S1 Performance Ratings Analysis

Here we analyse the results of the performance ratings defined by Kalin et al. (2010). Figure S1 shows the full distribution of classifications from both SWOT-DR and GRADES. We can see that while GRADES performs better under the $|pBias|$ metric, the number of reaches labelled Unsatisfactory is small for both methods, indicating that the factors having the largest impact on the results are the dynamics, scales and temporal variation, reflected by the NSE metric, rather than biases in the results.

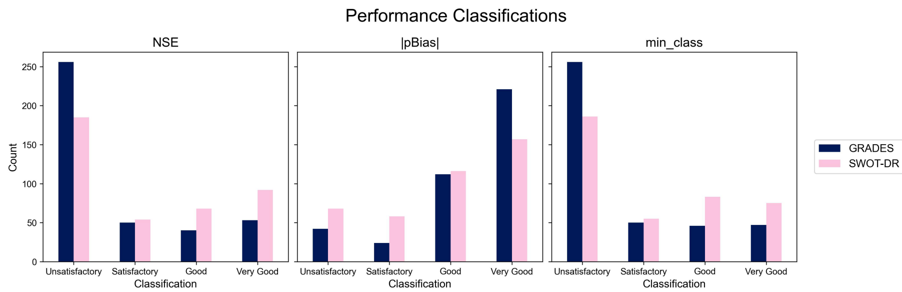

Figure S1. Distribution of performance rating classifications of 296 validation reaches, for each metric.

Figure S2 shows the increase in category from GRADES to our model; for example, +1 means that our model sits in a group one level higher than the GRADES discharge for a single reach. The most common outcome is no change, though we see the same differences between the NSE and $|pBias|$ metrics as above. The overall min_class metric shows a definite positive bias; that is, over all reaches we make more positive category changes than negative.

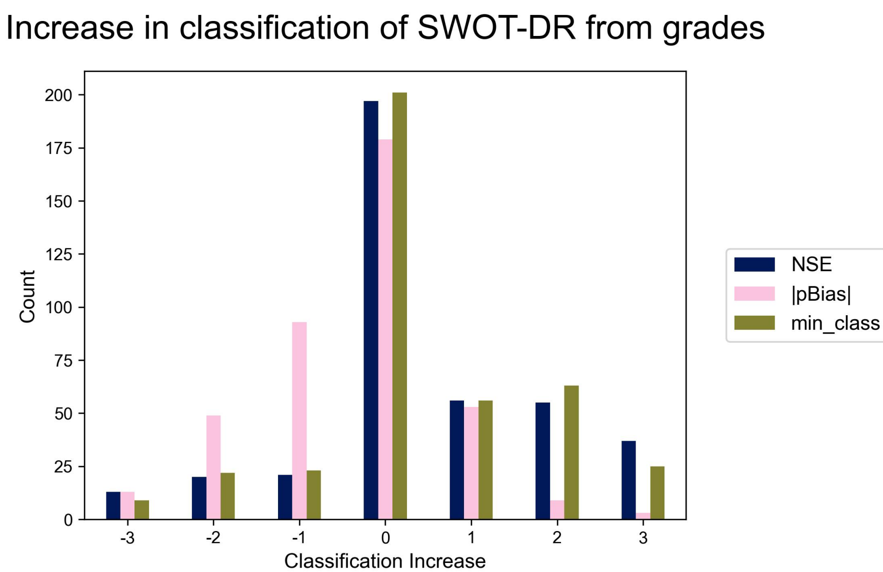

Figure S2. Distribution of category increases, for each metric. Positive means a reach is in a higher category in our model than in GRADES (i.e. performed better), negative means the opposite.

S2 Correlation of GRADES Time Series Analysis

Gleason et al. (In Review) show that there is a large variation in the correlation between GRADES discharge and water level observations, indicating a variance in dynamical skill of the GRADES model. We have suggested that this is an area in which our model makes an improvement, but we look to assess this in more detail. A lower correlation to observed water level indicates decreased reliability of GRADES data at these reaches, which could therefore have a knock-on negative impact on our model too. However, it also suggests that the integration of the highly-accurate SWOT WSE observations could make a more significant improvement in these locations.

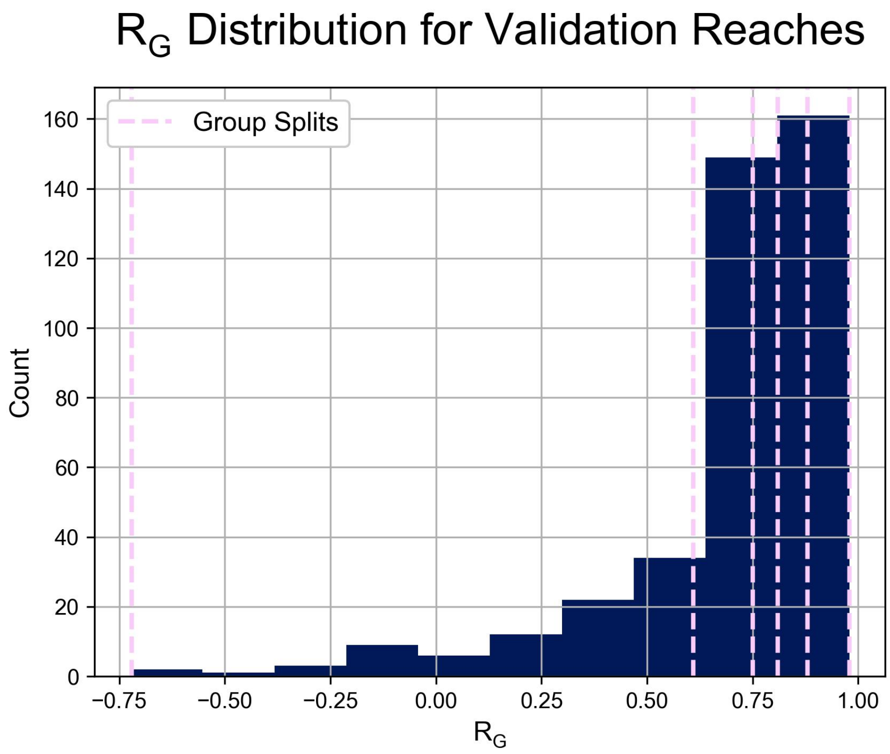

Figure S3. Distribution of R_G, the quintile group splits shown by the pink dashed lines.

For each of our validation reaches, we calculate the Spearman R value between SWOT WSE levels and GRADES discharge, as per Gleason et al. (In Review), which we label R_G. We split this into quintiles, with boundaries therefore at -0.72, 0.61, 0.75, 0.81, 0.88 and 0.98 as shown in Fig. S3, and summarise group performance.

Figure S4 shows that R_G is consistently correlated with skill in across all metrics, with more correlated groups showing better metrics in both GRADES and our model, as we would expect. Figure S5 shows how it effects the difference between the two models. A comparatively better performance is a curve that is shifted to the right (indicating a more-positive change in metric values) on metrics where a higher value is preferable (NSE, $nNSE$, KGE, R) and the opposite for metrics where lower values are preferable ($|pBias|$, $nRMSE$). Overall, we see the pattern that our model performs comparatively better at the central-to-lower R_G groups. In particular, our method significantly improves the R, $nRMSE$ and NSE values from $R_G \in (0.61, 0.75]$. Of course, we have not implied anything about the bias of the GRADES data over each group, so the minimal difference implies no strong correlation between R_G and the bias of GRADES. This allows us to improve the dynamics equally without a contrasting effect from any increased bias.

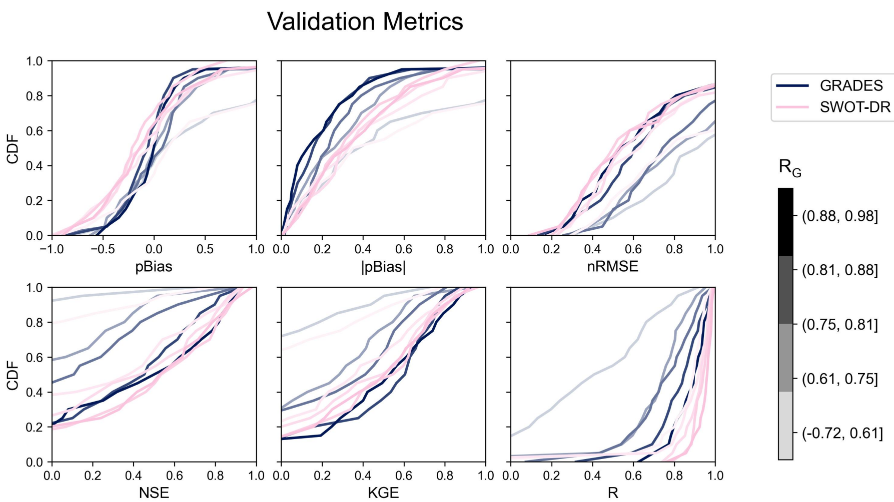

Figure S4. Cumulative distribution functions of $pBias$, $|pBias|$, $nRMSE$, NSE, KGE and R, calculated by comparing modelled to gauged discharge time series. Colour represents the model, shade shows different R_G.

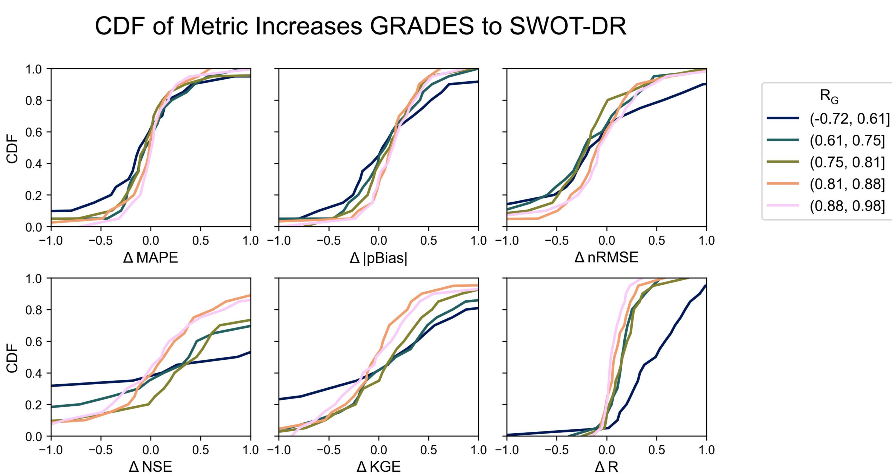

Figure S5. Cumulative distribution functions of the change in metric values between GRADES and our model, split by R_G group. Positive values means that our model has a higher value, and vice-versa.

This all indicates that there is a larger improvement, overall, to the reaches where GRADES struggles the most dynamically, which implies that this model can be a useful addition to global discharge datasets.

S3 Attribute-Metric Correlations

Table S1. Spearman R correlation values between SWORD reach attributes (rows) and a selection of metrics (columns).

Attribute	R	R_S	NSE	$nNSE$	$nRMSE$	$\|pBias\|$
lon	0.084	0.029	0.028	0.028	0.088	0.040
lat	-0.046	-0.072	0.030	0.030	-0.196	-0.151
n	-0.064	-0.021	0.048	0.048	-0.031	-0.087
n_h	0.006	-0.051	0.136	0.136	-0.114	-0.164
reach_len	-0.040	0.002	0.058	0.058	-0.081	-0.080
width	0.041	0.086	-0.123	-0.123	-0.071	-0.061
slope	-0.025	-0.053	0.020	0.020	-0.082	-0.076
facc	0.086	0.137	-0.249	-0.249	0.146	0.123

References

Gleason, C. J., Bates, P. D., Durand, M. T., Feng, D., Getirana, A., Lin, P., Oubanas, H., Pavelsky, T. M., Shen, C., Smith, L. C., Wang, J., Yamazaki, D., Yang, Y., An, H., Bennitt, F. B., Chuter, S. J., Fang, C., Flores, J. A., Friedmann, E., Jarugula, S., Ji, H., Langhorst, T., Long, D., Vu, D. T., Yin, Z., Allen, G. H., Andreadis, K. M., Biancamaria, S., Brown, C. M., David, C. H., Pan, M., Neal, J. C., Coss, S. P., Harlan, M. E., Simmons, T. T., and Tebaldi, N.: SWOT, empiricism, and river modelling, Geophysical Research Letters, In Review.

Kalin, L., Isik, S., Schoonover, J. E., and Lockaby, B. G.: Predicting Water Quality in Unmonitored Watersheds Using Artificial Neural Networks, Journal of Environmental Quality, 39, 1429–1440, https://doi.org/10.2134/jeq2009.0441, 2010.