Supplementary Material

How to use the impossible map – Considerations for a rigorous exploration of Digital Twins of the Earth

Supplementary Material A: Literature research

The records of the numbers of factors shown in Figure 2 of the main text are a combination of two existing studies ((Song et al., 2015) and (Vanrolleghem et al., 2015)) and additional records compiled with a Web of Science query shown in Fig. S1. The additional screened papers were scanned using the software ASReview (van de Schoot et al., 2021). ASReview uses a machine learning approach to increase the likelihood of including relevant studies and decreasing the number of studies that need to be analyzed.

Figure 2 does not show two data points (outside of the axis maximum of 120), which reported thousands of factors. With these two studies, it is unclear what number was reported. Some studies report a factor as one parameter that was changed for the whole model domain, even if thousands of model cells were changed (if it is a gridded model), and others may report that as thousands of factors even if that did not increase the complexity of the experiment. We also excluded these two data points from the shown regression line. But even if these studies were complex experiments with thousands of factors, they do not change the overall picture that most studies do not investigate such a high number.

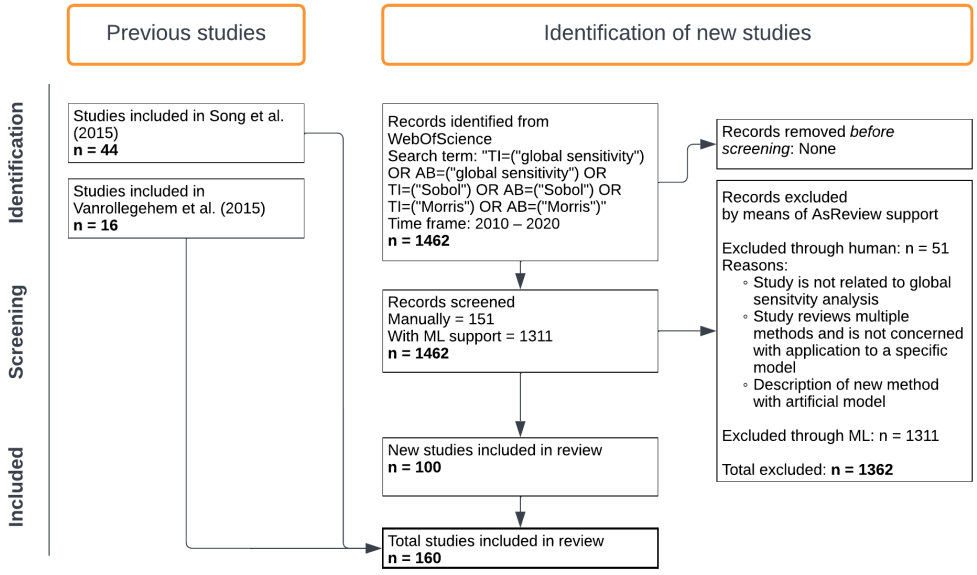

Figure S1: Visual PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) (Page et al., 2021) representation of studies used in Figure 2 of the main text.

Supplementary Material B: Automatic exploration of functional relationships

The code used to generate Figure 4 (see also code availability) searches recursively for the best possible split within the dataset. For each possible split, it determines which binary separation of an explanatory variable (e.g., air temperature above or below a certain threshold) would increase the correlation between an explanatory variable (e.g., annual mean precipitation or aridity index) and the variable under investigation (groundwater recharge) in at least one side (left or right) of the split. Possible splits are searched for based on

equally sized bins to reduce the search space into manageable pieces (though correlations are always calculated on the original data, not the bins). The algorithm tests all possible splits based on different threshold values, from small to large. To avoid selecting very small subspaces, a split requires each subspace to have at least 500 data points or 5% of the data of the parent node. Once the split has been defined, the functional relationship is determined by dividing the data in each leaf node into ten equally-sized bins and creating a line that connects the medians across the bins.

The model data originates from Gnann et al. (2023), who used 30-year averages of model simulations from the ISIMIP (Inter-Sectoral Impact Model Intercomparison Project; https://www.isimip.org) project. In addition, we use a set of explanatory variables that we assume to be potentially relevant in determining groundwater recharge in the model. We use long-term mean precipitation (P), long-term mean potential evapotranspiration (PET), an aridity index (AI) defined by PET/P, long-term mean temperature (T), an indicator of cold days per year (DB), and a land cover data set GlobCover (http://due.esrin.esa.int/page_globcover.php; aggregated to 0.5° with area-weighted Mode), which is closest to the information used in the model. All variables except land cover are 30-year averages from ISIMIP calculated in Gnann et al. (2023). In contrast to common forcing, the hydrological models used consider very different geological information, which is, therefore, hard to consider here.