Supplements S1 – Explanation of the uncertainty of the modelling components

The three components (climate modelling, hydrological modelling and the Flood Frequency Analysis) introduce uncertainty. The contribution of the sources is estimated by calculating the standard deviation for each component and averaging it over the other components.

Climate: We calculate the standard deviation across the 12 CMIP6 members for each of the 64 hydrological models and take the mean of these 64 values.

Hydrology: We calculate the standard deviation across the 64 hydrological models for each of the 12 CMIP6 members and take the mean of these 12 values.

Flood frequency analysis: We calculate the standard deviation of the 1,000 flood frequency curves for each combination of climate and hydrological models (12x64) and calculate the mean of these 768 values.

The mean prediction range provides information about the variability in each component's predictions. However, it is worth noting that the ranges overlap between the components and that the total ensemble's standard deviation is smaller than the sum of the component-wise standard deviations.

Supplements S2 – Comparing Flood discharges and magnitudes

The output of the hydrological model is flood discharge, while the Flood Frequency Analysis produces flood magnitudes. We compare the flood discharge and flood magnitude of events with similar magnitudes. We use the percentile of the ranked flood discharge (we call it the 'flood percentile') to determine the frequency of the flood discharge. The flood percentile (FP_i) describes the flood discharge of the *i*-th percentile where *i* is:

$$i = \frac{rank}{record\ length} * 100,$$

where the rank is the position of the flood discharge ordered from small to large. For example, the FP_{100} is the discharge of the highest flood event in the 40-year record. The flood percentiles provide a more meaningful measure to aggregate ensembles than a temporal aggregation. A temporal aggregation does not account for the internal climate variability, which would lead to the aggregation of smaller and larger events (i.e. the CMIP6 members predict large events in different years). Furthermore, the members of the hydrological ensemble may react differently to the climate input, and this information would be lost in the temporal aggregation. Instead, it is more meaningful to aggregate the flood predictions by frequency, which enables a meaningful comparison of flood discharges and flood magnitudes.

For example, the $FP_{92.5}$ is the 4th highest flood prediction in the 40-year record ensemble member, and hence, is predicted statistically once every ten years and can be compared with the flood magnitude of the 1-in-10-year (10%-AEP) event. The $FP_{97.5}$ is the second-highest flood prediction and is hence predicted statistically once every twenty years, and can be compared with the magnitude of the 1-in-20-year (5%-AEP) event.

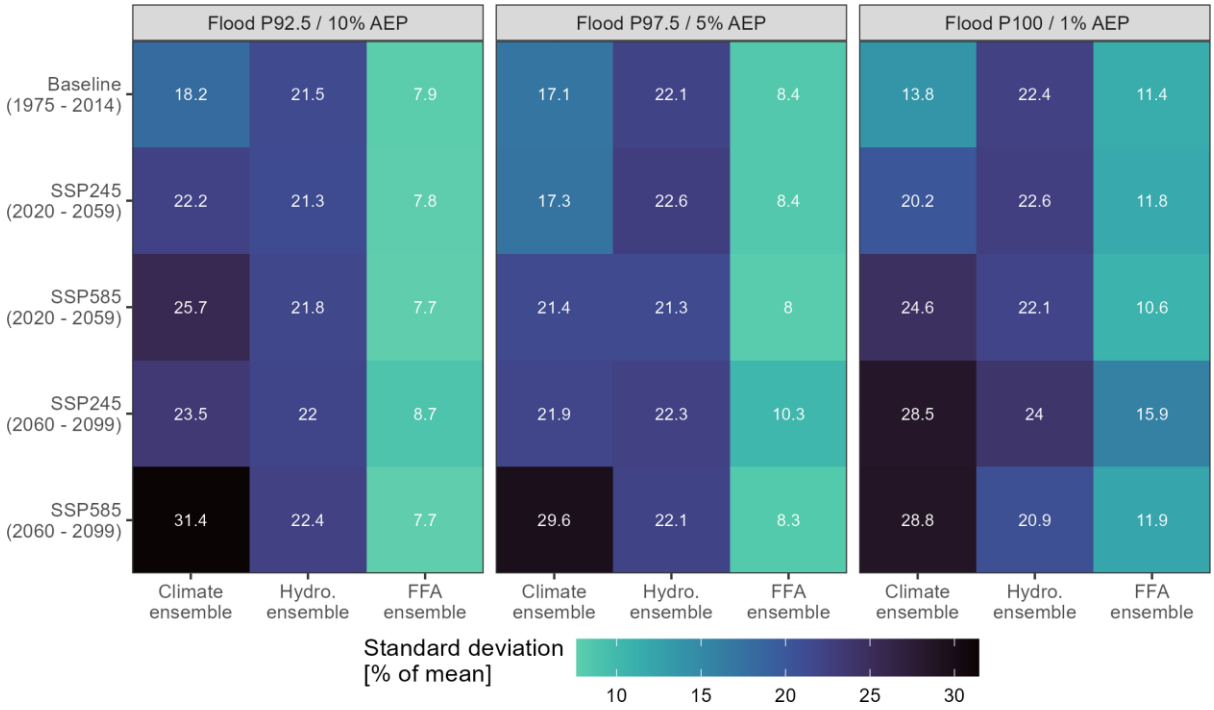

Figure 1: The mean standard deviation of the modelling components. See S1 and S2 for details about the calculation and the frequencies.