examples: 1) a perfect model that simulates observed atmospheric water supply and demand in a water-limited catchment with zero bias, and 2) an imperfect model that is biased in its simulation of atmospheric water supply and demand (E_p/P is too small), so that an energy-limited catchment is wrongly simulated instead of the correct water-limited catchment. We rely on the assumption that ΔP and ΔE_p are correctly simulated by both models. In reality, the mean hydrological state of the surface will have some influence on the simulated future-minus-present changes in ΔP and ΔE_p, due to land-atmosphere feedbacks. For example, a trend towards more arid conditions can be amplified by feedbacks of soil moisture decrease on surface temperature, relative humidity and P [*Berg et al.*, 2016].

If a model perfectly simulates observed atmospheric water supply and demand (aridity) and E_p/P is equal to 2.0 (long-term mean P and E_p equal to 700 mm year^{-1} and 1400 mm year^{-1}, for example), then the Budyko formula [*Budyko*, 1974]:

$$\frac{E}{P} = \left\{ \frac{E_p}{P} \tanh\left(\frac{P}{E_p}\right) \left[1 - \exp\left(-\frac{E_p}{P}\right)\right] \right\}^{1/2}, \tag{S1}$$

returns long-term mean Q and E of 74 mm year^{-1} and 626 mm year^{-1}. E/P is large in this arid environment (0.9). We assume that the true ΔP and ΔE_p are -35 mm year^{-1} and +70 mm year^{-1} (future long-term mean P and E_p equal to 665 mm year^{-1} and 1470 mm year^{-1}; representing an increase in E_p/P from 2.0 to 2.2), so that future long-term mean Q and E are equal to 58 mm year^{-1} and 607 mm year^{-1}. Therefore Q has decreased by 16 mm year^{-1}, or 22 %.

If the same catchment is incorrectly simulated as having a present day E_p/P of 0.9 (long-term mean P and E_p equal to 1100 mm year^{-1} and 1000 mm year^{-1}, for example) then long-term mean Q and E are 375 mm year^{-1} and 725 mm year^{-1}. There is a five-fold overestimate in Q. We therefore expect Q to be overly-sensitive to changes in atmospheric water supply and demand [*Roderick and Farquhar*, 2011]. Taking the true ΔP and ΔE_p of -35 mm year^{-1} and +70 mm year^{-1} (future long-term mean P and E_p equal to 1065 mm year^{-1} and 1070 mm year^{-1}; representing an increase in E_p/P from 0.9 to 1.0) returns future long-term mean Q and E of 324 mm year^{-1} and 741 mm year^{-1}. The decrease in Q in this example is 51 mm year^{-1}. Adopting a simple bias correction [e.g., *Hempel et al.*, 2013] and assuming that the difference between simulated and observed Q in the future is the same as at present (+301 mm year^{-1}) gives a bias corrected Q in this example of 23 mm year^{-1} (324 mm year^{-1} minus 301 mm year^{-1}). This is a decrease from the true observed Q at present (74 mm year^{-1}) of 51 mm year^{-1}, or 69 %.

Section S2

Precipitation time series derived from widely used datasets are open to inhomogeneities due to discontinuities in station/spatial coverage. Here, we use the interpolated version of CRU TS3.23 because it offers complete global terrestrial coverage and so allows direct comparison to our catchment masks. We can also consider a raw version of CRU TS3.23 that only considers grid boxes where actual observations from at least one station are available. It is worth investigating the changes in spatial coverage over the historical period studied here (1951–2000) and the sensitivity of our conclusions to the version of CRU TS3.23 used.

We do not simply consider all months and grid boxes for which there is station data in the raw version of CRU TS3.23. There must be at least one station and it must be possible to construct a climatology (each calendar month must have at least 10 years' worth of data in the period 1961–1990). Further, for a water year (October-September) to be "complete", all 12 months worth of data are required. Figure S1 shows, for grid boxes across East Asia, the number of years in each decade of the 1951–2000 period that meet the above conditions.

The bottom right panel shows how the spatial coverage of the CRU TS3.23 raw precipitation dataset changes. For both of the Yangtze and Yellow catchments coverage increases dramatically during the 1950s, stays relatively constant from 1960 to 1990, before decreasing during the 1990s. Although we were restricted to the 1951–2000 period (highlighted on the figure panel by the gray shading) by the availability of river discharge measurements, this is also when the spatial coverage of precipitation observations is greatest. Indeed, during this period there is little difference between the time series for the CRU TS3.23 interpolated precipitation dataset and the CRU TS3.23 raw precipitation dataset. The linear trends in Yellow river P between 1951 and 2000 are 0.17 ± 0.21 mm day^{-1} century^{-1} and 0.16 ± 0.22 mm day^{-1} century^{-1} for the CRU TS3.23 interpolated precipitation dataset and the CRU TS3.23 raw precipitation dataset, respectively. Values for the Yangtze river P are 0.02 ± 0.34 mm day^{-1} century^{-1} and -0.02 ± 0.41 mm day^{-1} century^{-1}, respectively. There are some obvious differences between the time series in the early 20th century when spatial coverage of precipitation observations is poor. However, with analyses restricted to 1951–2000 our conclusions are unaffected by the choice of precipitation dataset.

Section S3

Throughout our work we take ω to be constant, first in using the Budyko framework to quantify the aridity change contribution to the observed change in Q, then in using the Budyko framework to constrain projections of water availability. But what about changes in climatic factors besides aridity, such as seasonality, snow dynamics and storminess, and non-climatic factors besides direct human impacts, such as land surface characteristics and the physiological response of plants to increasing CO_2 (CO_2 fertilization, CO_2 stomatal closure and water-use efficiency), which are integrated by ω? It is known that the sensitivity of Q to these other factors increases in arid climates [*Gudmundsson et al.*, 2016] and so the Yellow catchment is more likely to be affected. The LPJ LSM represents changes in land-use and land cover, the response to stomatal conductance due to rising CO_2 concentrations, CO_2 fertilization and soil moisture controls on transpiration [*Sitch et al.*, 2013]. Notably, it includes a realistic representation of vegetation, which is known to be a useful integrated indicator of these other factors that are integrated by ω [*Li et al.*, 2013]. Q simulated by the LPJ LSM strongly agrees with the aridity change contribution to the observed change in Q calculated using the Budyko framework.

The primary aim of our work is to produce physically consistent projections of Q by using CMIP5 simulated P and E_p with the models in the correct region of the Budyko space. By taking ω to be constant there is the possibility that we ignore robustly simulated ecohydrological changes that shape equally robust changes in ω and, in turn, Q. We can actually test this by verifying the Budyko framework on the CMIP5 models themselves. We use model simulated P and E_p, without correction, combined with model-specific constant ω, again calculated for the period 1951–2000. All of the subsequent analyses are for RCP8.5. Figure S2 shows the results from this verification. If the partitioning of P into Q and E is dependent on aridity according to the Budyko framework and constant ω is a valid approximation, then we would expect the multi-model mean Budyko verified ΔQ (0.16 ± 0.38 mm day^{-1} and 0.15 ± 0.17 mm day^{-1} for the Yangtze and Yellow, respectively) and the multi-model mean model simulated ΔQ (0.14 ± 0.40 mm day^{-1} and 0.09 ± 0.14 mm day^{-1} for the Yangtze and Yellow, respectively) to match. The 5-95 % range remains largely unchanged, but the multi-model means are noticeably different, especially for the Yellow catchment.

The discrepancy is almost certainly due to time-varying ω in CMIP5 models. This is tested by calculating ω for moving 10-year windows between 1951 and 2100 for all 34 CMIP5 models, using the objective function described in the main text. As already described, a number of factors determine the value of ω, but if it is indeed correlated with vegetation coverage, then as a first guess we would expect this parameter to increase in magnitude, since 21st-century increases in total vegetation coverage are projected [*Schneck et al.*, 2015]. Figure S3 shows the evolution of ω anomalies in the 34 CMIP5 models. The Yellow river shows a near-constant increase in the multi-model mean from about 1990 until the end of the 21st century. The Yangtze river shows an increase in the first half of the 21st century, followed by a smaller magnitude decrease in the second half. Interestingly, no obvious trends are apparent during the historical period (end of the 20th century) lending support to all other factors besides aridity change and direct human impacts having a negligible contribution to Yellow river Q change.

Following the Budyko framework, a 21st-century increase in ω means that for the same aridity the evaporative index, E/P, increases. Therefore, for the same aridity, more P leaves the catchment as E rather than Q. This explains why the Budyko verified ΔQ with constant ω is greater than the model simulated ΔQ, especially for the Yellow catchment. Unsurprisingly, considering model-specific time-varying ω, the multi-model mean Budyko verified ΔQ (0.14 ± 0.40 mm day^{-1} and 0.09 ± 0.14 mm day^{-1} for the Yangtze and Yellow, respectively) and the multi-model mean model simulated ΔQ (0.14 ± 0.40 mm day^{-1} and 0.09 ± 0.14 mm day^{-1} for the Yangtze and Yellow, respectively) match (Fig. S4).

We next test the sensitivity of our results (see Fig. 10 of the main text) to the choice of ω (constant or time-varying). We incorporate the CMIP5 changes in ω with an additive correction, adding temporally constant offsets (the absolute differences between observed and simulated climatological ω) in the same way that we correct P and E_p in the main text. Budyko corrected runoff with time-varying ω is shown in Fig. S5. The multi-model mean Budyko corrected ΔQ^* with time-varying ω (0.16 ± 0.42 mm day^{-1} and 0.07 ± 0.08 mm day^{-1} for the Yangtze and Yellow, respectively) and the multi-model mean Budyko corrected ΔQ^* with constant ω (0.18 ± 0.39 mm day^{-1} and 0.09 ± 0.09 mm day^{-1} for the Yangtze and Yellow, respectively) are closely matched, with small decreases in the multi-model mean ΔQ^* in both catchments. This does not change the conclusions of the main text. Recall that the aim was to generate refined runoff projections that account for the CMIP5 models, on average, being in the incorrect region of the Budyko space (particularly for the Yellow catchment).

Section S4

We use data from 34 GCMs participating in CMIP5 [*Taylor et al.*, 2012]; ACCESS1-0, ACCESS1-3, bcc-csm1-1, bcc-csm1-1-m, BNU-ESM, CanESM2, CCSM4, CESM1-BGC, CESM1-CAM5, CMCC-CM, CNRM-CM5, CSIRO-Mk3-6-0, FGOALS-g2, GFDL-CM3, GFDL-ESM2G, GFDL-ESM2M, GISS-E2-H, GISS-E2-H-CC, GISS-E2-R, GISS-E2-R-CC, HadGEM2-CC, HadGEM2-ES, inmcm4, IPSL-CM5A-LR, IPSL-CM5A-MR, IPSL-CM5B-LR, MIROC5, MIROC-ESM, MIROC-ESM-CHEM, MPI-ESM-LR, MPI-ESM-MR, MRI-CGCM3, NorESM1-M and NorESM1-ME.

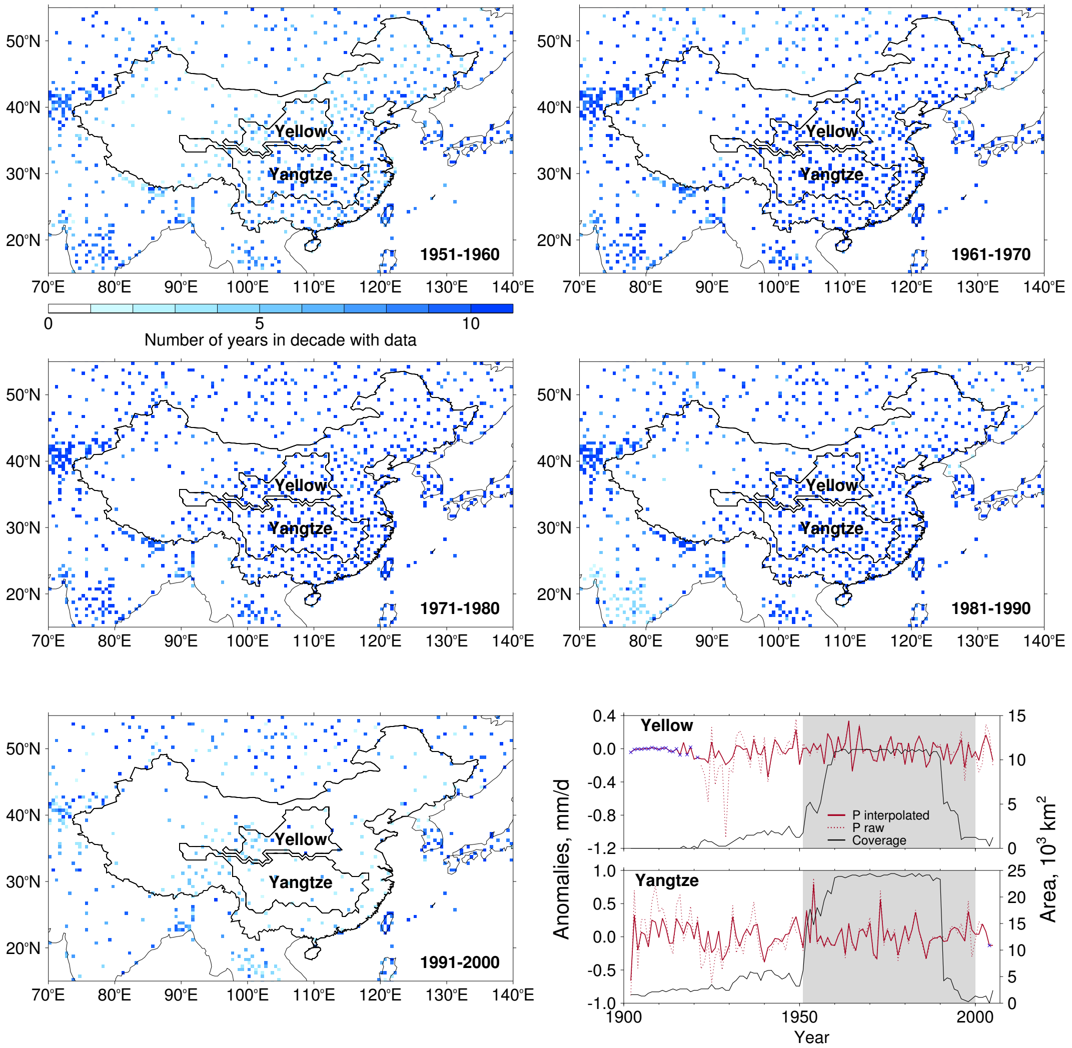

Figure S1: The spatial coverage of stations within the Yangtze and Yellow catchments and the wider East Asia region. The bottom right panel shows precipitation time series for both the CRU TS3.23 interpolated precipitation dataset (as used in the manuscript) and the CRU TS3.23 raw precipitation dataset (see Sect. S1 for information on how this is derived). Also shown, using the right y-axis, is the land area coverage of the CRU TS3.23 raw precipitation dataset. The blue crosses represent years where observations within the given catchment are missing (i.e. land area coverage is equal to zero) and a comparison between products is not possible. The gray shaded area represents the period 1951–2000 considered in our work.

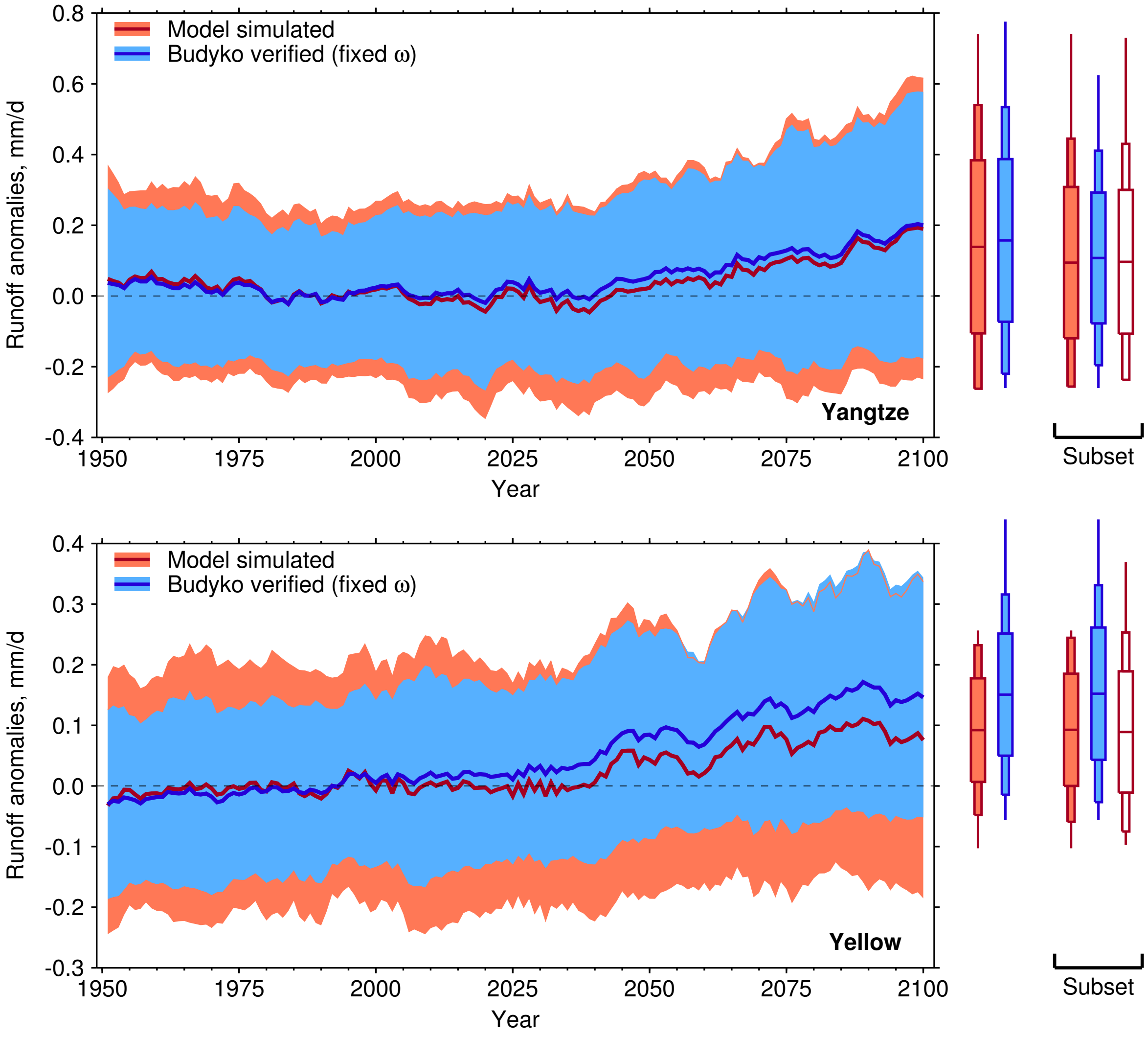

Figure S2: CMIP5 model simulated (orange) and CMIP5 Budyko verified with fixed ω (blue) runoff anomalies for 1951–2100, relative to 1980–1999, for the Yangtze (top) and Yellow (bottom) river catchments in the historical and RCP8.5 experiments. Shown are the 5-year running multi-model mean (thick line) and 5–95 % ranges (shading) across the CMIP5 ensemble. The box plots (mean, ± one standard deviation ranges, 5–95 % ranges, and minimum to maximum ranges) are given for 2080–2099. Also shown, for comparison, are box plots for a subset of 28 (from 34) CMIP5 models for which Q is directly simulated (not limited to being calculated as $P - E$). The unfilled box plot shows Q as directly simulated in these models.

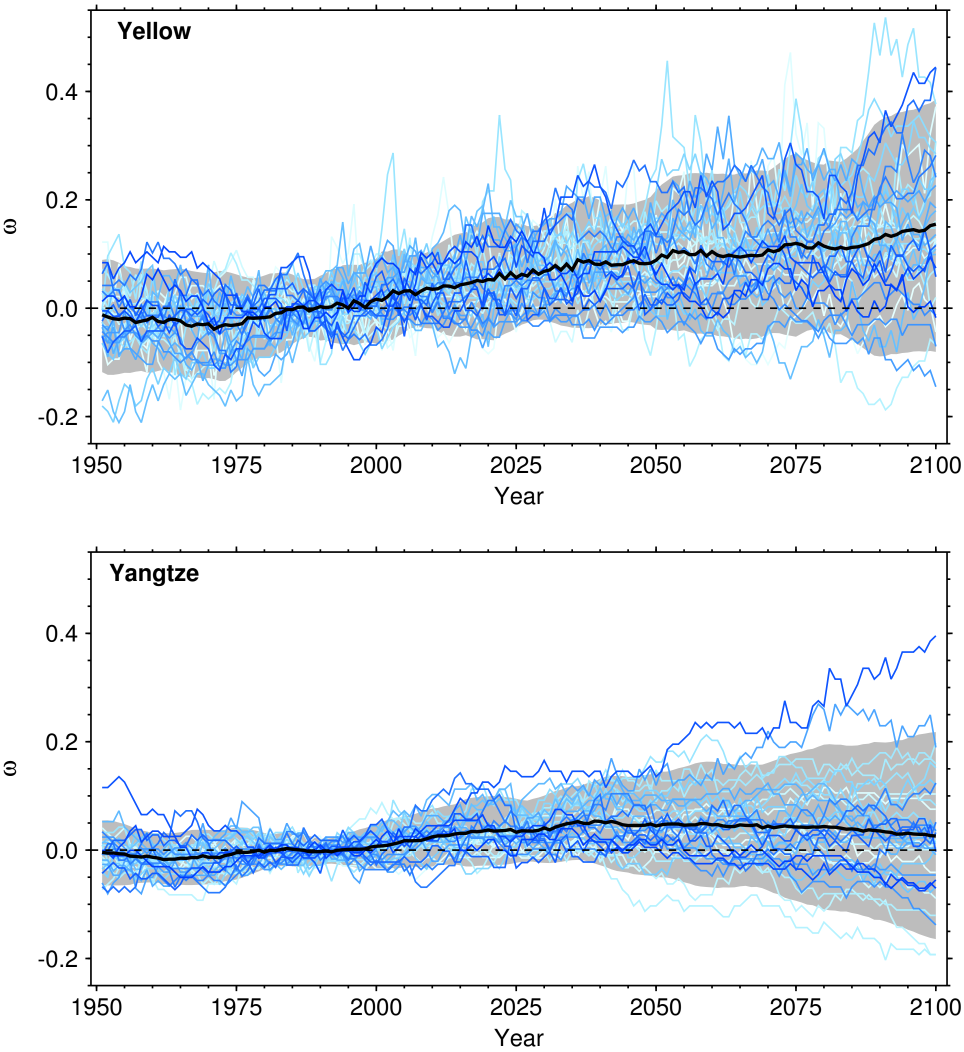

Figure S3: ω anomalies in the 34 CMIP5 models (blue lines). Anomalies are for 1951–2100, relative to 1980–1999, for the Yellow (top) and Yangtze (bottom) river catchments in the historical and RCP8.5 experiments. Shown are the multi-model mean (thick black line) and 5–95 % ranges (gray shading) across the CMIP5 ensemble.

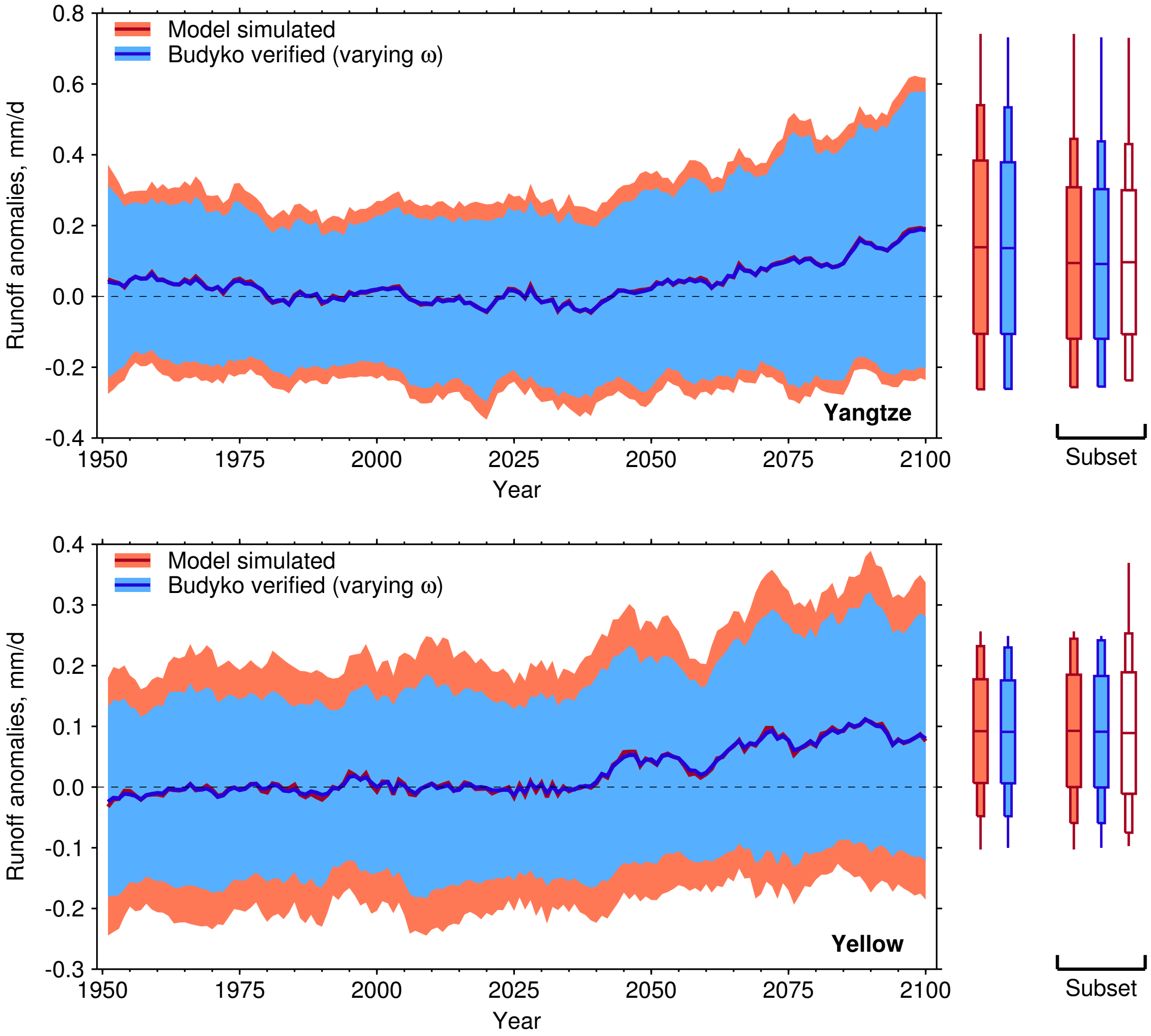

Figure S4: As in Fig. S2 but using time-varying ω for the Budyko verification, rather than fixed ω.

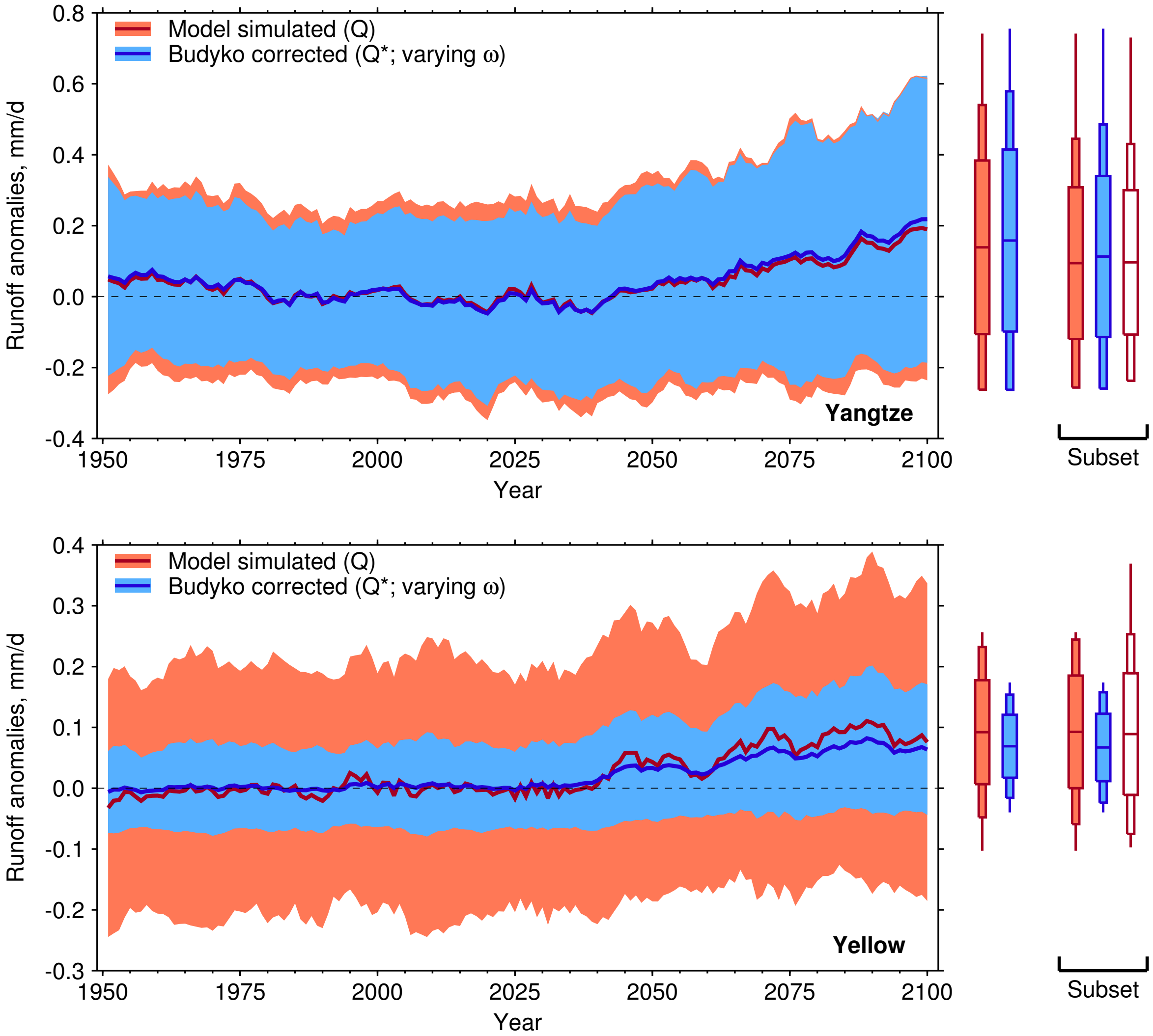

Figure S5: As in Fig. 10 of the main text but using time-varying ω for the Budyko correction, rather than fixed ω.