Supplementary Information 1: RCM/bias-correction information

Table S1. Snowmelt model parameters

Parameter	Definition	Value	Units
lapse_rate	Temperature lapse rate (reduction in air temperature with increasing elevation)	0.0059	$^{o}C\ m^{-1}$
tsnow	Threshold temperature, below which precipitation is snow	1.0	^{o}C
tmelt	Temperature threshold for snowmelt	0.0	^{o}C
mfac	Melt factor	6.0	$mm\ ^{o}C^{-1}\ day^{-1}$
tdrel	Threshold temperature for drainage release	0.0	^{o}C
k1 k2	Storage time constants of two-outlet liquid water store	0.5 0.9	day^{-1}
Scfac	Critical water retention capacity	0.18	-
snowfrac	Under-catch factor for gauged rainfall falling as snow (not applied to RCM data)	1.0	-

Table S2. RCM variables used in the calculation of PET.

Variable	Description	Units
huss	Specific humidity	-
rls	Radiation, net long wave	$W\ m^{-2}$
rss	Radiation, net short wave	$W\ m^{-2}$
sfcWind	Wind speed	$m\ s^{-1}$
tas	Mean temperature	^{o}C
psl	sea level air pressure (used to derive surface air pressure at gridbox altitude, psurf)	hPa
pr_bc	precipitation (after bias-correction)	$mm\ day^{-1}$

Supplementary info 2: PDM Calibration

Model configurations

A full list of the options tested in the PDM model for each catchment is given in Table S3 and summarised below following the notation of Moore (2007).

- *Runoff generation and groundwater recharge.* In the "Full" version of the model, drainage to the groundwater store is described by a recharge time-constant and exponent (k_g, b_g) and, optionally, a soil tension storage capacity (S_t). For the "Reduced" form of model, surface runoff is simply split so that a fixed fraction (α) enters the surface store while the remainder enters the groundwater store. In both cases the water absorption capacity of the soil is described by a Pareto distribution characterised by a shape parameter (b), a maximum storage (c_{max}), and optionally a minimum storage (c_{min}). A "Classic" version of the Reduced model employs a rectangular distribution for the soil's water absorption capacity ($c_{min} = 0$, $b = 1$).
- *Surface water routing.* The surface runoff component of total flow is related to the volume of water in the surface store using a time constant (k_1) and exponent (m). Exponents of $m = 1, 2, 3$ are trialled, as is two linear ($m = 1$) stores in series in a discretely-equivalent transfer function form.
- *Groundwater routing.* The baseflow component of total flow is related to the volume of water in the groundwater store using a time constant (k_b) and exponent (m). Values of $m = 2, 3$ are trialled.
- *Groundwater (GW) extension.* A standard implementation of PDM conserves water throughout, albeit with the option of applying a multiplicative factor (*rainfac*) to the precipitation input. Conceptually, this factor might compensate for a lack of representativeness in the data used to estimate catchment precipitation. It may also serve to account for losses or gains of water affecting the catchment itself. Alternatively, functionality within the GW extension can be considered to address catchment water conservation issues. This extension, subject to data availability, allows modelling of underflows at the catchment outlet, external springs, pumped abstractions, and the incorporation of well level data. Under eFLaG, only the Spring Factor option (*springfac*) is invoked and repurposed to infer unknown net water exchanges affecting the catchment via the groundwater storage. It serves as a multiplicative factor representing either net losses from ($1 \geq springfac > 0$), or net gains to ($springfac < 0$), the baseflow.

Calibration process

A three-stage calibration process was applied independently for each of the model configurations in Table S3, each starting from a number of different choices for the initial parameter. The design of this process was motivated by the desire to find a procedure that could be applied automatically across many disparate catchments without a tendency to either get blocked in local optimums or produce unphysical models.

The following three calibration stages were employed for all model configurations, except those employing the GW extension.

- *Stage 1.* Four or five dominant parameters were segregated according to whether they were judged to control mainly the slow response (c_{max}, k_b for Reduced Models; k_b, k_g, b_g for Full Models) or the fast response (k_1, α for Reduced Models; k_1, c_{max} for Full Models). Then, (i)

the slow parameters were calibrated to optimise the KGE'_{log}, (ii) the fast parameters were calibrated to optimise the KGE'_{sqrt}, and then (iii) *rainfac* was calibrated to achieve zero bias. These steps were iterated six times to achieve convergence.

- *Stage 2.* All parameters calibrated in Stage 1 are re-calibrated simultaneously to maximise the KGE'_{sqrt}. The *rainfac* was then recalibrated to achieve zero bias. This process was iterated three times to ensure convergence.
- *Stage 3.* Additional parameters controlling the distribution of the soil water absorption capacity (S_t, b and c_{min} for the Full Model; b and c_{min} for the Reduced Model; none for the "Classic" model), and one parameter (b_e) controlling the sensitivity of the conversion of Potential Evaporation (PE) to Actual Evaporation (AE) with available soil moisture, were each calibrated separately to optimise KGE'[sqrt]. Stage 2 was then repeated.

When the GW extension was employed, the three stages were modified so that (i), the Spring Factor was used to achieve zero bias, (ii), greater emphasis was placed on obtaining suitable ground- and soil-water storage parameters (beginning in Stage 1), and (iii) initial parameters where chosen that were more suitable for slowly responding catchments. While calibrations employing the GW extension were tested at all sites, they were only judged to be appropriate for final model selection for the 26 catchments with a Base Flow Index (BFI) greater than 0.7. Selection of the GW extension at Leven at Linnbrane (85001) was also excluded.

Calibrated model selection

A calibrated PDM model is produced for each model configuration, each initial parameter choice, and at each of the three calibration stages, yielding a total of 46×3 = 138 possible calibrations per catchment. Figure S1 shows the KGE'_{sqrt} (colours) and KGE'_{log} (grey) values for each of these calibrations for catchment 2001 (Helmsdale at Kilphedir, North West Scotland) and catchment 39089 (Gade at Bury Mill, Hertfordshire and North London area). Any calibrations yielding extreme parameters, including those found to be storing excessive quantities of water, are automatically judged to be unphysical and are shown in black.

The Helmsdale catchment demonstrates several features that are typical across most catchments with low or medium Base Flow Index (BFI) (Figure S1a, BFI = 0.47). These are, (i) calibrations with different model configurations or different initial parameter choices often yield similar metric values, (ii) there are a small number of calibrations that produce unphysical models and poorer metric values, and (iii) the calibration with the best KGE'_{sqrt} value does not necessarily have the best KGE'_{log} value. This last observation motivated the use of a weighted sum of the two metrics (weights 0.8 and 0.2 respectively) as the criteria for selecting among the different calibrations. For the catchments with very high BFI, such as the Gade catchment (Figure S1b, BFI = 0.89), the GW extension is often essential: other model configurations typically produce poor and variable metric values, or unphysical models.

Part of the quality control for the PDM model was the examination of RCM flows (simrcm) for each catchment. For the Misbourne at Little Missenden (catchment 39127, BFI = 0.96), this revealed unphysically smooth multi-decade recessions beginning in the immediate future for some RCM ensemble members. The cause of this behaviour was found to be a very large minimum point soil water capacity combined with the use of the Reduced Model with no soil drainage to groundwater: a combination that inhibited runoff generation when, due to climate change, the soil water store became depleted. Because of this, the best performing Full Model (F-GW322) was chosen as offering a better hydrological representation of the catchment and more realistic predictions under climate change. This highlights the possibility of unphysical calibrations achieving good metric values against historical

river flows ($KGE'_{sqrt} = 0.927$ was achieved for Misbourne), while producing unphysical results when climate change pushes hydrological conditions outside of their historical regime. This possibility is expected to be more associated with high BFI catchments as these will have greater sensitivity to the longer-term average trends in the weather that become apparent under climate change. By using multiple disparate hydrological models in the eFLaG project (PDM, GR4J, GR6J and G2G), over-reliance on a single model can be avoided.

The number of times each model was selected for one of the 200 catchments is listed in Table S3.

Table S3. Full list of PDM models trialled for each catchment, and the number of catchments for which each model was selected. For the "Surface routing exponent" column, "22" indicates use of two linear reservoirs in series.

Recharge form	Model code	Groundwater routing exponent	Surface routing exponent	Initial parameter choices	Final selections (out of 200)
Reduced	**R322**	3	22	6	25
Reduced	**R33**	3	3	2	44
Reduced	**R32**	3	2	2	34
Reduced	**R31**	3	1	2	3
Reduced (Classic)	**C31**	3	1	2	5
Reduced	**R222**	2	22	2	4
Reduced	**R23**	2	3	2	9
Reduced	**R22**	2	2	2	16
Reduced	**R21**	2	1	2	2
Full	**F322**	3	22	6	5
Full	**F33**	3	3	2	2
Full	**F32**	3	2	2	6
Full	**F31**	3	1	2	8
Full	**F222**	2	22	2	5
Full	**F23**	2	3	2	4
Full	**F22**	2	2	2	7
Full	**F21**	2	1	2	2
Reduced GW extension	**R-GW322**	3	22	2	15 (out of 26)
Full GW extension	**F-GW322**	3	22	2	4 (out of 26)

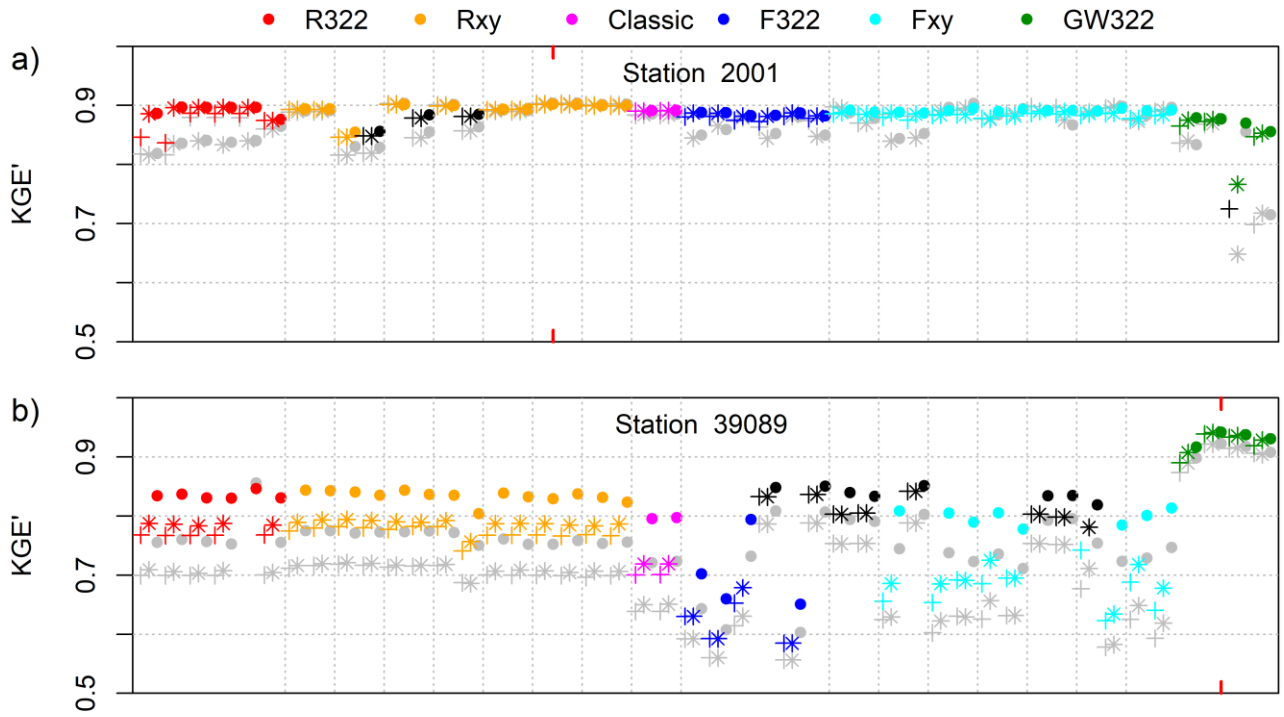


Figure S1. Performance of modelled river flow for each model configuration, calibration stage, and initial parameter choice. Performance is measured by KGE'_{sqrt} (coloured symbols) or KGE'_{log} (grey symbols). Crosses, asterisks and circles indicate performance at calibration stages 1, 2, and 3 respectively. Different colours and dashed lines are used to separate different model configurations. Black is used to show calibrations that resulted in unphysical model parameters. Red ticks on the upper and lower x-axis indicate the final model selection. Catchments are: (a) Helmsdale at Kilphedir (2001), BFI = 0.47, and (b) Gade at Bury Mill (39089), BFI = 0.89.

Supplementary Info 3: maps of model evaluation against observed data

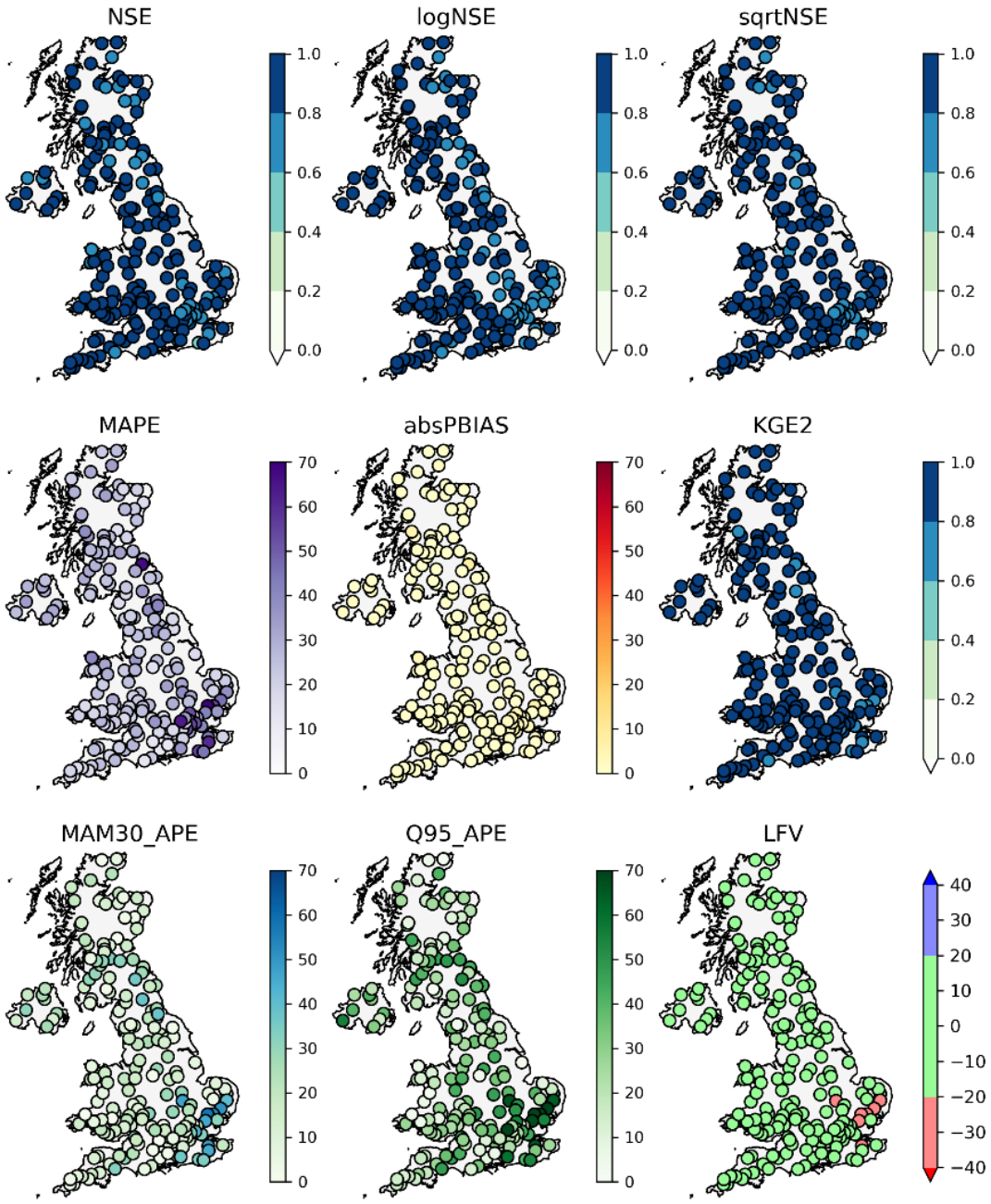
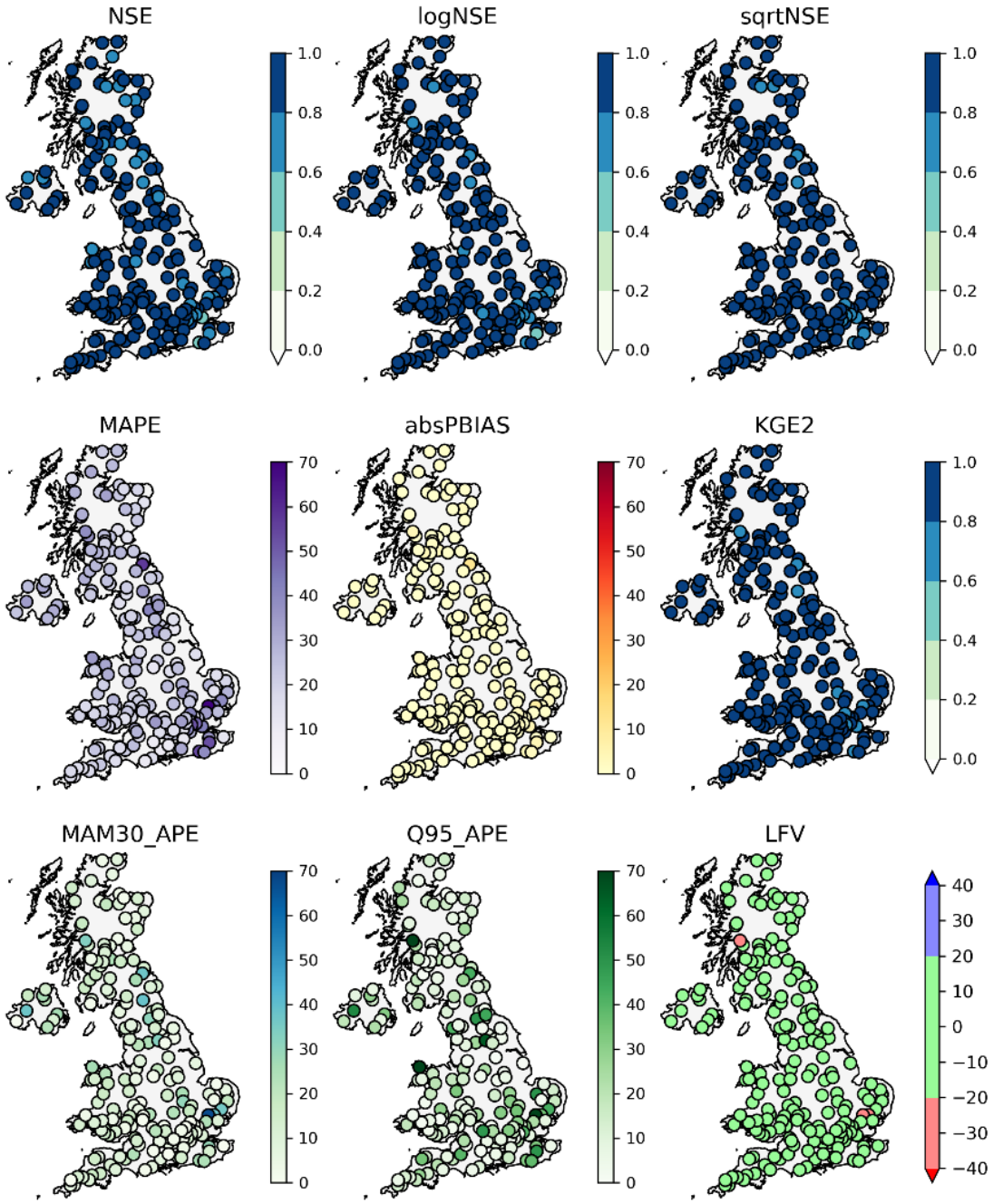


Figure S2: Performance results for GR4J (for metrics see Table 3 in main paper)



147

148

149 **Figure S3: Performance results for GR6J**

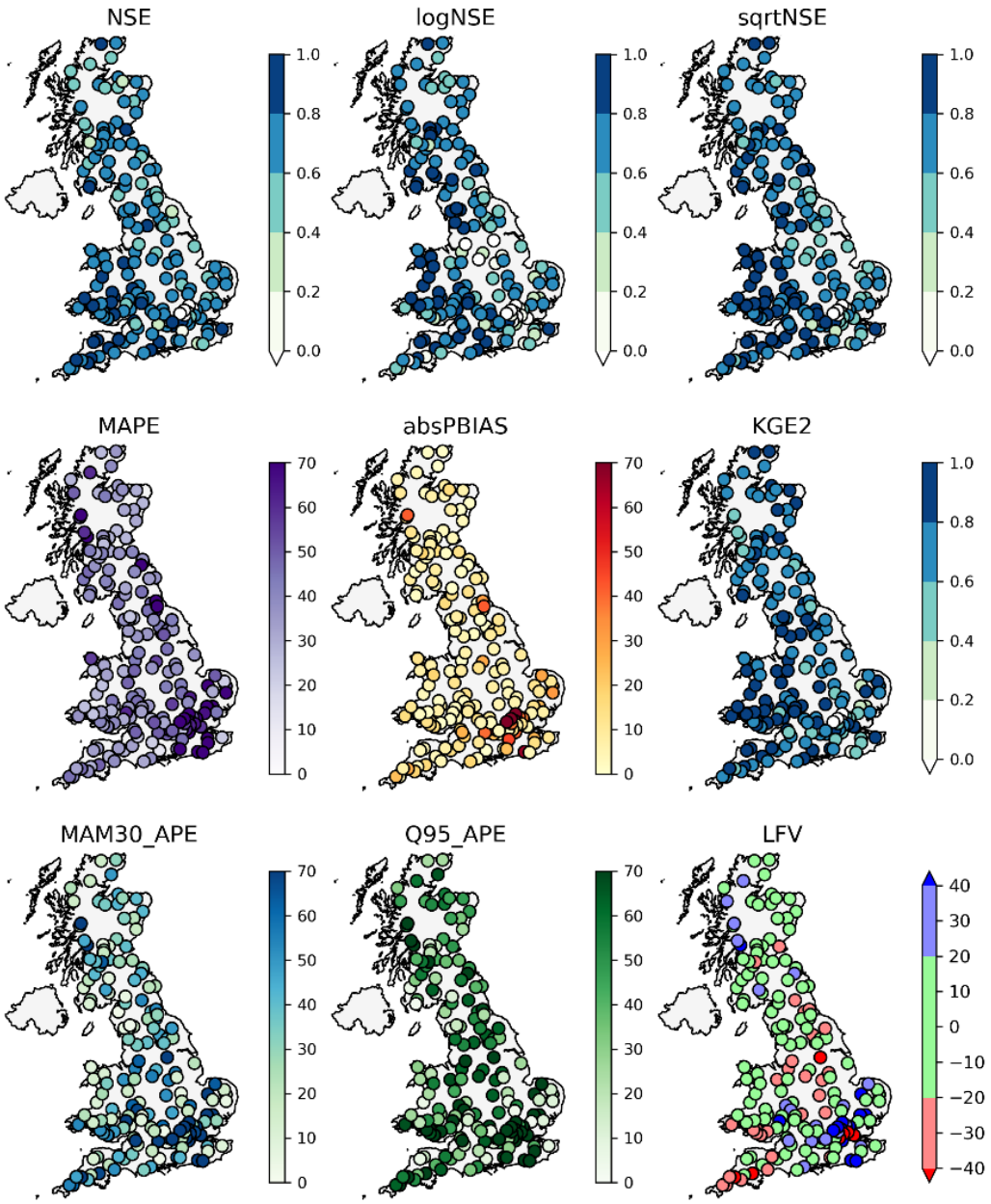


Figure S4: Performance results for G2G

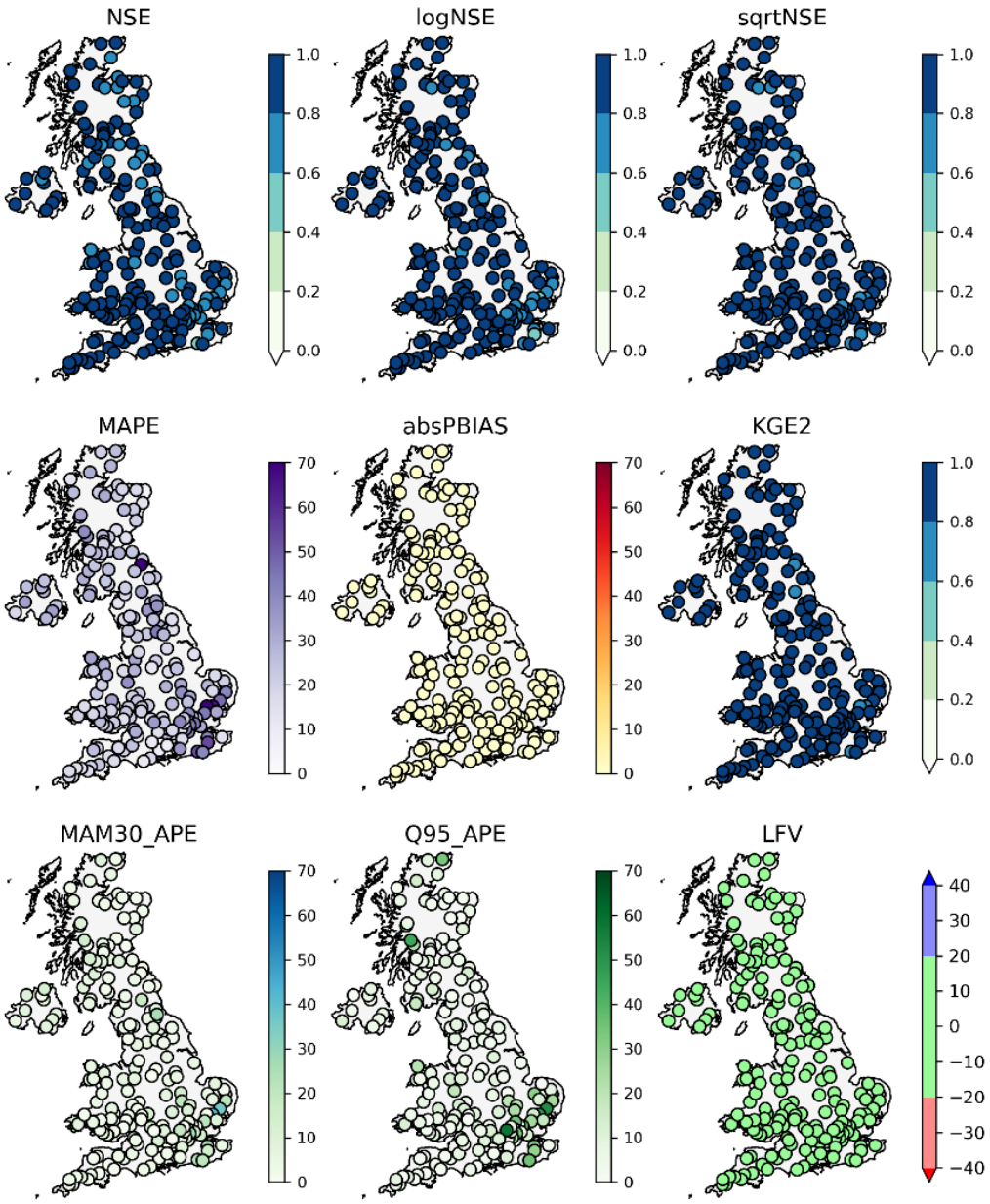


Figure S5: Performance results for PDM

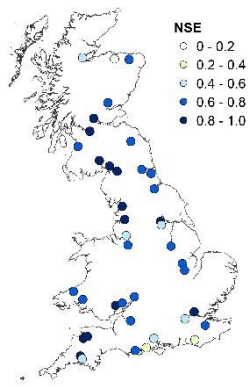
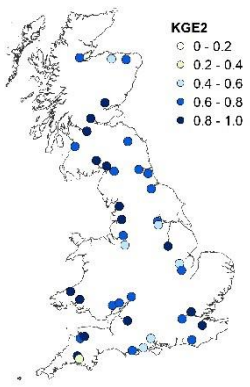
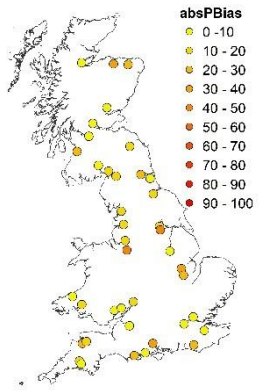
NSE	KGE2	absPBias
		

Figure S6: Performance results for the distributed recharge model (ZOODRM)

Supplementary Information 4: River Flow and Groundwater level Duration Curves

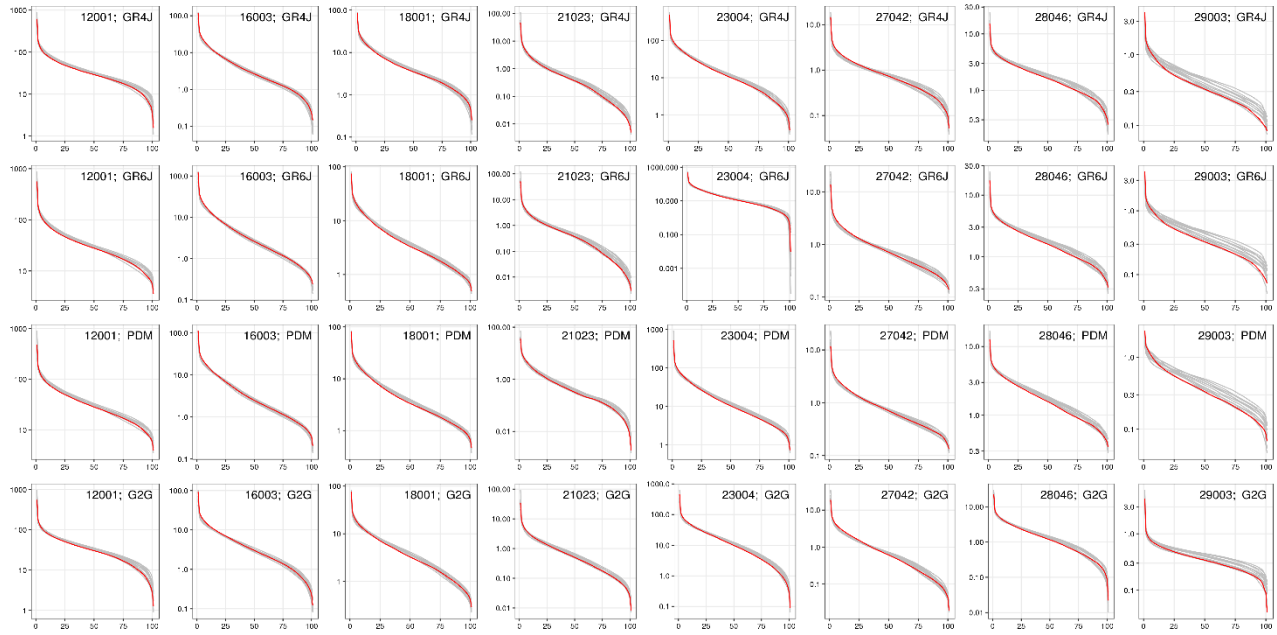


Figure S7 -- Flow duration curves (FDCs) comparing the baseline flow regime in the 12 RCM ensemble members (simrcm, grey lines) to model observations (simobs, red line), 1989-2018. FDCs are featured for four hydrological models (GR4J, GR6J, PDM, G2G; rows) and eight catchments in eastern Scotland and north-east England (12001 Scottish Dee, 16003 Ruchill Water, 18001 Allan Water, 21023 Leet Water, 23004 South Tyne, 27042 Yorkshire Dove, 28046 Derbyshire Dove, 29003 Lud; columns). The y-axis represents river flows (cumecs) on a logarithmic scale.

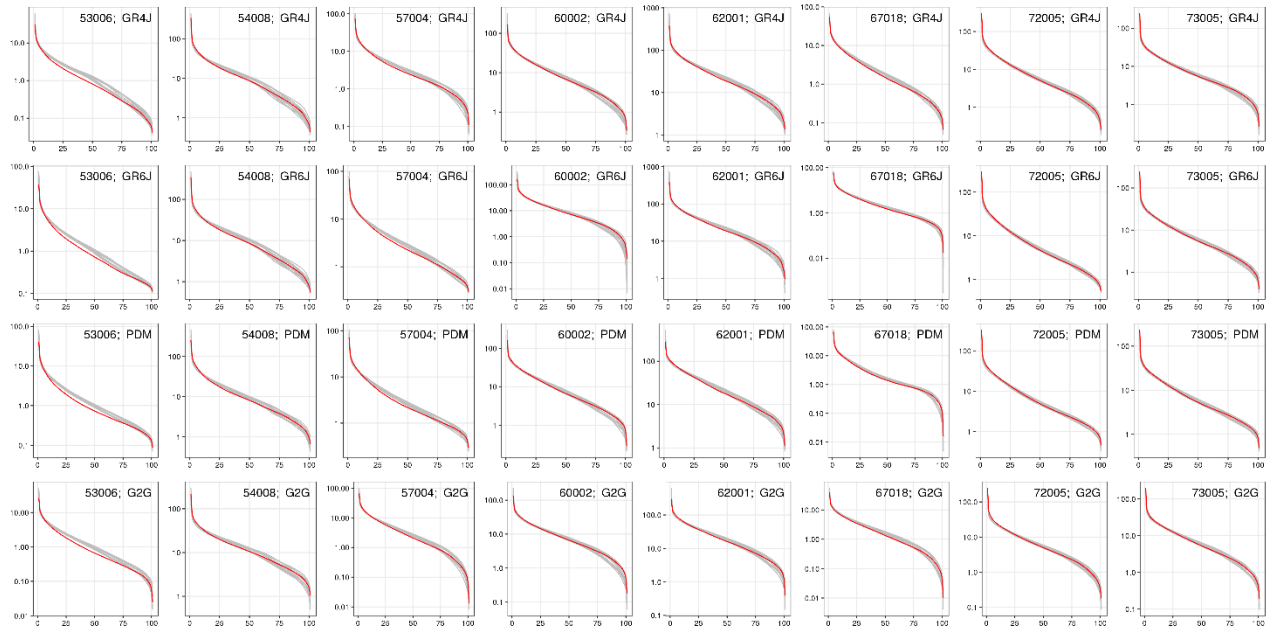


Figure S8 -- Flow duration curves (FDCs) comparing the baseline flow regime in the 12 RCM ensemble members (grey lines) to model observations (red line), 1989-2018. FDCs are featured for four hydrological models (GR4J, GR6J, PDM, G2G; rows) and eight catchments in Wales and north-west England (53006 Bristol Frome, 54008 Teme, 57004 Cynon, 60002 Cothi, 62001 Teifi, 67018

Welsh Dee, 72005 Lune, 73005 Kent; columns). The y-axis represents river flows (cumecs) on a logarithmic scale.

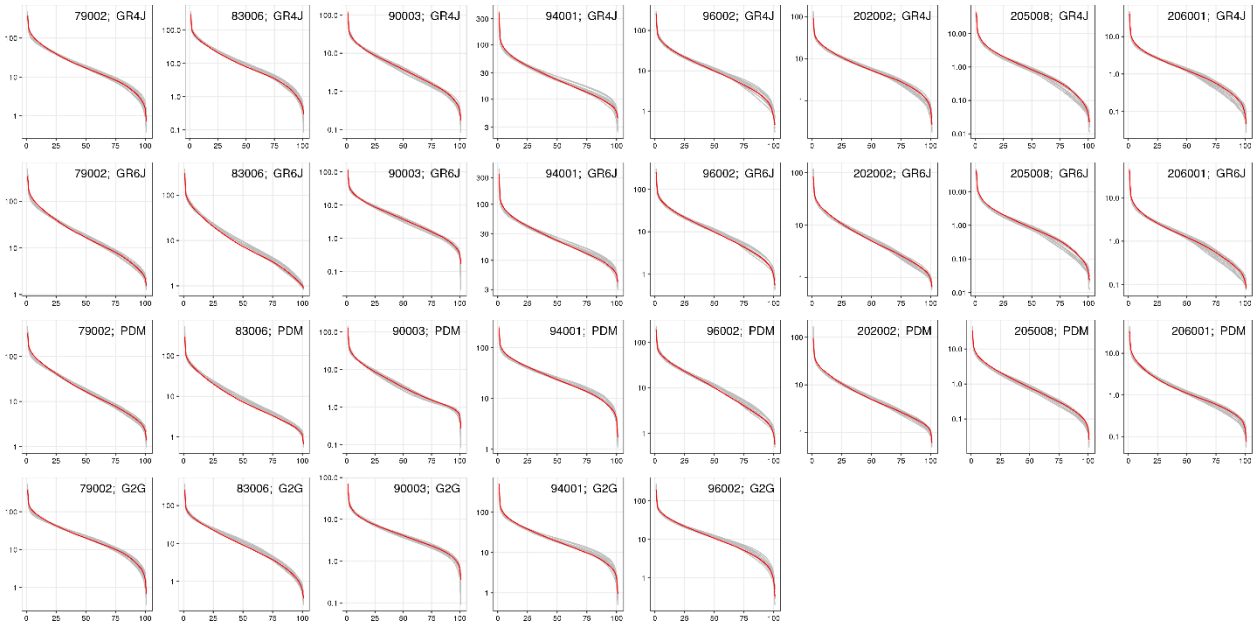


Figure S9 -- Flow duration curves (FDCs) comparing the baseline flow regime in the 12 RCM ensemble members (grey lines) to model observations (red line), 1989-2018. FDCs are featured for four hydrological models (GR4J, GR6J, PDM, G2G; rows) and eight catchments in western Scotland and Northern Ireland (79002 Nith, 83006 Ayr, 90003 Nevis, 94001 Ewe, 96002 Naver, 202002 Faughan, 205008 Lagan, 206001 Clanrye; columns). The y-axis represents river flows (cumecs) on a logarithmic scale. The absence of FDCs for G2G for 202002, 205008 and 206001 is because G2G does not cover Northern Ireland.

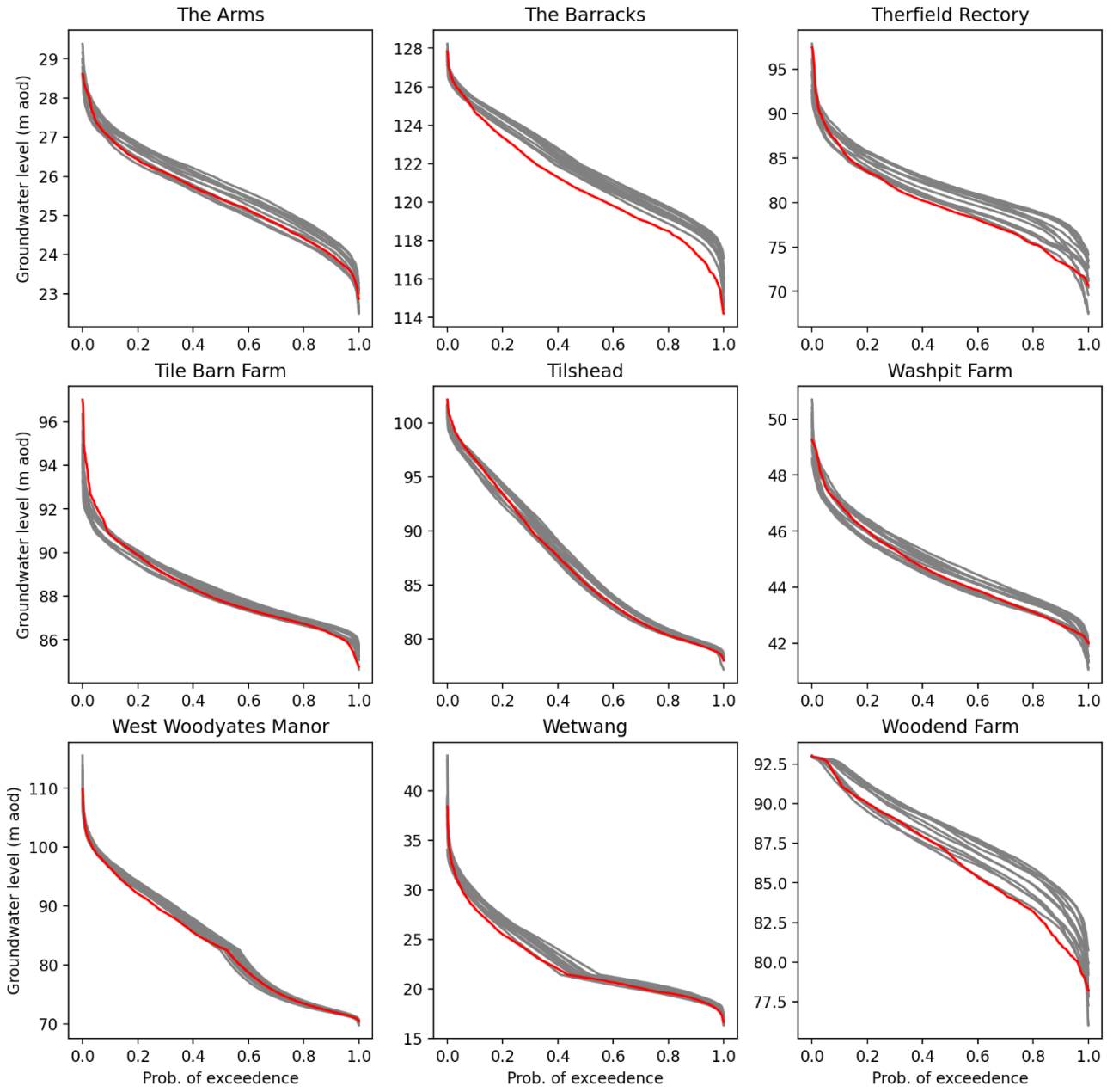


Figure S10 – Groundwater level duration curves (GLDCs) for the period 1989-2018 using the simrcm (grey lines) simobs (red line) simulations.

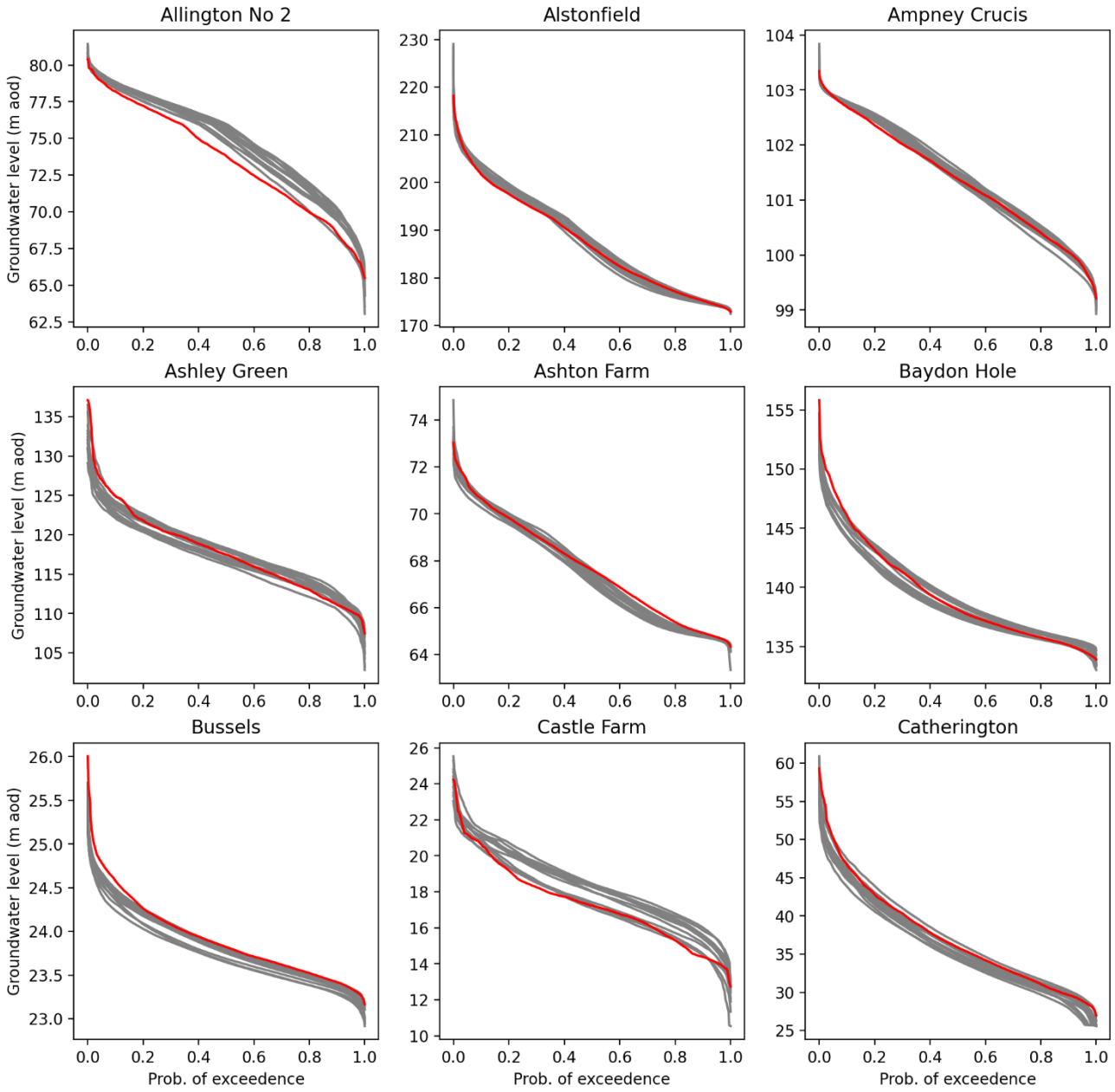


Figure S11 – Groundwater level duration curves (GLDCs) for the period 1989-2018 using the simrcm (grey lines) simobs (red line) simulations.

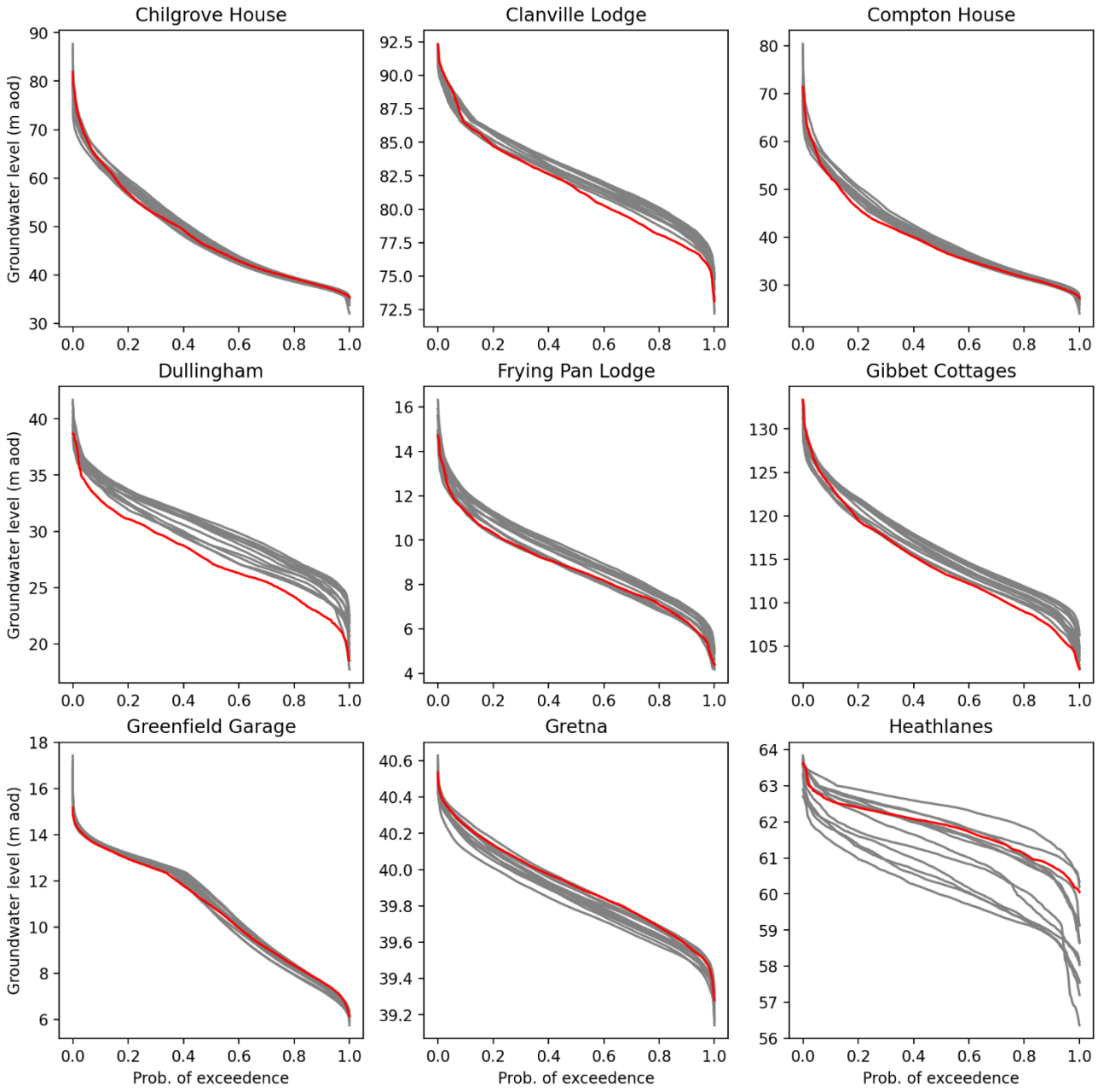


Figure S12 – Groundwater level duration curves (GLDCs) for the period 1989-2018 using the simrcm (grey lines) simobs (red line) simulations.

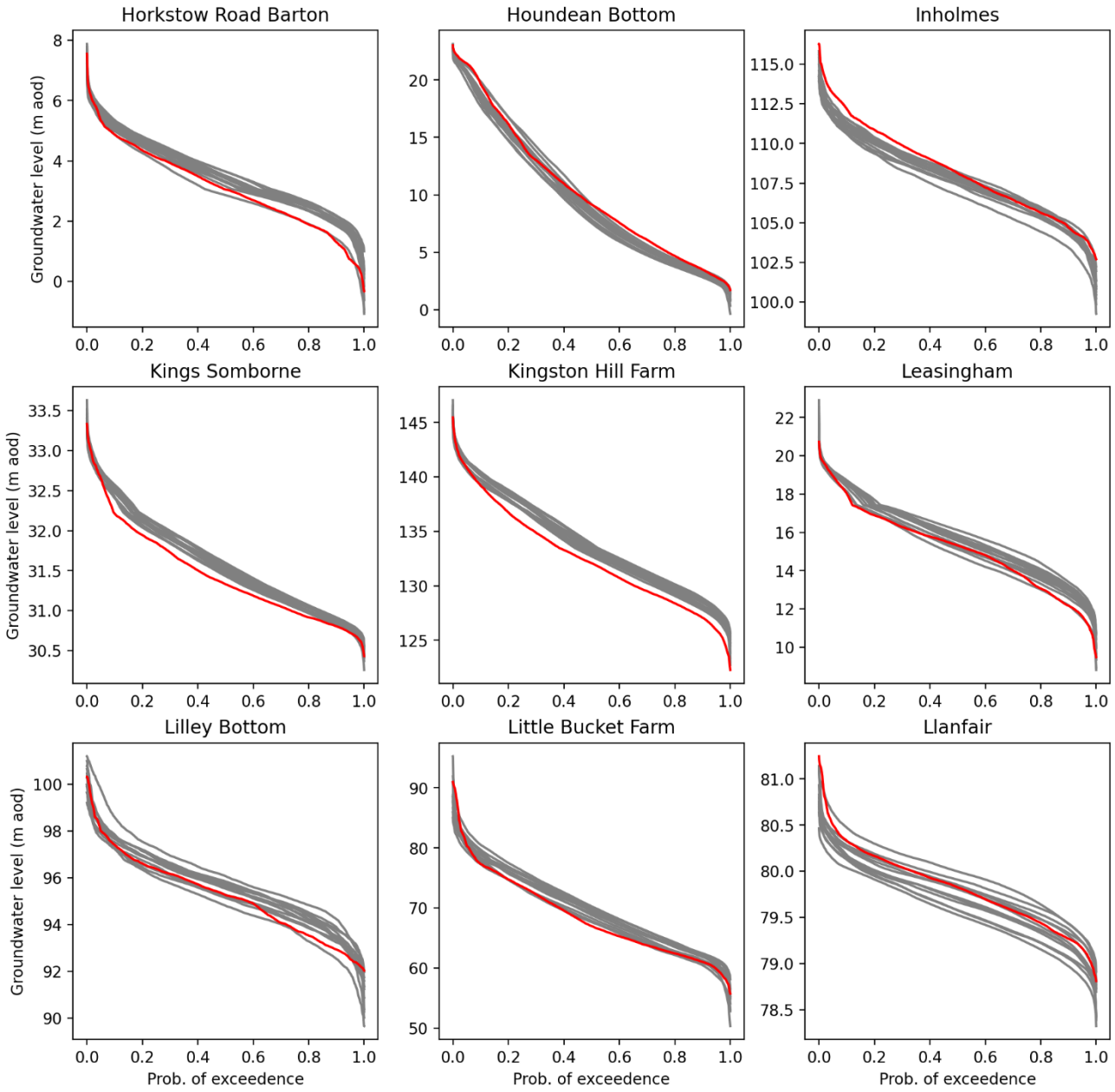


Figure S13 – Groundwater level duration curves (GLDCs) for the period 1989-2018 using the simrcm (grey lines) simobs (red line) simulations.

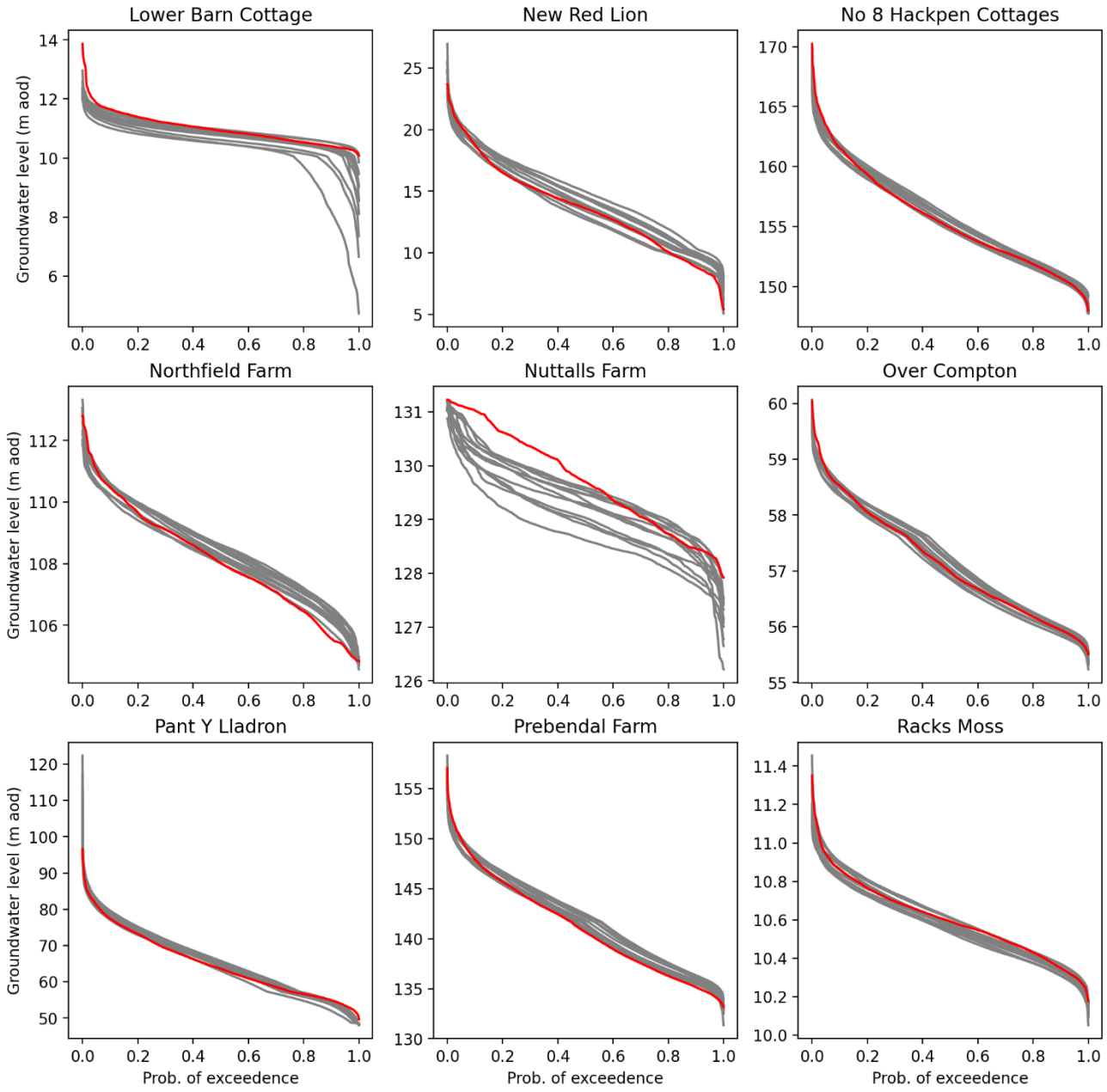


Figure S14 – Groundwater level duration curves (GLDCs) for the period 1989-2018 using the simrcm (grey lines) simobs (red line) simulations.

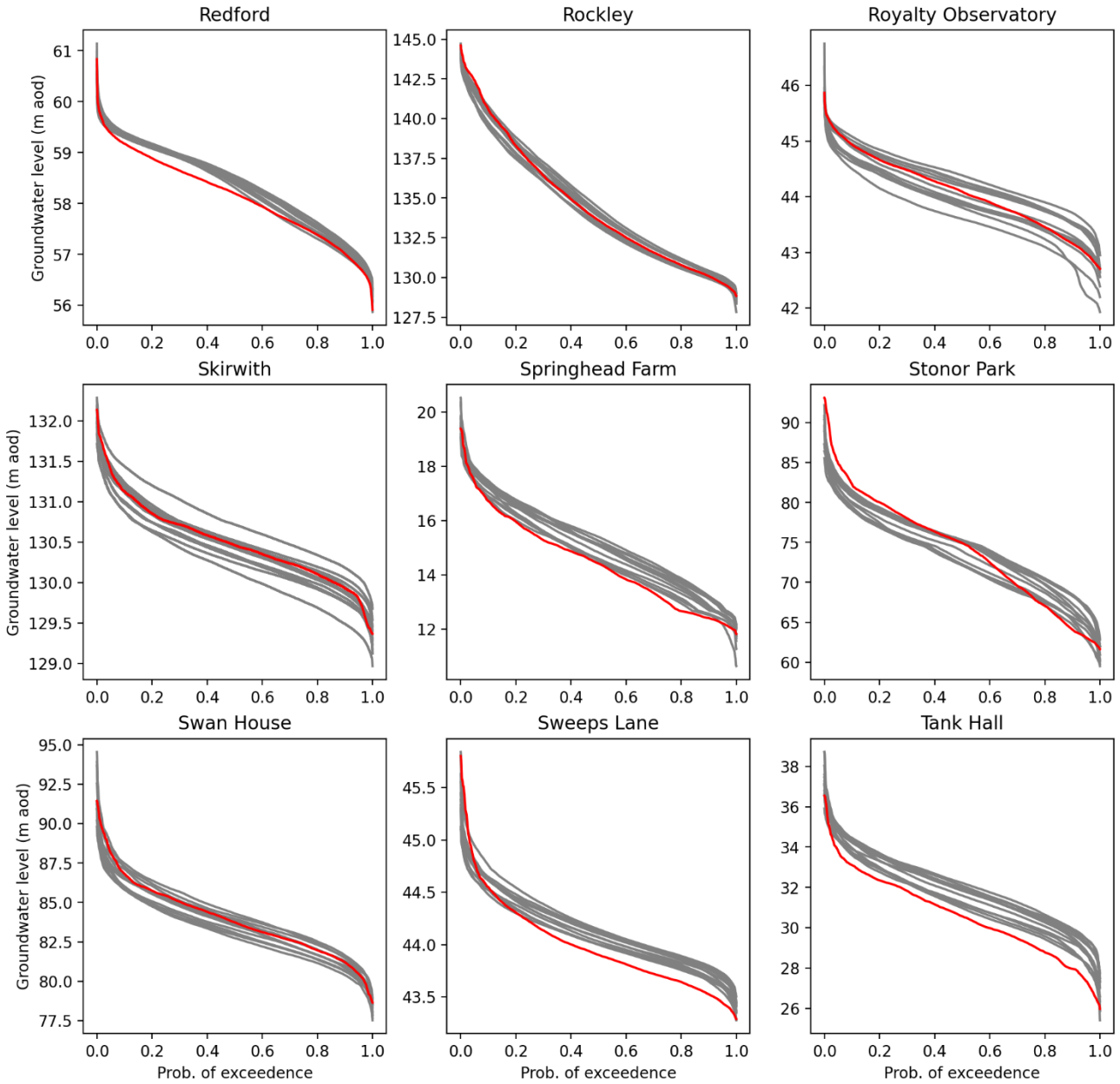


Figure S15 – Groundwater level duration curves (GLDCs) for the period 1989-2018 using the simrcm (grey lines) simobs (red line) simulations.

Supplementary Information 6: Low River Flows and Groundwater Levels

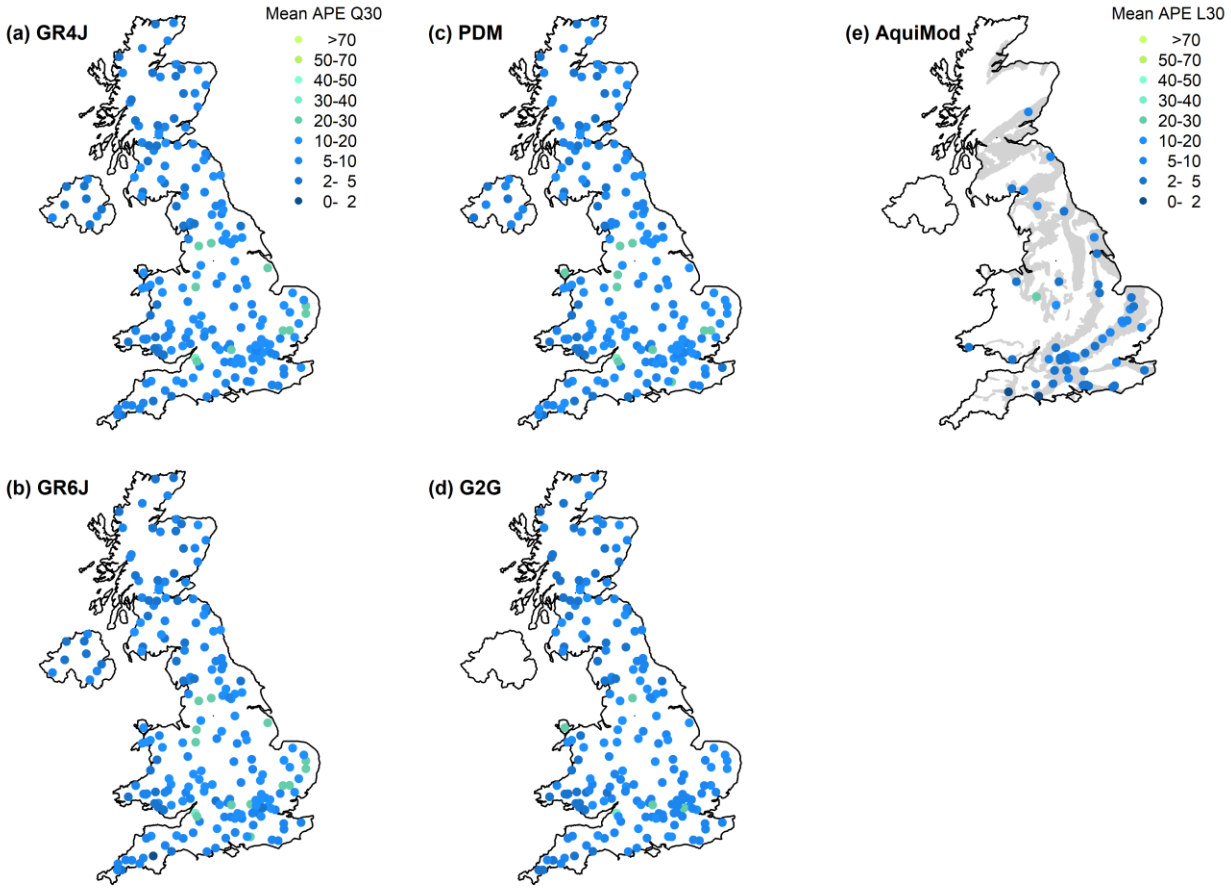


Figure S16 -- Comparison of river flows and groundwater levels exceeded 30% of the time (Q30) in model observations and RCM ensemble baseline, 1989-2018. Colour scale indicates the mean of 12 absolute percent errors (APEs) between Q30 in model observations and Q30 in each of 12 RCM ensemble members. Results are presented for each of the four hydrological models and one borehole model: (a) GR4J; (b) GR6J; (c) PDM; (d) G2G; (e) AquiMod. Note: AquiMod levels expressed relative to the minimum level prior to calculating APEs, to remove influence of arbitrarily high datums.

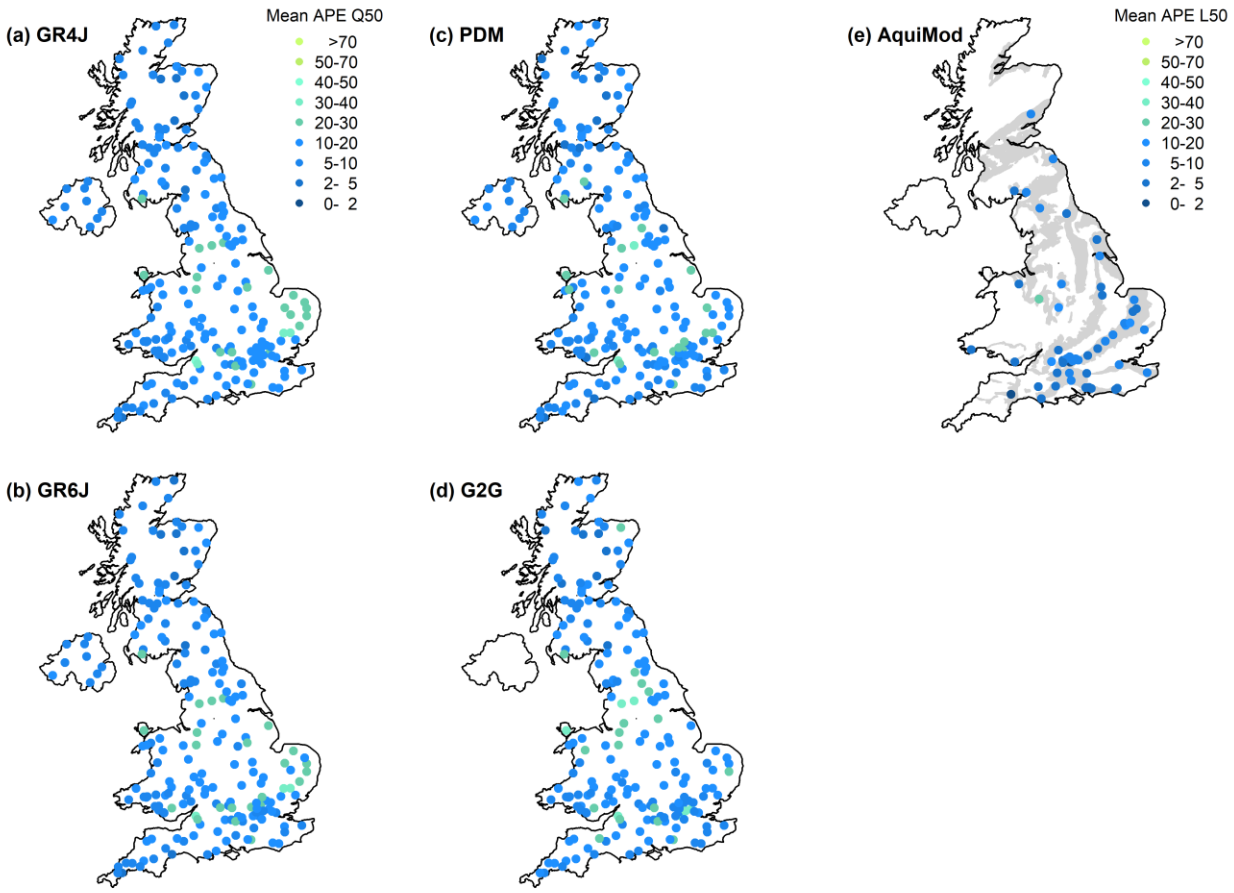
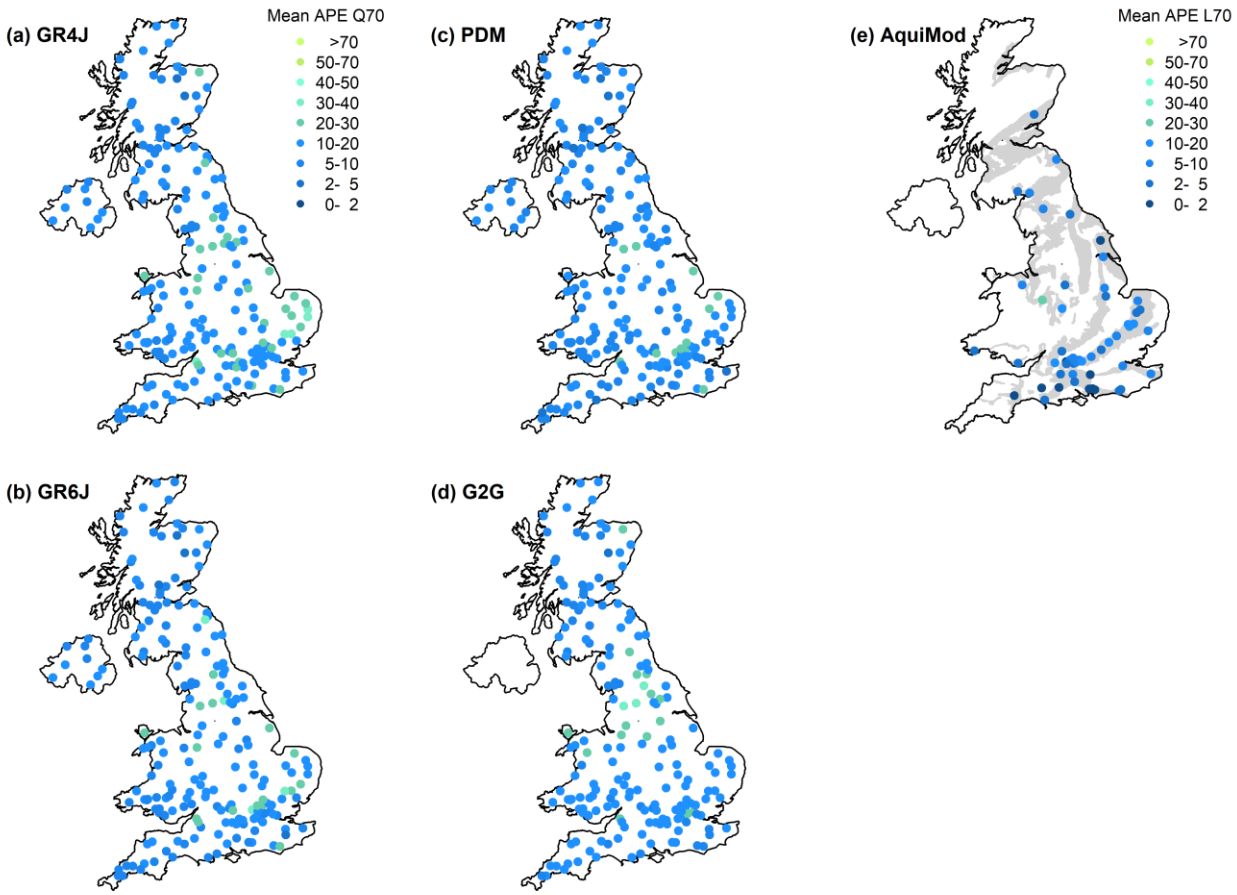


Figure S17 -- Comparison of river flows and groundwater levels exceeded 50% of the time (Q50) in model observations and RCM ensemble baseline, 1989-2018. Colour scale indicates the mean of 12 absolute percent errors (APEs) between Q50 in model observations and Q50 in each of 12 RCM ensemble members. Results are presented for each of the four hydrological models and one borehole model: (a) GR4J; (b) GR6J; (c) PDM; (d) G2G; (e) AquiMod. Note: AquiMod levels expressed relative to the minimum level prior to calculating APEs, to remove influence of arbitrarily high datums.



253

254 *Figure S18 -- Comparison of river flows and groundwater levels exceeded 70% of the time (Q70) in*
255 *model observations and RCM ensemble baseline, 1989-2018. Colour scale indicates the mean of 12*
256 *absolute percent errors (APEs) between Q70 in model observations and Q70 in each of 12 RCM*
257 *ensemble members. Results are presented for each of the four hydrological models and one borehole*
258 *model: (a) GR4J; (b) GR6J; (c) PDM; (d) G2G; (e) AquiMod. Note: AquiMod levels expressed relative*
259 *to the minimum level prior to calculating APEs, to remove influence of arbitrarily high datums.*

260

261