Supplementary material

1 Basin information

Basin name	Center lon.	Center lat.	Glacierization degree (%)	GC99 [-]	GRDC station no.	GRDC station name	Obs. start	Obs. End
Alsek	-137	60	19.76	0.827	4102050	Near Yakutat	2000	2012
Amazon	-64	-6	0.03	0.003	3629001	Obidos - Linigrafo	2000	2007
Clutha	169	-45	0.31	0.114	5868050	Clyde	2000	2008
Columbia	-116	46	0.28	0.259	4215210	International Boundary (Canada)	2000	2012
Copper	-143	61	20.01	0.698	4102710	Million Dollar Bridge Near Cordova Ak.	2003	2011
Danube	18	46	0.05	0.04	6742201	Bazias	2000	2007
Dramselv	9	61	0.19	0.018	6731310	Dovikfoss	2000	2012
Fraser	-122	52	1.04	0.262	4207900	Hope	2000	2012
Gloma	11	61	0.63	0.141	6731403	Solbergfoss	2000	2012
Irrawaddy	96	23	0.02	0.006	2260400	Katha	2000	2009
Joekulsa	-16	65	15.03	0.898	6401702	Grimsstadir	2000	2012
Kalixaelven	22	67	0.22	0.001	6233850	Raektfors	2000	2012
Kuskokwim	-156	61	0.87	0.349	4102100	Crooked Creek, Alas.	2000	2012
Lule	18	67	0.98	0.211	6233750	Bodens Krv (+ Vattenverk, Trangfors)	2002	2011
Mackenzie	-120	61	0.09	0.029	4208025	Arctic Red River	2000	2012
Nass	-129	56	6.3	0.535	4206100	Above Shumal Creek	2000	2012
Negro	-68	-39	0.05	0.083	3275990	Primera Angostura	2000	2012
Nelson	-101	51	0.03	0.02	4213711	Long Spruce Generating Station	2000	2012
Ob	75	55	0.03	0.028	2912600	Salekhard	2000	2009

Table S1: Information on each basin and their corresponding observations. For the Rhone, data from the Hydrobanque was used instead of the GRDC as an exception. Only basins with more than 5 years of observation records between 2000 and 2012 were selected. The Colorado (Argentina) complied with these requirements but its GRDC observations were defected. See Appendix 5 for the five basins in this list assumed to be invalid for analysis.

2   Aletsch glacier runoff calibration

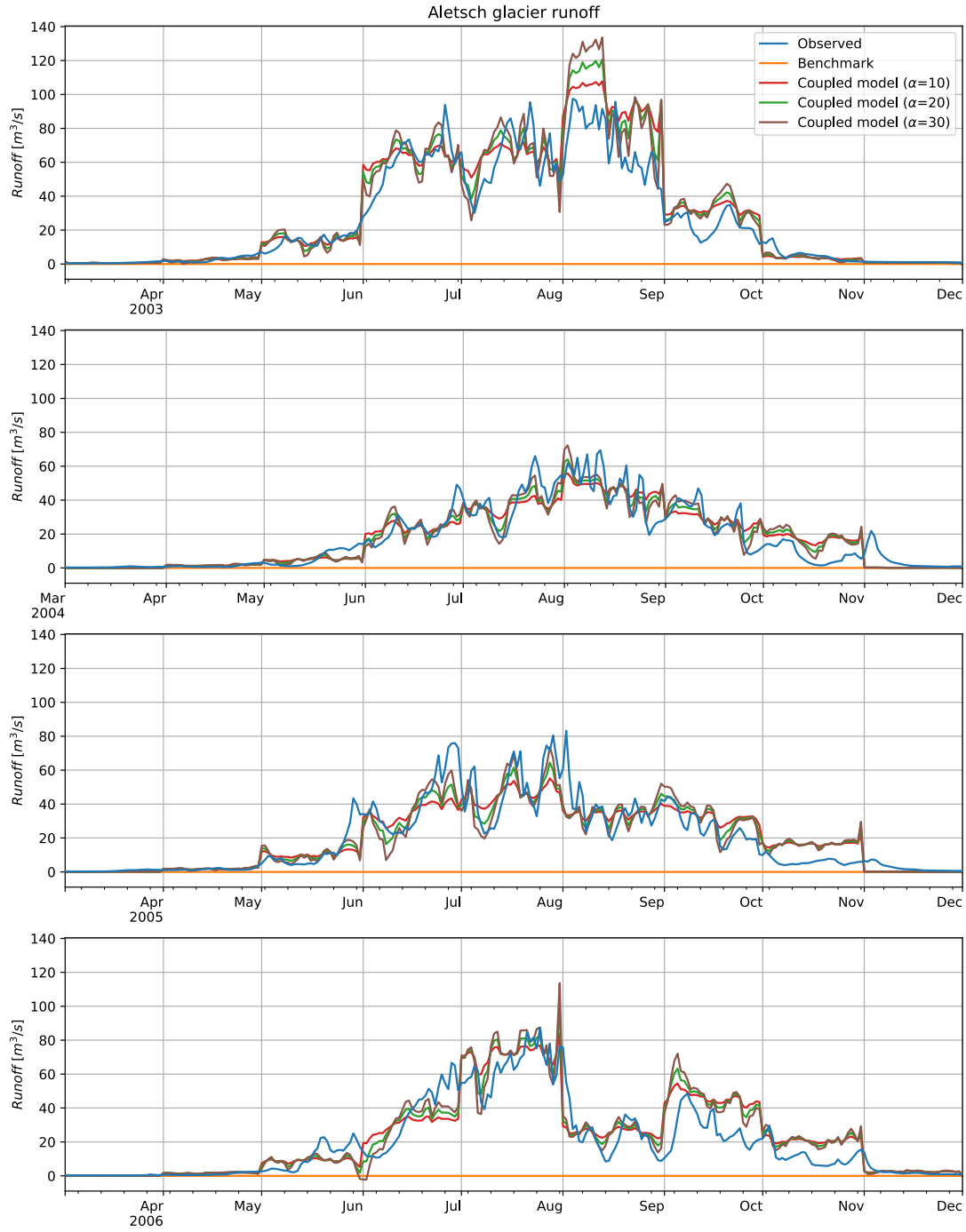



Figure S1: To determine the optimal weighting factor α to use in the weighting function (section 2.2 of the main text), a simple calibration was performed on the runoff downstream of the Aletsch glacier (BAFU). An α-value of zero gives a monthly step-wise function, while a higher α-value leads to a higher sensitivity to daily temperature. An α-value of 20 produces the lowest RMSE over the 10 years considered and was therefore selected, although the sensitivity analysis suggested an α-value of 30 to possibly be better on the basin scale. The benchmark simulates a runoff of zero, which can likely be attributed to the formation of snow towers.

3 Benchmark independent evaluation

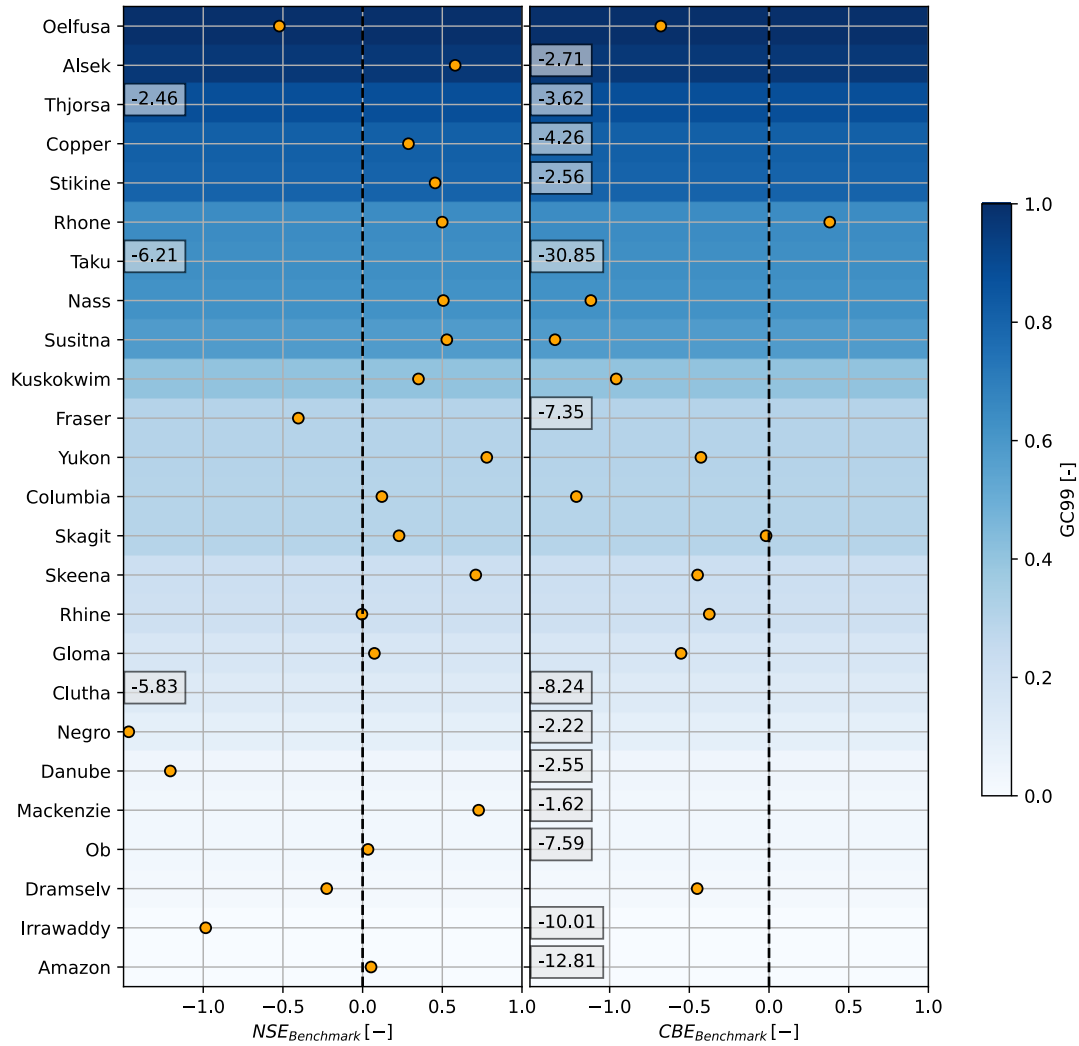



Figure S2: Evaluation of the benchmark (standard PCR-GLOBWB 2) against observations. The performance is expressed both in the Nash-Sutcliffe efficiency (NSE, Nash and Sutcliffe (1970)) and calendar day benchmark efficiency (CBE, Schaefli and Gupta (2007)). In seasonal runoff regimes the mean flow is a poor benchmark, resulting in high NSE-values. The CBE is more suitable for seasonal regimes and was therefore included in the evaluation. While in many basins the benchmark performs better than the mean flow, it only performs better than the mean flow of each calendar day in the Rhone.

4 Overall metrics

In this appendix three other metrics are applied on each of the 25 basins over the whole time range for an overall evaluation. Additionally, an explanation is provided on the choice of the relative RMSE difference over these overall metrics.

Overall metrics

Since no significant runoff timing difference is involved between the benchmark and the coupled model, we only consider metrics evaluating the value differences. Firstly, a benchmark efficiency (BE) is applied as follows:

$$BE = 1 - \frac{\sum_{t=1}^{N}(Q_{Obs} - Q_{Coupled})^2}{\sum_{t=1}^{N}(Q_{Obs} - Q_{Benchmark})^2}$$

where Q_{obs} is the observed basin runoff as reported in the GRDC and N is the number of data points. With this metric, a value of 1 indicates perfect correlation with the observations and a value of 0 or lower indicates equal or worse performance compared to the benchmark respectively. This benchmark efficiency is similar to the Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) and to the benchmark efficiency defined by Schaefli and Gupta (2007), but while those metrics need an artificial benchmark to compare the model against, the benchmark in this study is already present. Additionally, while its use would facilitate comparison with other studies given its widespread use in the hydrological modeling community, the NSE is a poor metric choice in highly seasonal flow regimes (Schaefli and Gupta, 2007, van Tiel et al., 2020).

Secondly and thirdly, the flow-duration benchmark efficiency (FDBE) and total flow benchmark efficiency (TFBE) are applied analogous to the above-defined BE:

$$FDBE = 1 - \frac{\sum_{t=1}^{N}(FD_{Obs} - FD_{Coupled})^2}{\sum_{t=1}^{N}(FD_{Obs} - FD_{Benchmark})^2}$$

$$TFBE = 1 - \frac{|\sum_{t=1}^{N}Q_{Obs} - \sum_{t=1}^{N}Q_{Coupled}|}{|\sum_{t=1}^{N}Q_{Obs} - \sum_{t=1}^{N}Q_{Benchmark}|}$$

in which FD represents the flow-duration curve. These metrics assess the ability of the model to reproduce the flow regime and the total basin runoff of the observations as compared to the benchmark.

Results

For all three BE's a low sensitivity can be observed in lowly glaciated basins, owed to the limited influence the glacial runoff has on the total runoff, flow-duration curve and total flow respectively. The opposite is true for highly glaciated basins (see Figure S3). For both the BE and the FDBE, around half of the basins score positively, with seemingly no correlation to glaciation degree. For TFBE only a small number of basins score positively, since PCR-GLOBWB 2 often already overestimates the basin runoff in many cases and the additional basin runoff in the coupled model only exacerbates this.

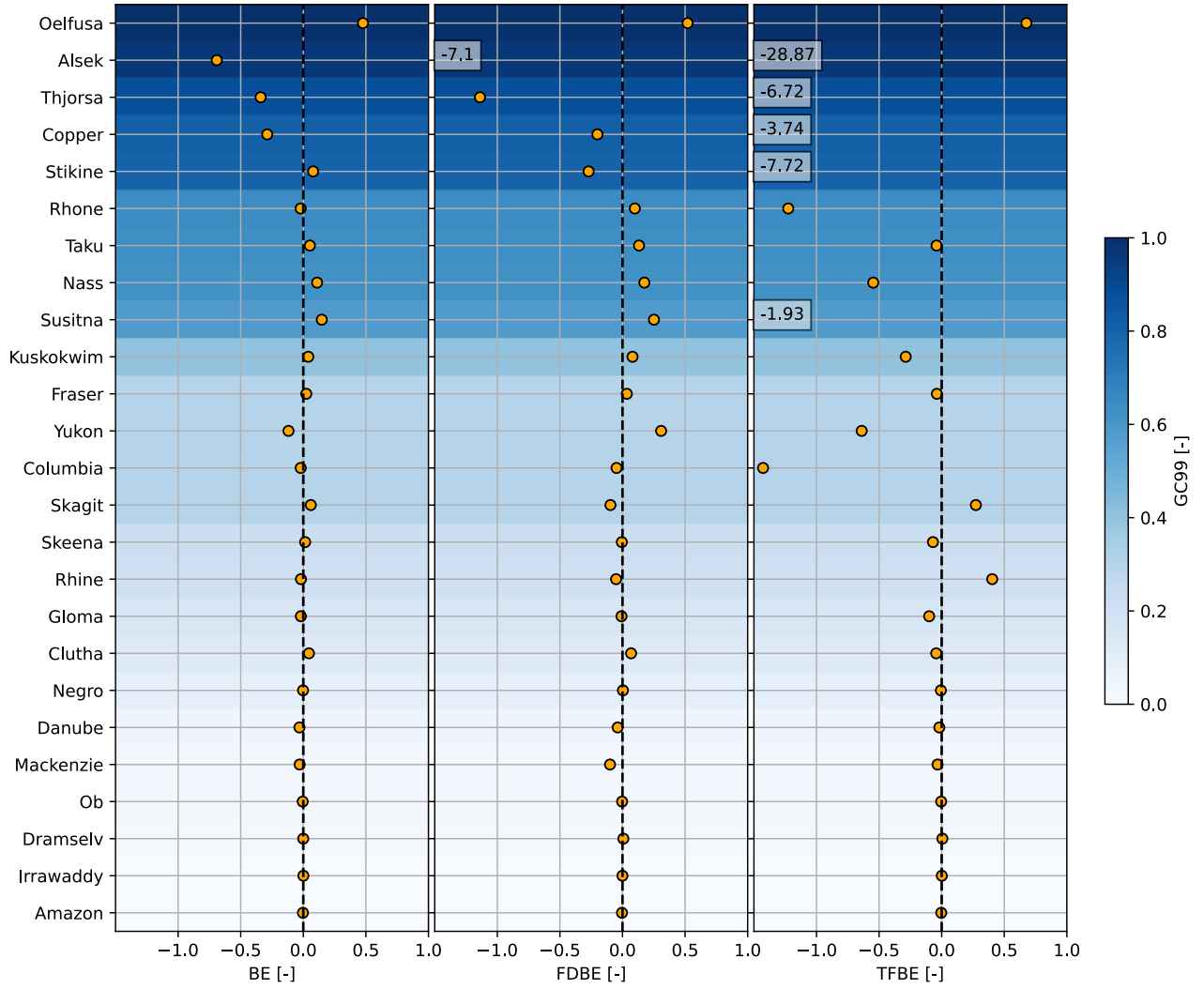



Figure S3: Results of the three overall metrics: the benchmark efficiency (BE), the flow-duration benchmark efficiency (FDBE) and the total flow benchmark efficiency (TFBE). The basins are sorted based on the 99th quantile of the contribution of the routed GloGEM glacier runoff to the coupled model runoff. A value of 1 indicates perfect correlation of the coupled model with the observations, while a negative value indicates a lesser performance compared to the benchmark.

Metric choice

While the above-mentioned metrics provide a good overall evaluation, their capacity of interpreting the data is limited for three reasons. Firstly, since the melt simulated by the models is highly seasonal, the difference between the models is also likely to be highly seasonal. It is therefore worth looking at the average monthly performance of the coupled model as compared to the benchmark, instead of only at the entire time range. Secondly, since the mean basin runoff and the fraction of glacial meltwater to basin runoff are different for each basin, an absolute error metric such as the BE does not allow for a fair comparison between the coupled model and the benchmark and between the different basins. In a lowly glaciated basins such as the Amazon, an additional error caused by the coupling of GloGEM will be only a fraction of the total error and will cause the BE-score to deviate only minimally from zero, and vice versa for highly glaciated basins such as the Oelfusa. Finally, the BE fails to express the performance change (+/-) between the coupled model and the benchmark relative to the maximum possible performance change (Seibert et al., 2018). In other words, the BE misses out on the fact that the same error decrement is worth more on a day with little melt than on a day with a high melt rate. The relative RMSE difference (RRD) introduced in section 3.4.2 meets these three criteria and was therefore chosen for this particular study. To avoid the introduction of an entirely new metric, the RMSE was used as a basis and the calculation was kept simple.

5 Discarded basins

Although in total 30 basins with runoff observations data were found, 5 of them proved to be unsuited for further analysis. Firstly, the flow of the Kalix was routed upstream towards a bifurcation of the Torne that in reality forks into the Kalix. This is likely caused by the reduced DEM quality above the polar circle (Lehner et al., 2008). Secondly, the Santa Cruz, the Lule and the Nelson contain lakes upstream of the GRDC station, making any meaningful evaluation of daily streamflow impossible. Finally, the Joekulsa is simulated to contain an endless reservoir just downstream of its glaciers which fills up during the summer and drains slowly over the course of the winter and spring. The cause of this misrepresentation likely has to do with the routing module settings of PCR-GLOBWB. It should be noted that in other basins, such as the Copper and Skeena basins, this problem might be partly present as well but to a much lesser degree and they are therefore not excluded from analysis.

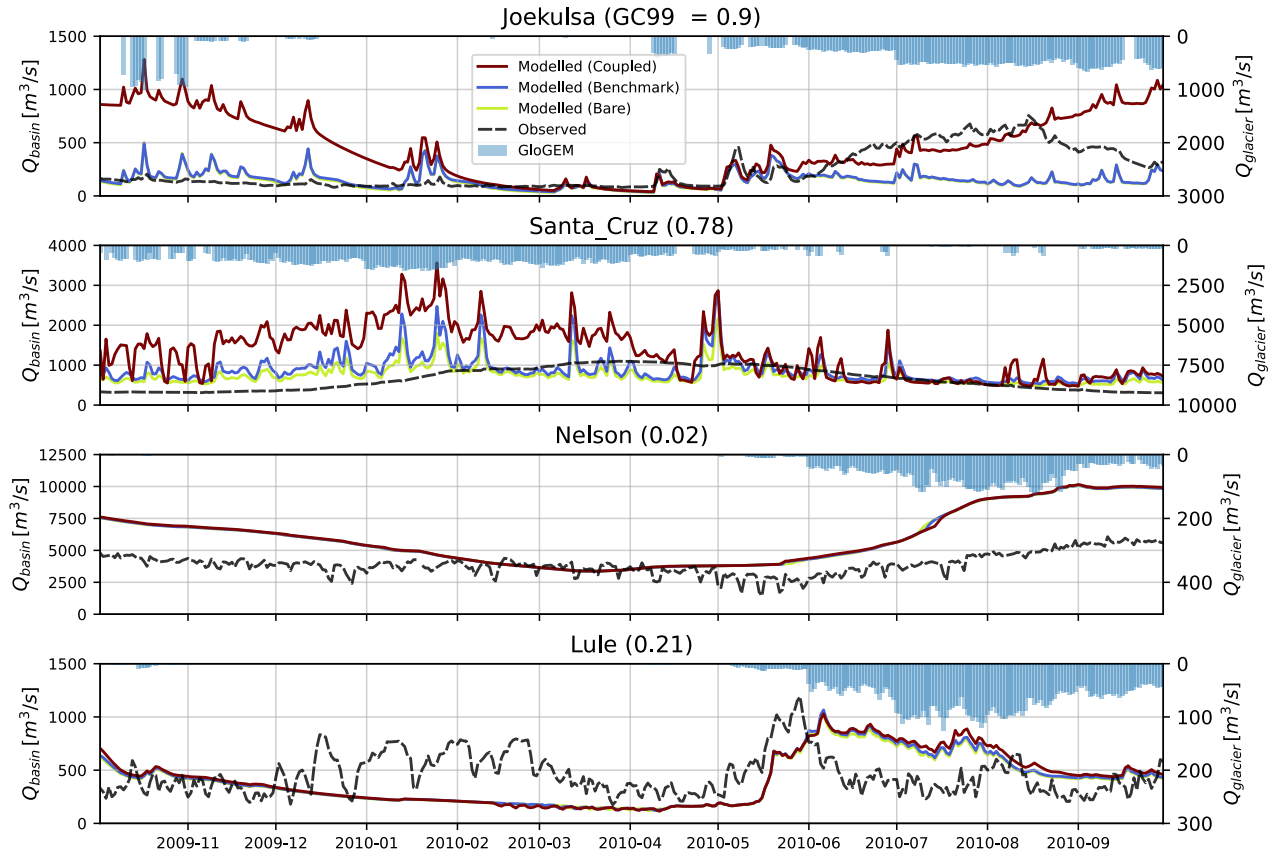



Figure S4: Four of the five discarded rivers: the Joekulsa due to routing problems, the remaining three due to large lakes in the river course. The Kalixa is not shown as the modelled discharge is simply zero, since all discharge is routed into a neighboring river.

6 Weighting factor sensitiviy

Given its importance in accurately capturing the daily fluctuations of glacier runoff, we quantified the sensitivity of the weighting factor in the resampling function of GloGEM from monthly to daily resolution (section 2.2). For the same four basins as in section 5.1 the coupled model runs were repeated with α-values of 10 and 30 instead of 20. Compared to an α-value of 20, an α-value of 30 lead to better RRD scores in 3 out of 4 basins (see S2). Additionally, it ensured a smoother glacier runoff transition between months where jumps were sometimes visible in the runs with an α-value of 20 (e.g. Columbia and Mackenzie in figure 2). Nonetheless, the maximum mean differences in RRD are -0.009 for α=10 and +0.0176 for α=30. The sensitivity of the weighting factor is therefore limited compared to other factors of the glacier parameterization, such as snow redistribution and mass balance calculations.

Table S2: Sensitivity analysis of the resampling weighting factor α. Difference in RRD between the α=10 and α=30 cases with the reference case (α=20).

α	Alsek	Columbia	Oelfusa	Rhone
10	-0.00905	-0.00144	-0.00587	0.00015
30	0.01364	0.00247	0.01759	-0.00232

7 PCR-GLOBWB 2 SWE evolution

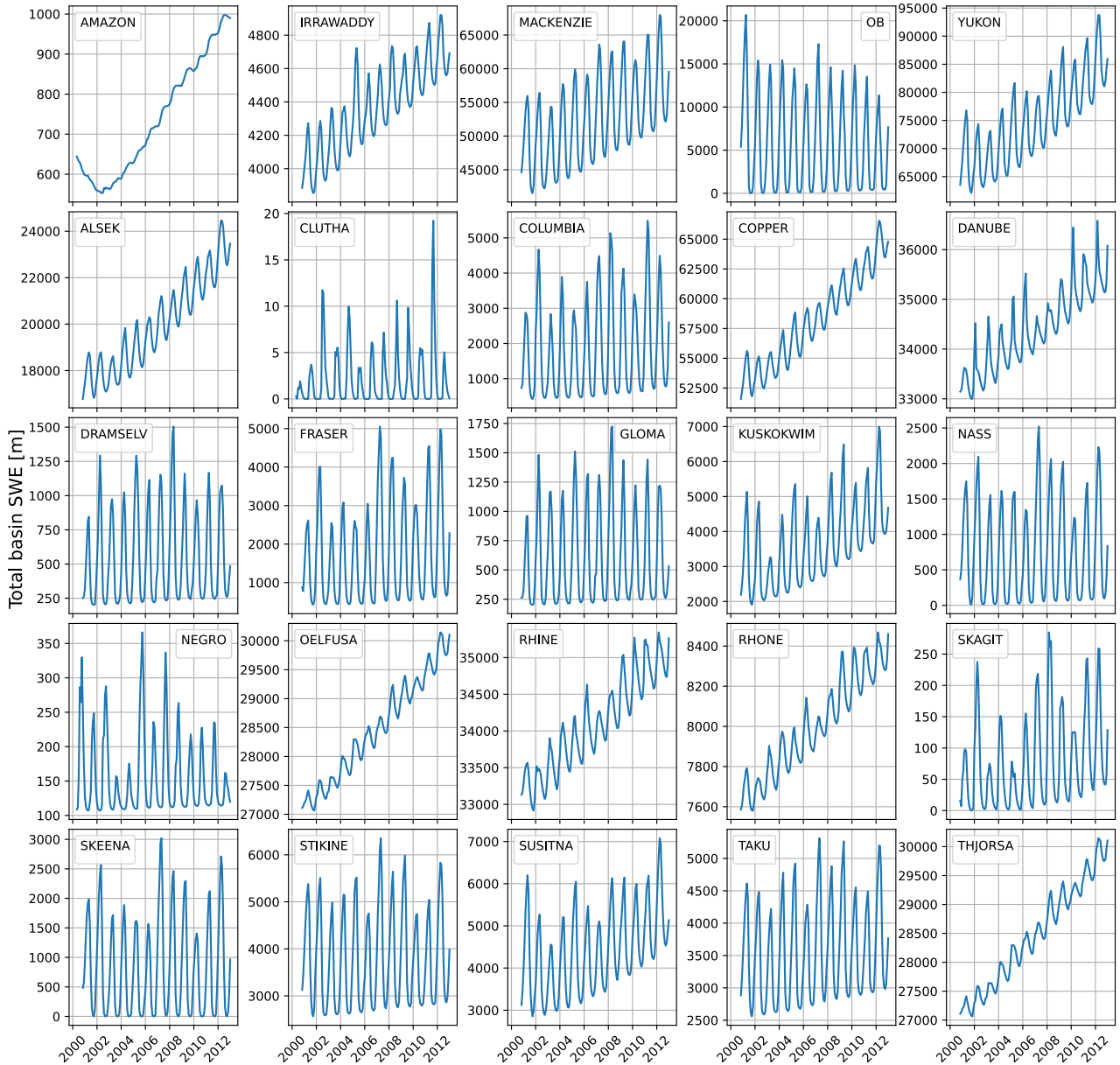



Figure S5: Basin-wide snow water equivalent (SWE) evolution. The multi-year accumulation of 'snow towers' due to a lacking snow redistribution representation in PCR-GLOBWB 2 is present in all basins with an increasing annual trend (16/25).

References

BAFU. Hydrologische Daten und Vorhersagen. URL `https://www.hydrodaten.admin.ch/de`.

B. Lehner, K. Verdin, and A. Jarvis. New global hydrography derived from spaceborne elevation data. *Eos*, 89 (10):93–94, 2008. ISSN 00963941. 10.1029/2008EO100001.

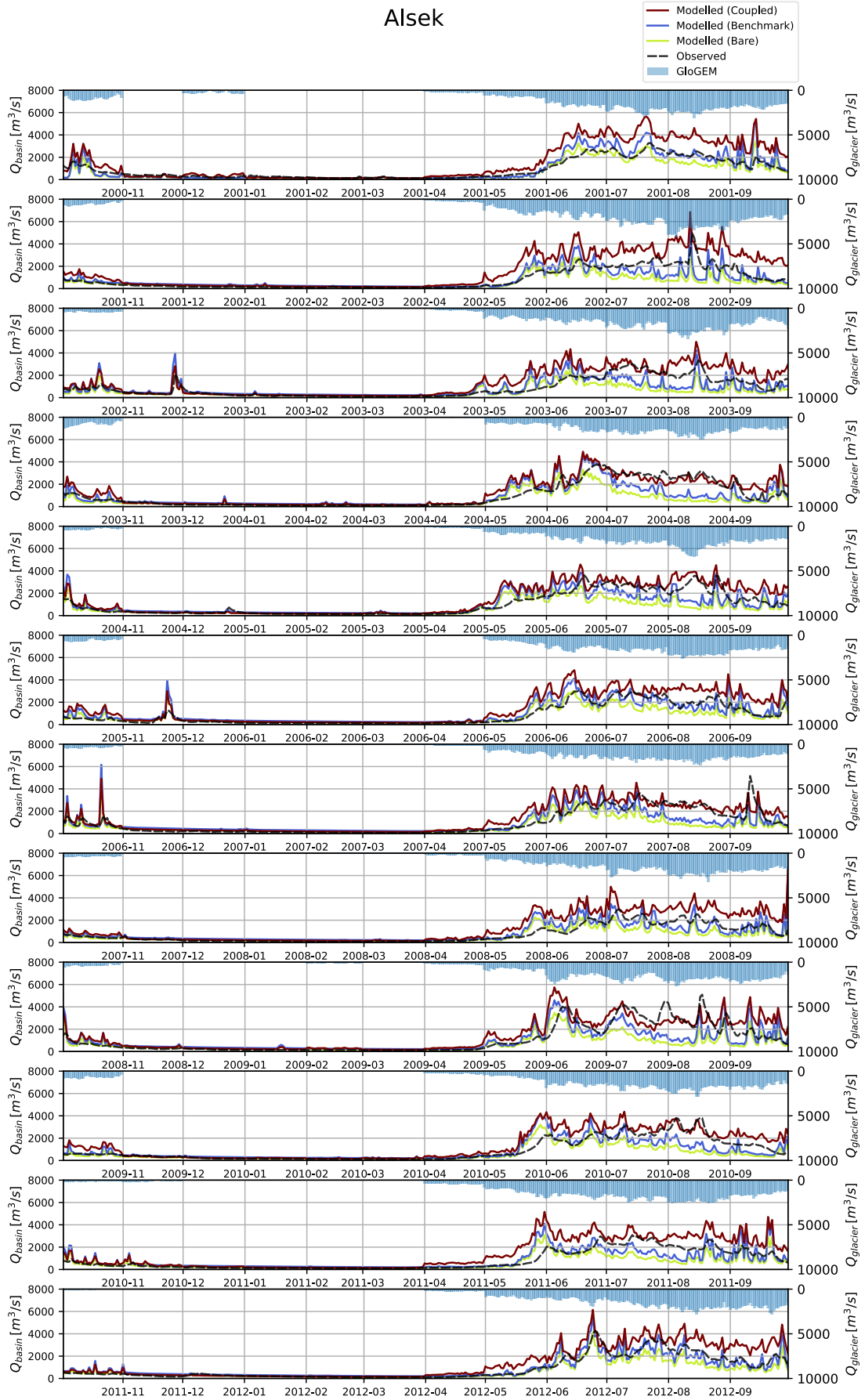
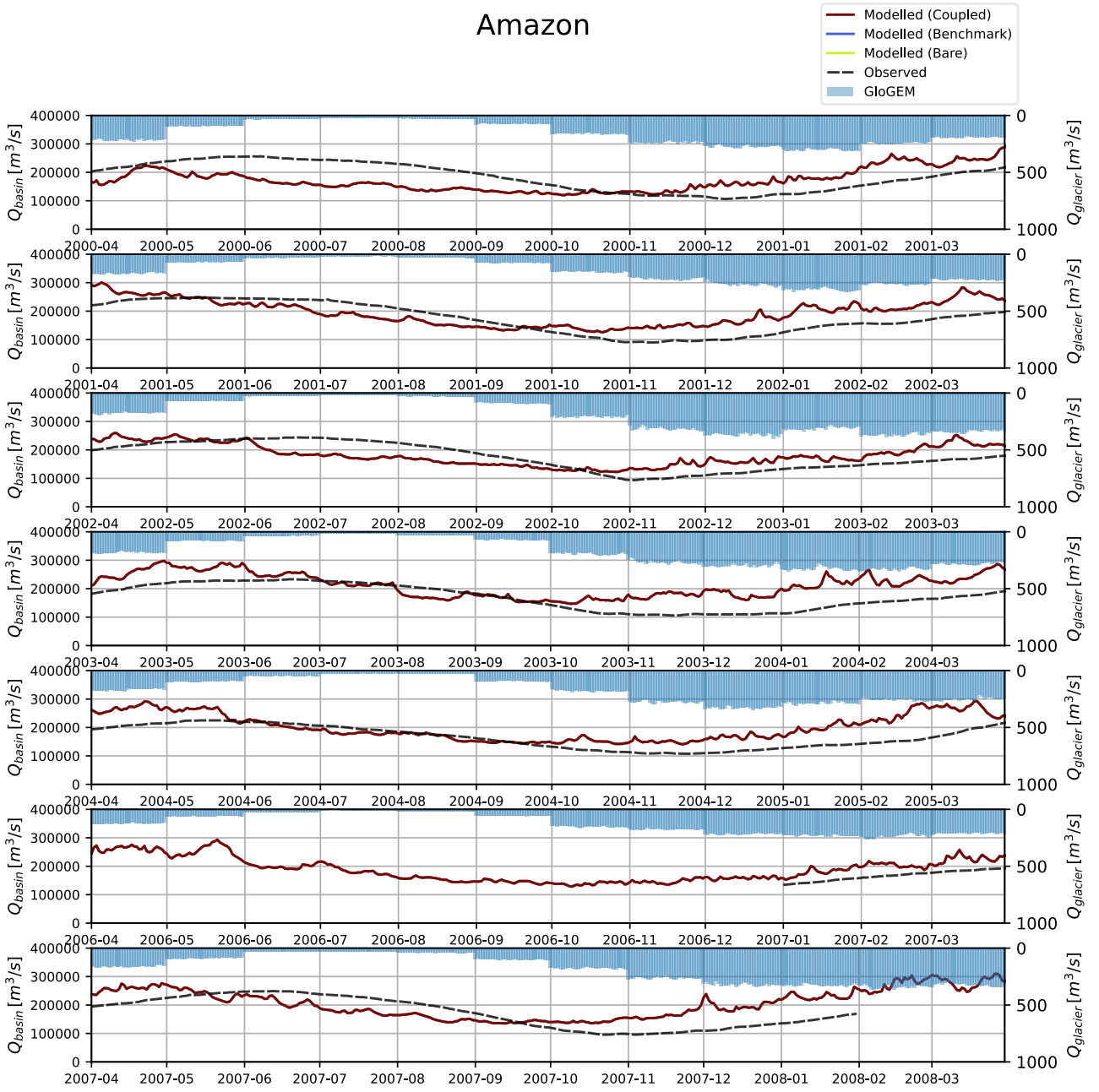
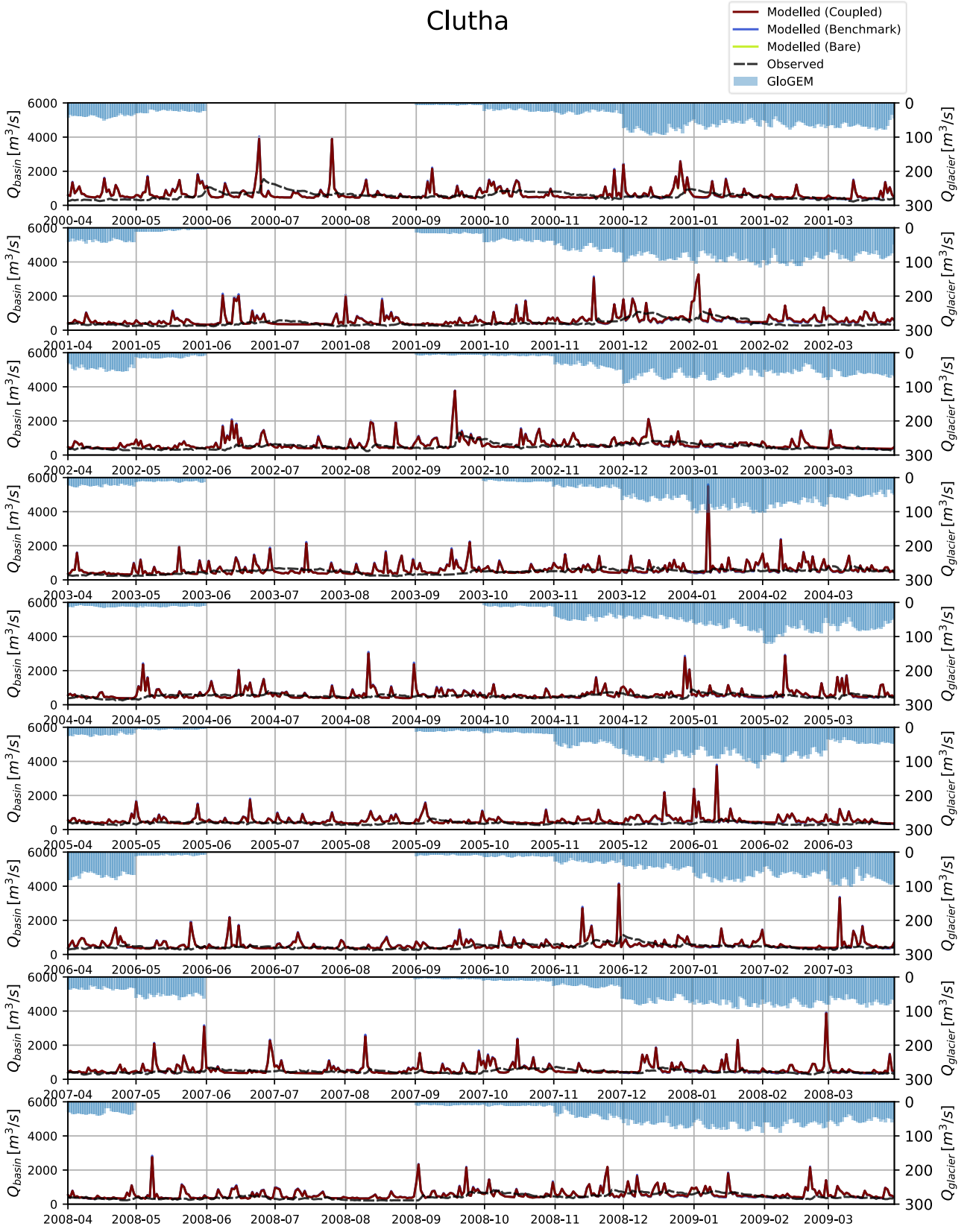
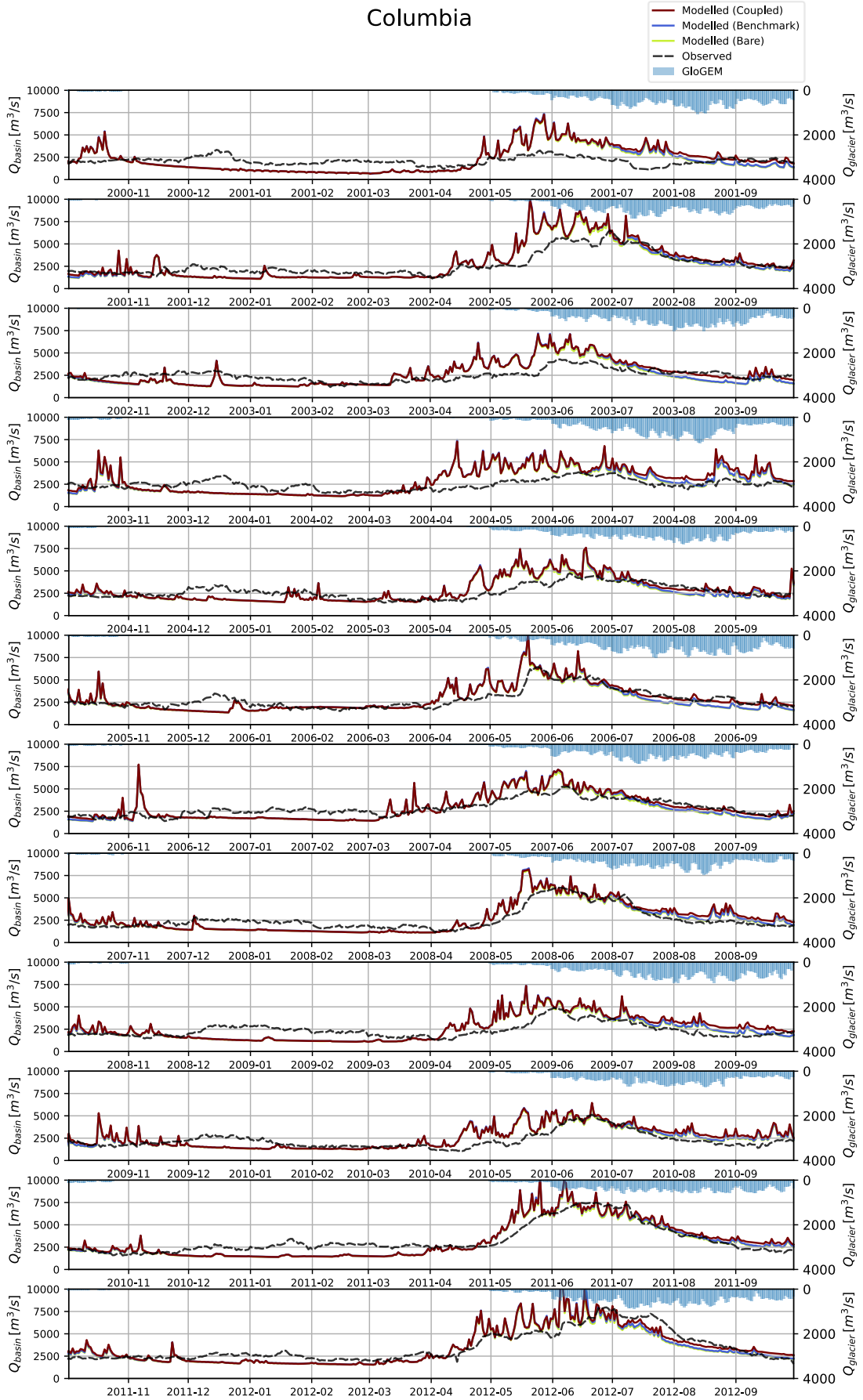
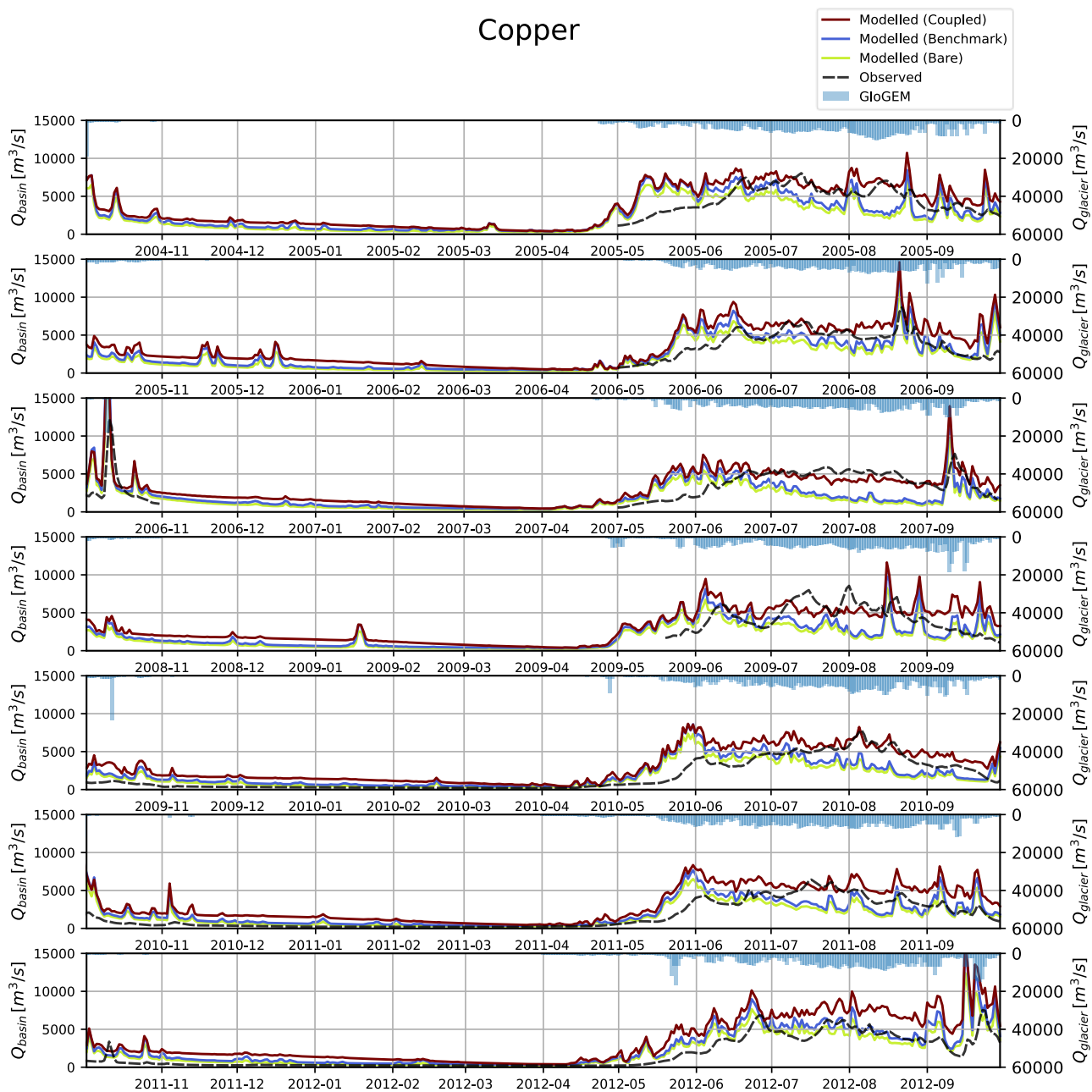
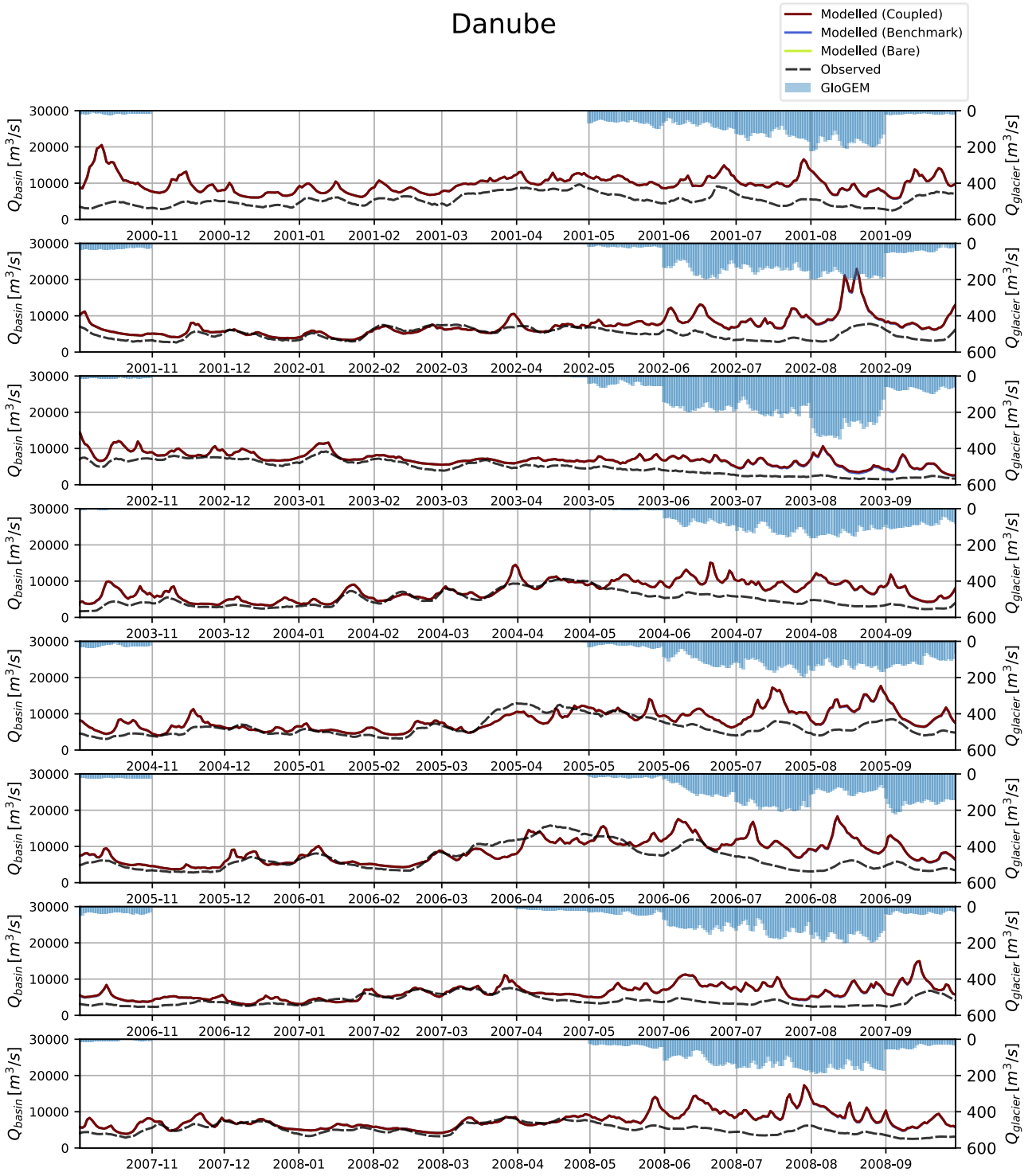
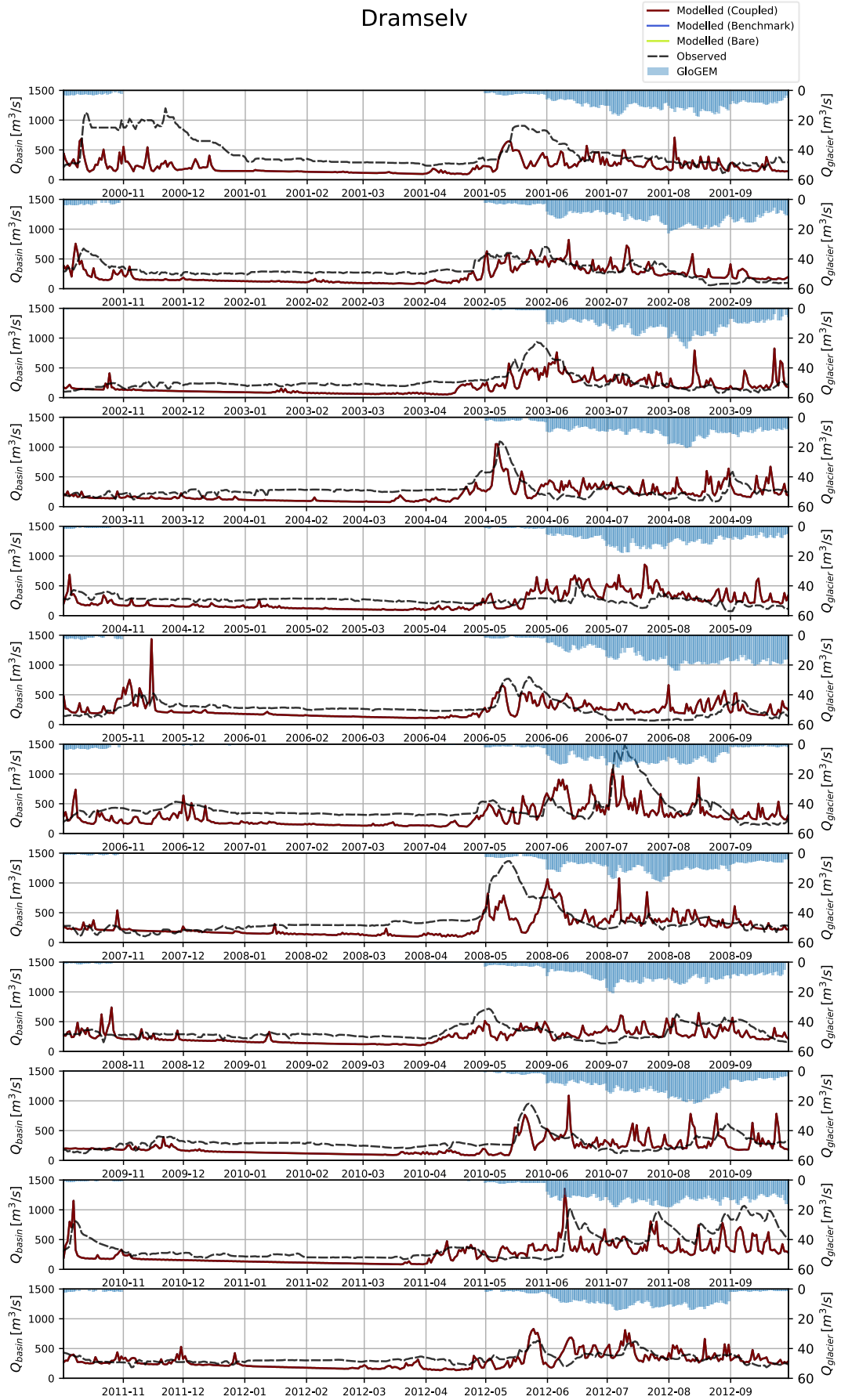
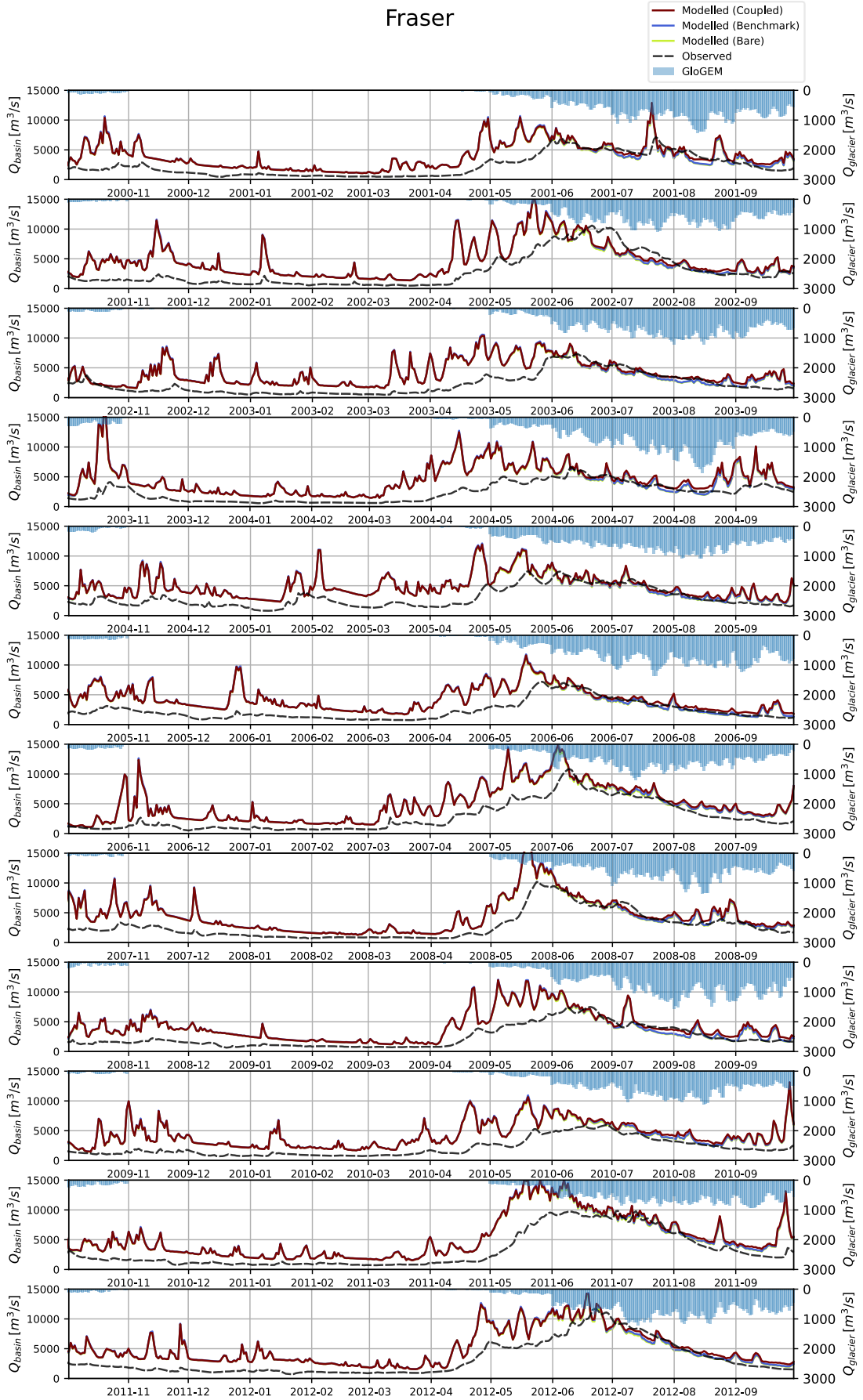
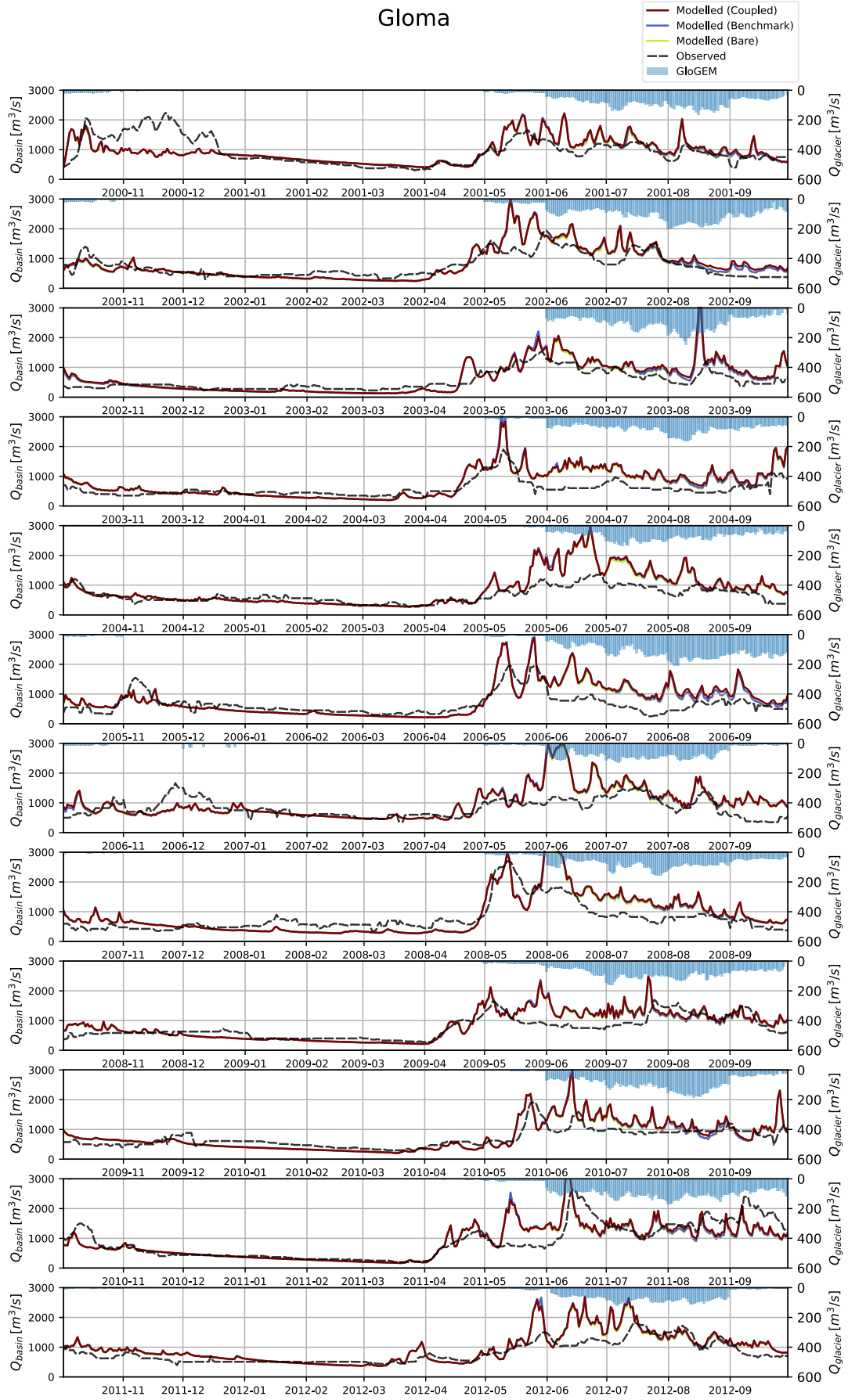
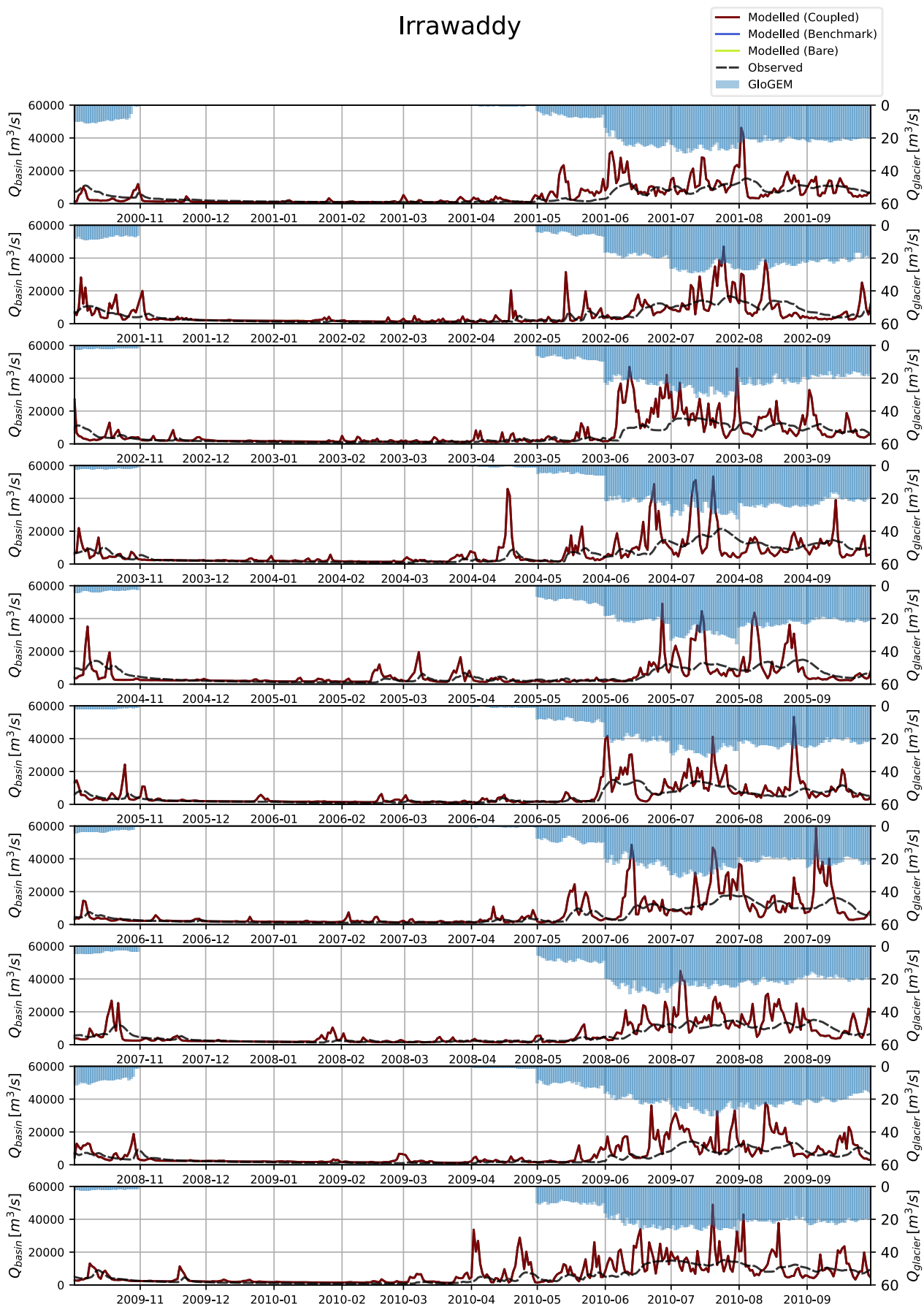
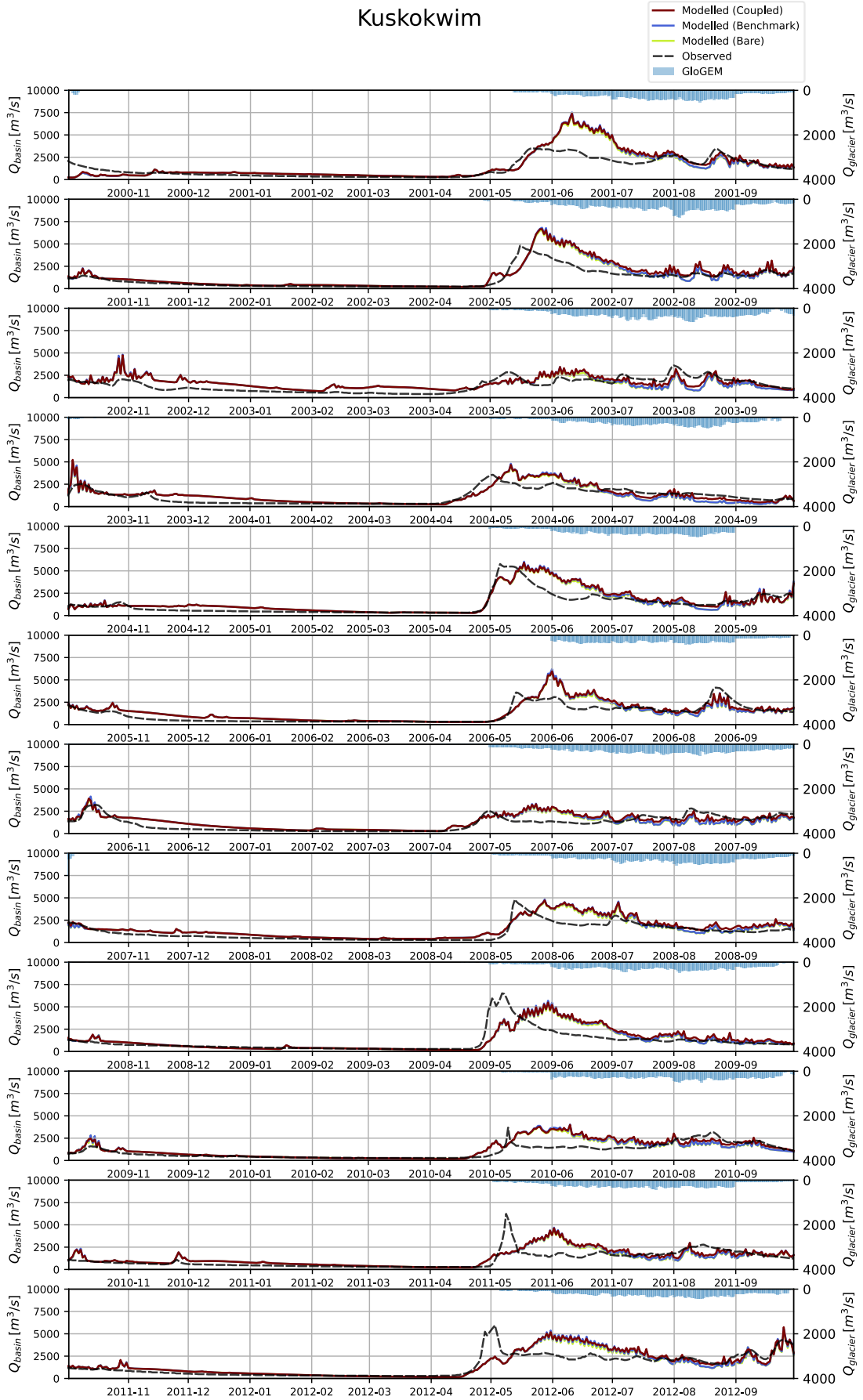
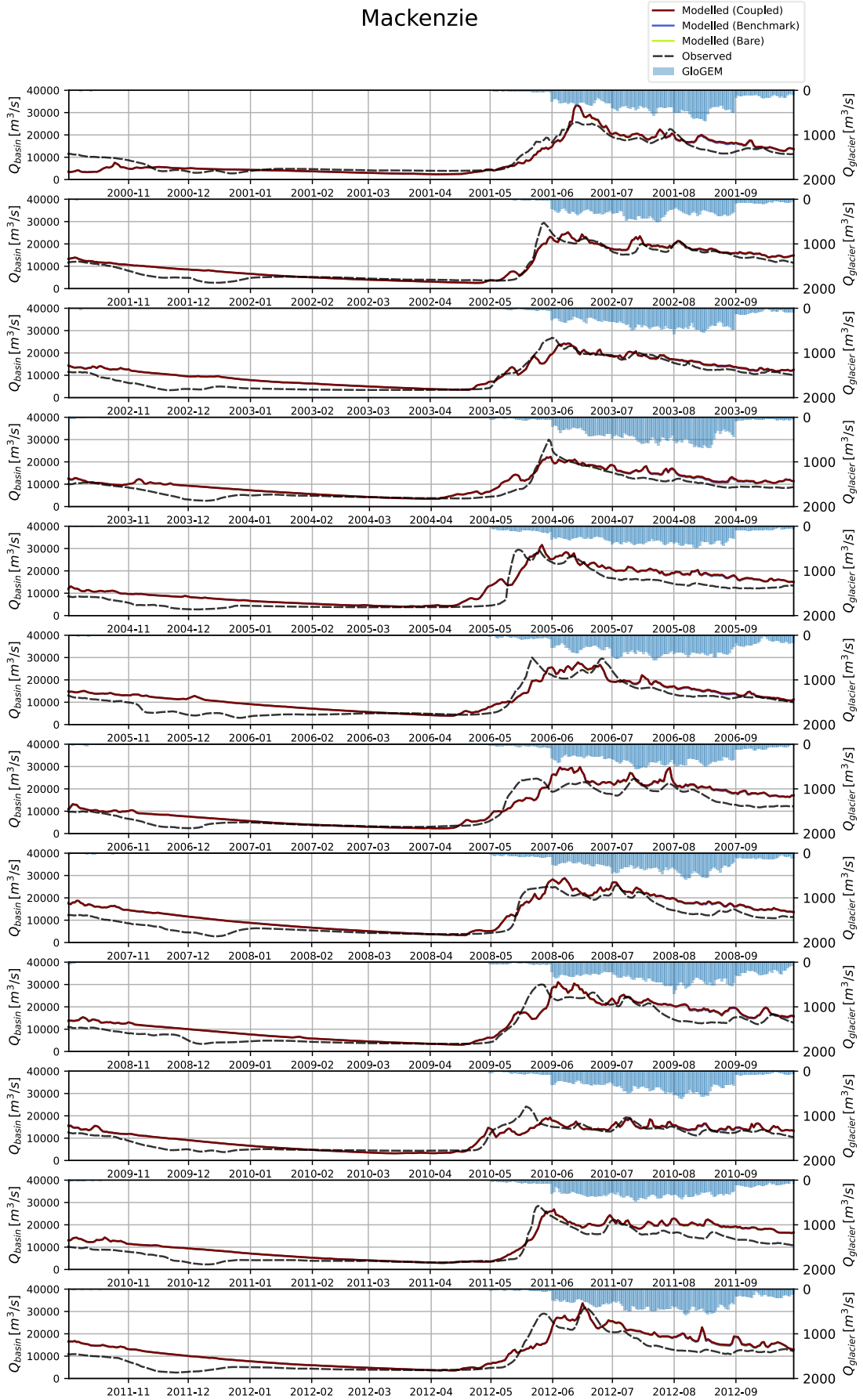
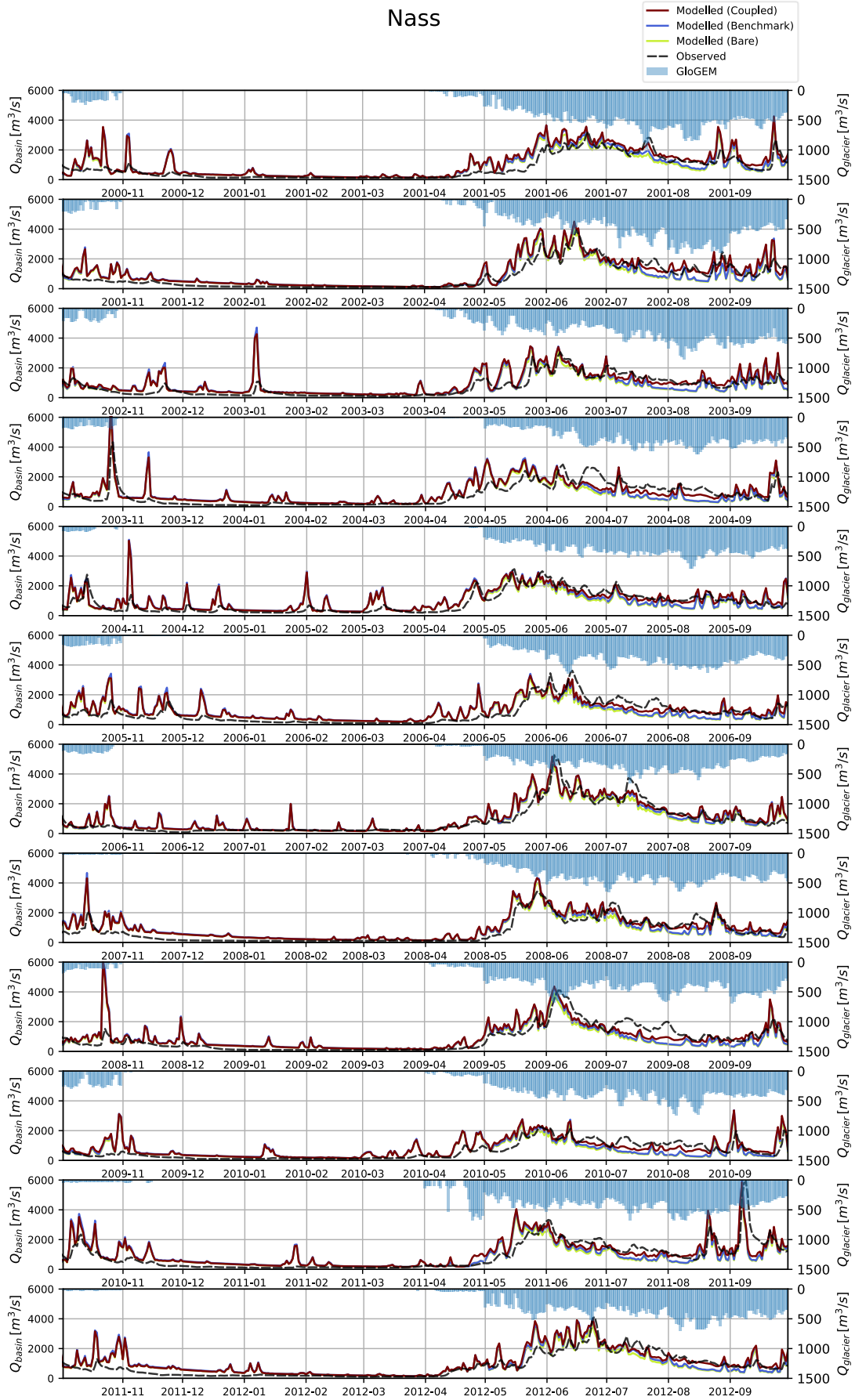
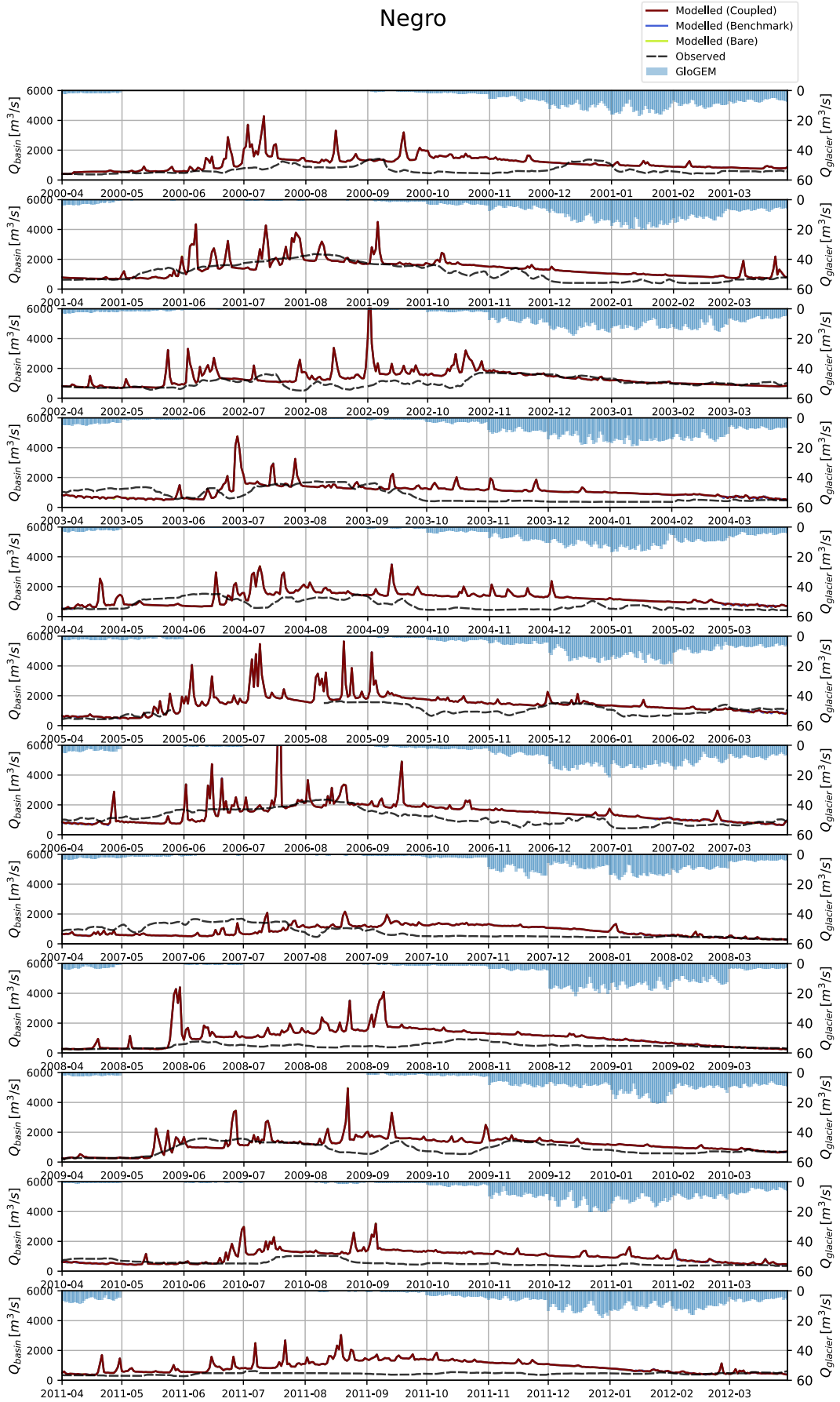
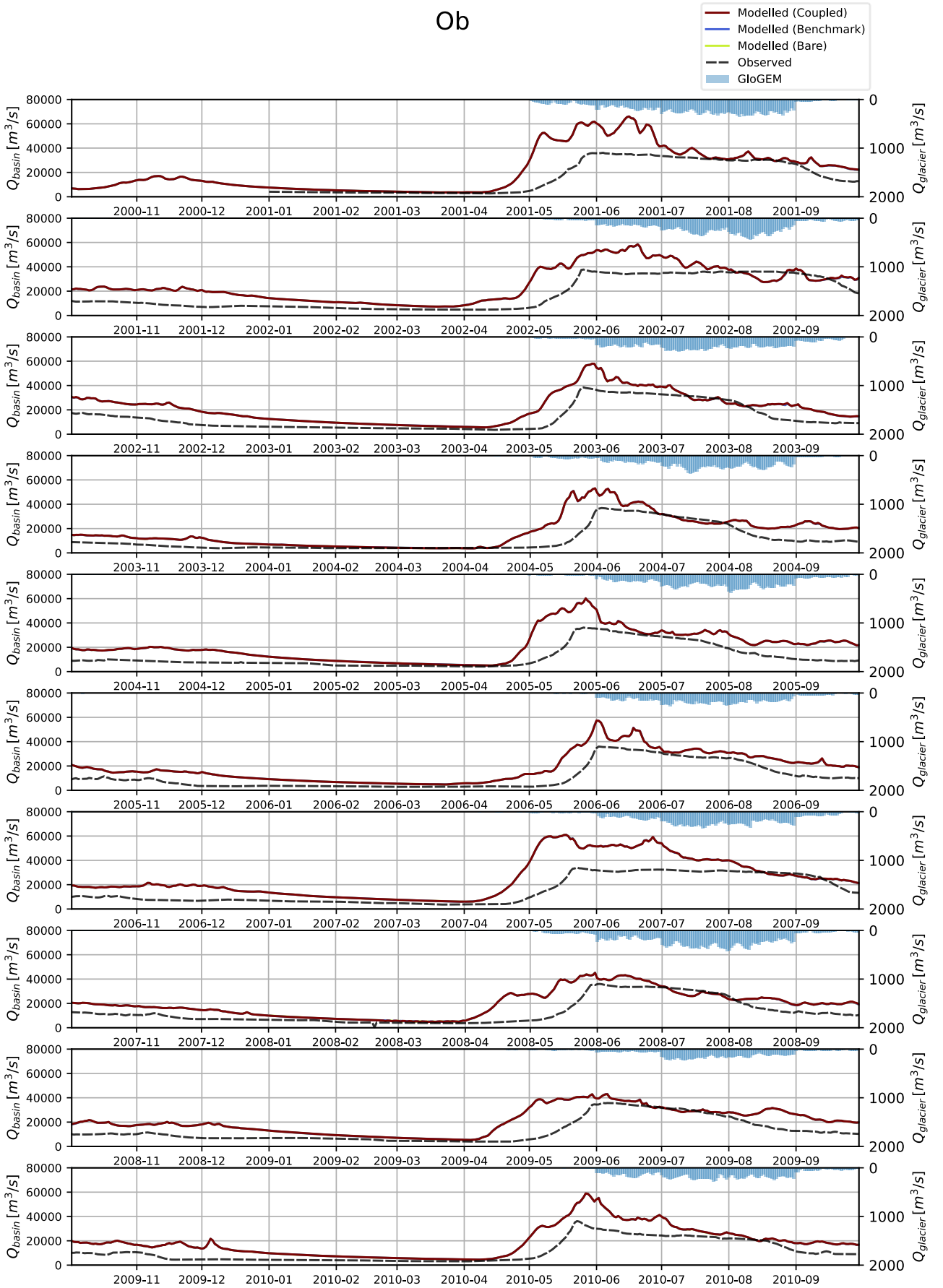
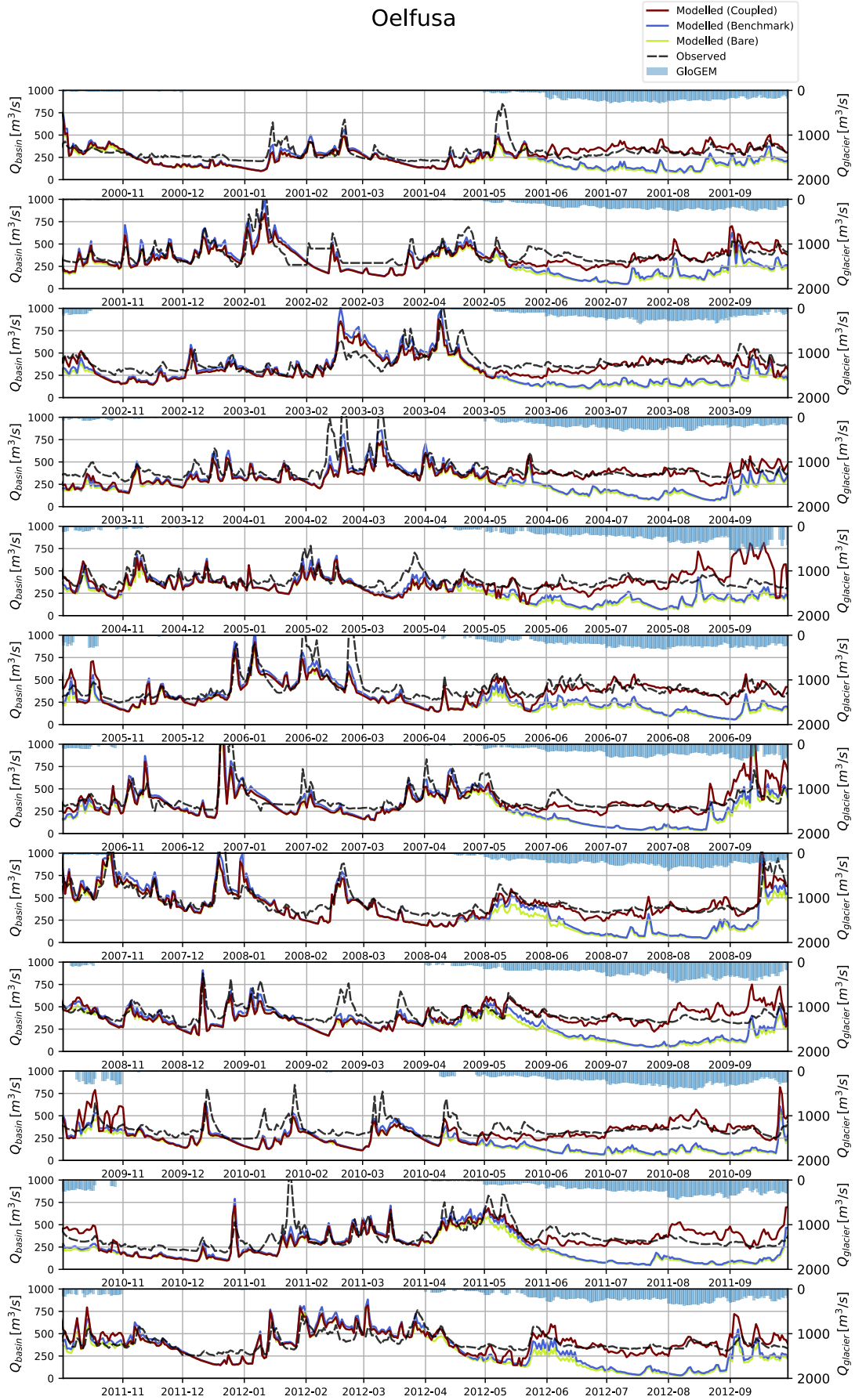
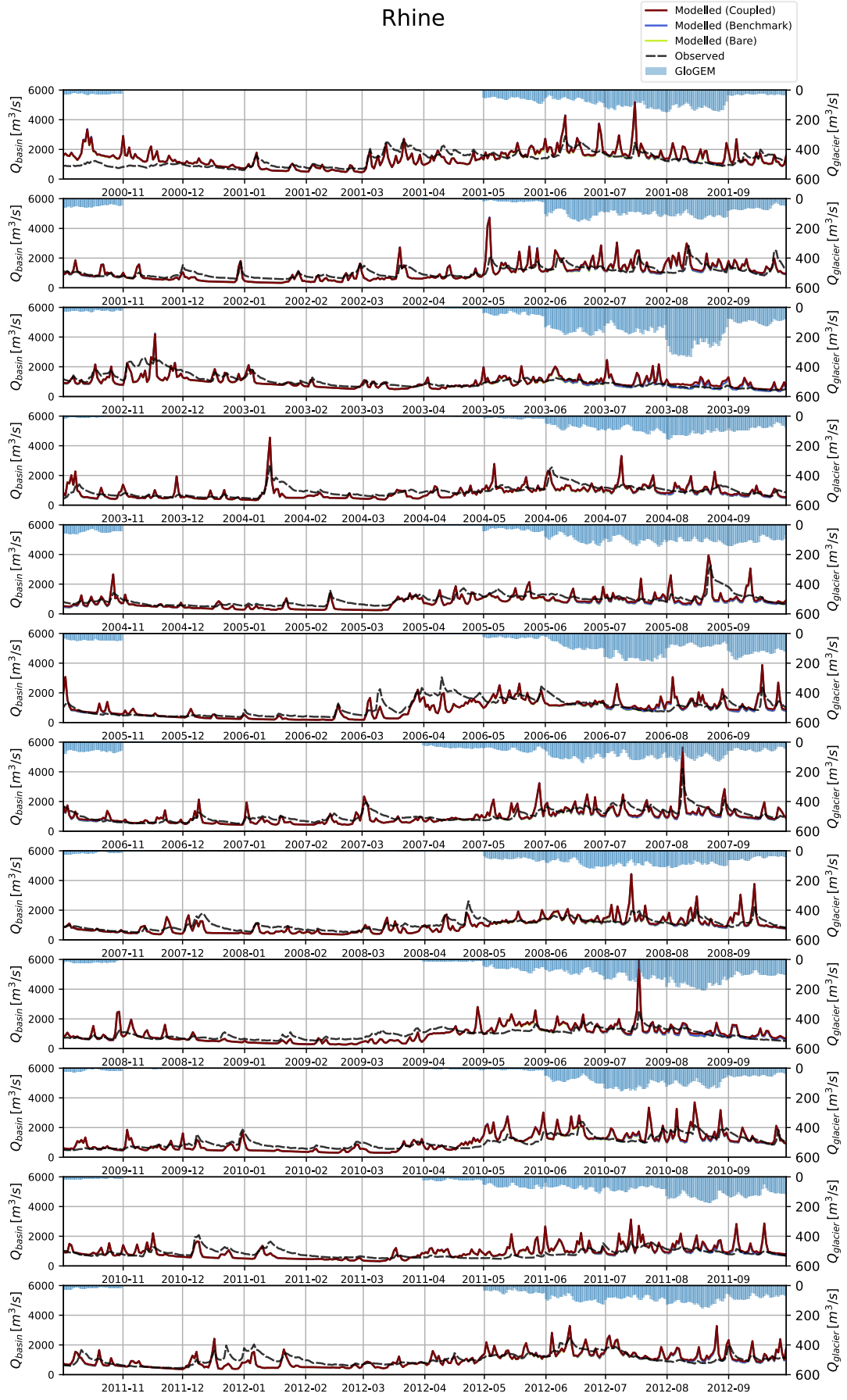
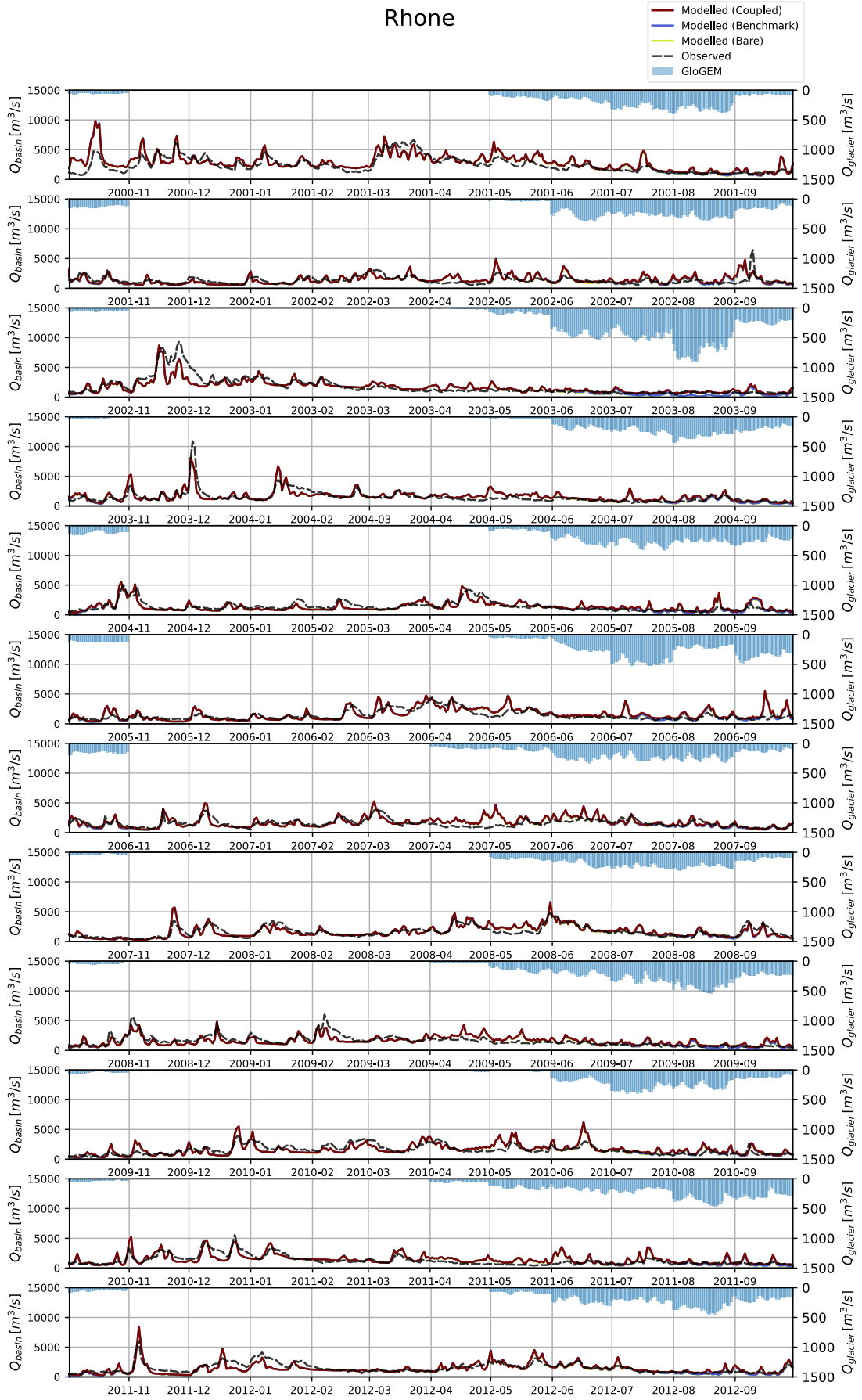
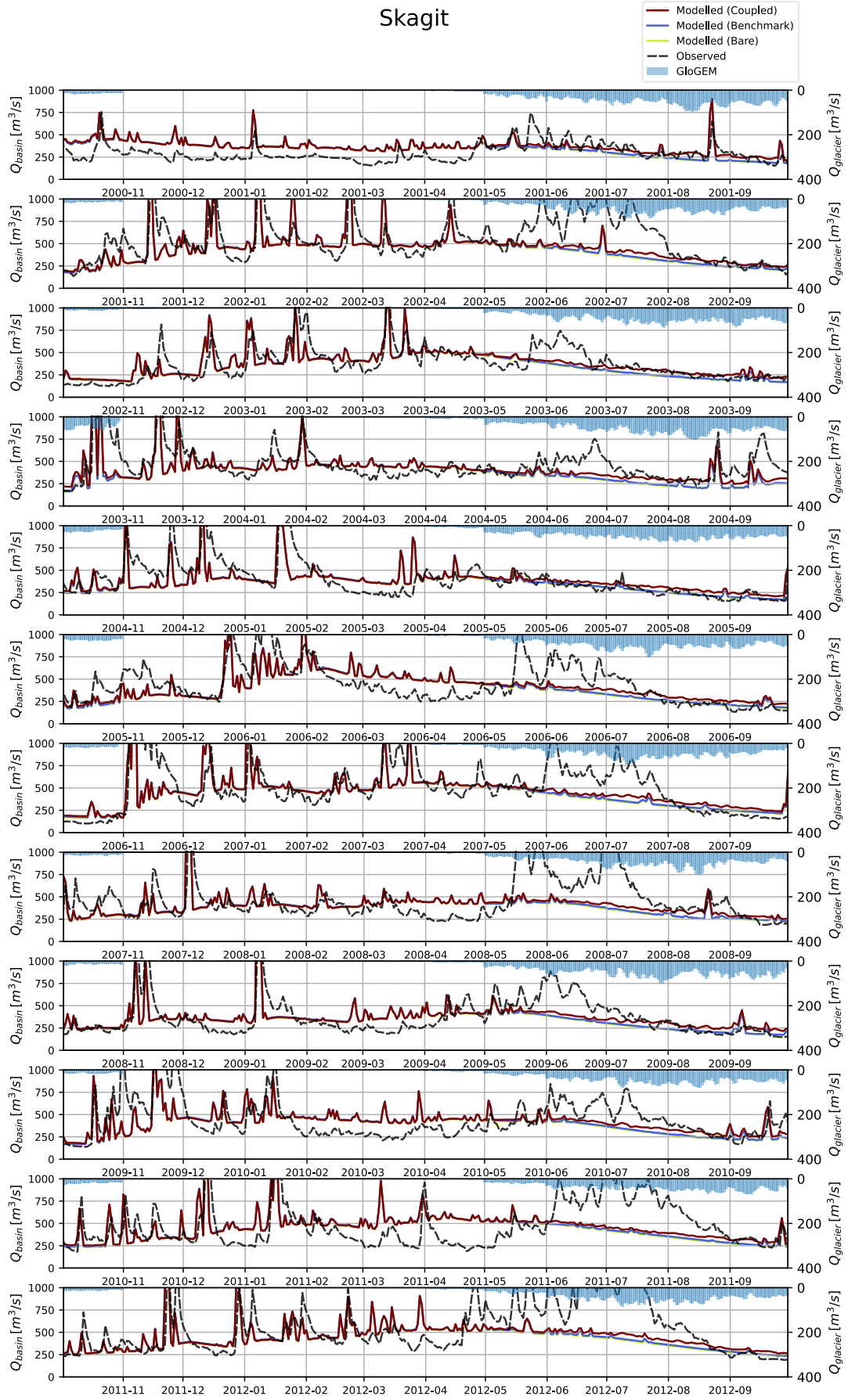
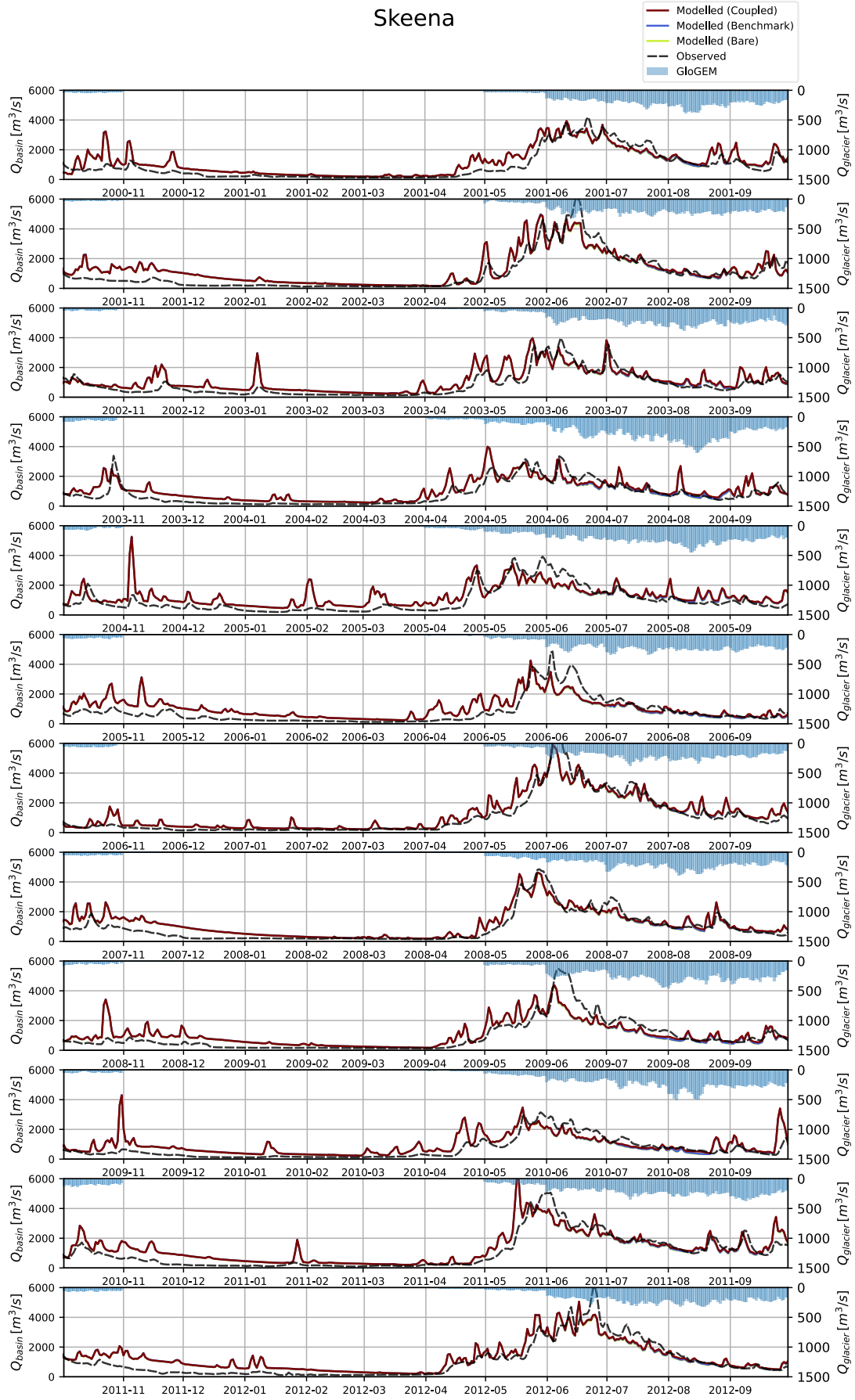
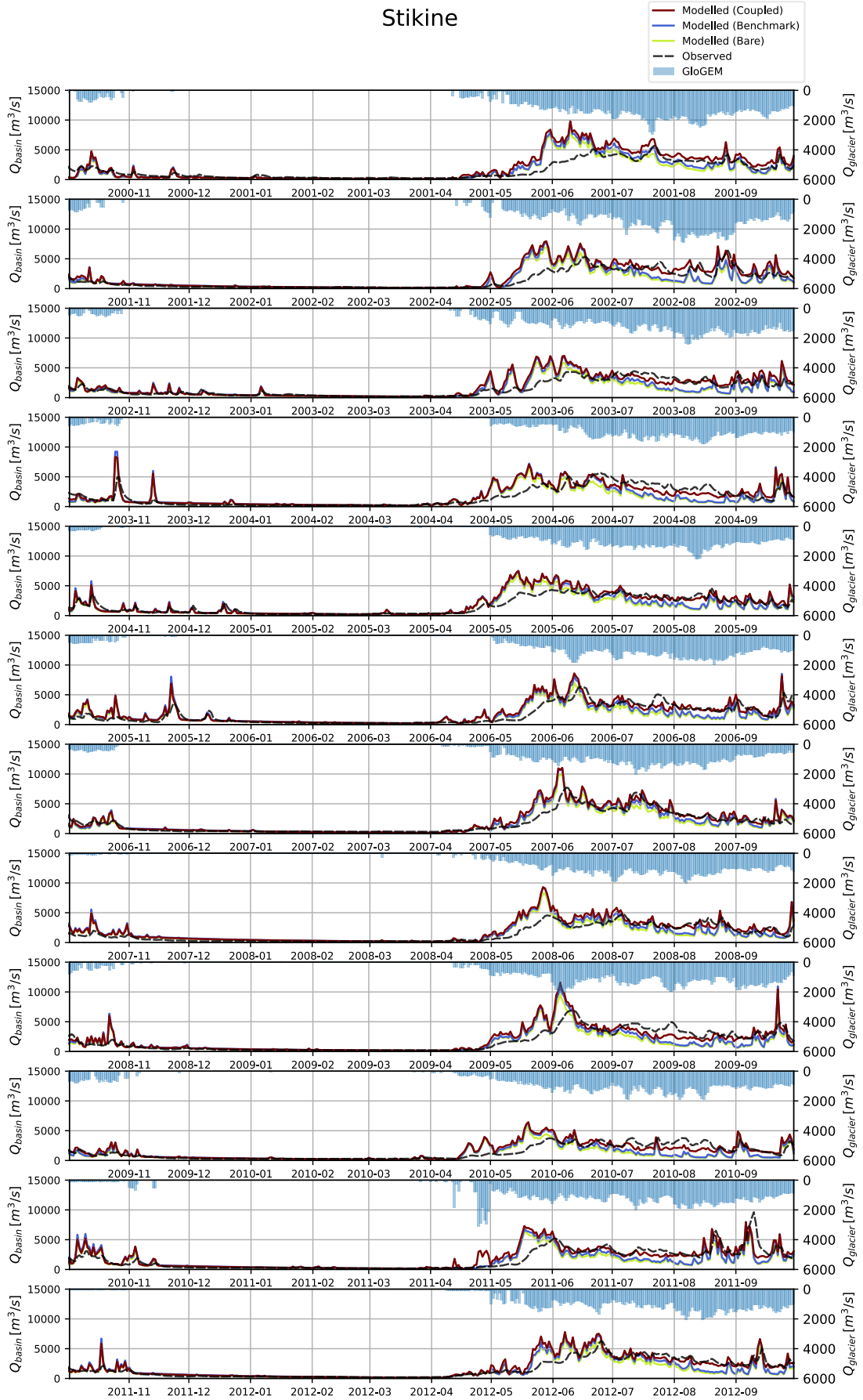
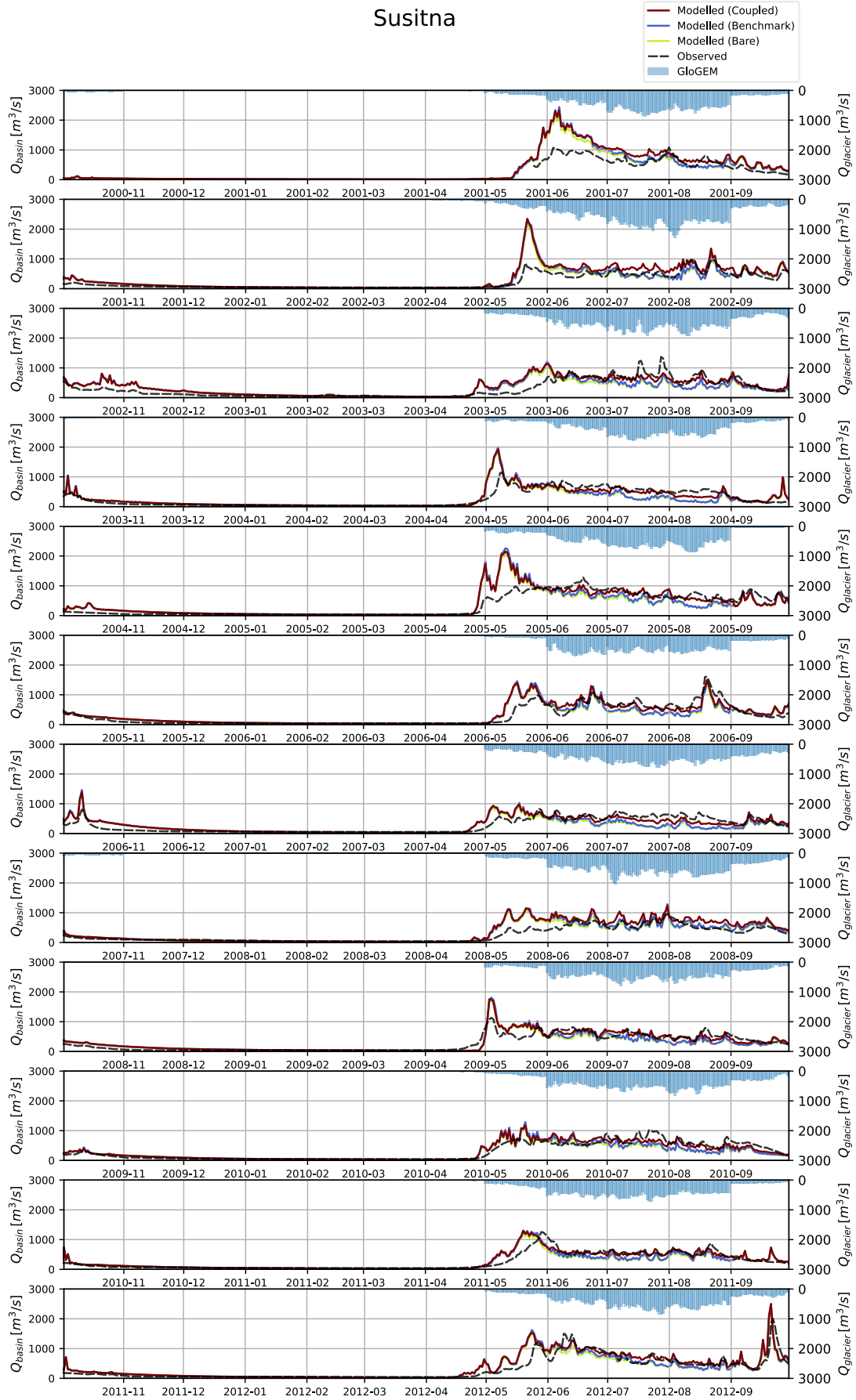
J. E. Nash and J. V. Sutcliffe. River flow forecasting through conceptual models part I—A discussion of principles. *Journal of hydrology*, 10(3):282–290, 1970.

B. Schaefli and H. V. Gupta. Do Nash values have value? *Hydrological Processes*, 21(ARTICLE):2075–2080, 2007.

J. Seibert, M. J. Vis, E. Lewis, and H. J. van Meerveld. Upper and lower benchmarks in hydrological modelling. *Hydrological Processes*, 32(8):1120–1125, 2018. ISSN 10991085. 10.1002/hyp.11476.

M. van Tiel, K. Stahl, D. Freudiger, and J. Seibert. Glacio-hydrological model calibration and evaluation. *Wiley Interdisciplinary Reviews: Water*, 7(6), 2020. ISSN 20491948. 10.1002/wat2.1483.

8 Hydrographs

Alsek

Amazon

Clutha

Columbia

Copper

Danube

Dramselv

Fraser

Gloma

Irrawaddy

Kuskokwim

Mackenzie

Nass

Negro

Ob

Oelfusa

Rhine

Rhone

Skagit

Skeena

Stikine

Taku

Thjorsa

Yukon

9 Basin maps