The protocol is designed to test whether EC-Earth3 gives replicable results under two computing environments, named “A” and “B” hereinafter.

1. ***Deterministic check.*** This initial step is performed to confirm that EC-Earth is fully deterministic. Two 1-year integrations are conducted consecutively under the same computing environment (same executable, same machine, same domain decomposition). The results are required to be bit-for-bit identical.
2. ***Accounting for internal variability.*** Ensemble simulations are conducted so that the imprint of internal variability on climate indices like time means can be gauged. In an attempt to find the right balance between statistical power and limited computational resources, we run a five-member, 20-year integration for both “A” and “B” computing environments. In the subsequent steps, these simulations are referred to as “ensembles”.
3. ***Generation of simulations***. The five members of ensembles A and B always start from unique atmospheric and sea ice restarts, obtained from a long equilibrium simulation conducted on one computer. An oceanic restart is also obtained from this equilibrium simulation, and five random but deterministic perturbations are added to the sea surface temperature of this restart (Gaussian perturbation, standard deviation: 10^{-4} K). The introduction of these tiny perturbations allows ensemble spread to develop in ensembles A and B. Note that by the deterministic nature of the perturbations, pairs of members always start from the same triplet of atmospheric, oceanic and sea ice restarts: the first member of ensemble A and the first member of ensemble B start from identical initial conditions, and so on. Ensembles A and B are conducted under an annually repeating pre-industrial constant forcing. The ensembles start in 1850 and extend to 1869. The method of perturbation does not allow sampling the entire distribution of climatic states, as simulations are only 20 years long. If differences in computing environments cause differences in model output that are further enhanced by differences in climatic states, then there is a chance that our approach misses the detection of cases of replicability. However, differences in surface properties like SSS and SST develop rather rapidly in the ensemble, with inter-model spread comparable to interannual variability well before 20 years.
4. Calculation of standard indices. Due to the large amount of output produced by each simulation, the outputs from ensembles A and B are first post-processed in an identical way. Based on the list of standard metrics proposed by Reichler and Kim (2008), we record for each ensemble standard ocean, atmosphere, and sea ice parameters: 3D air temperature, humidity, and components of the wind; 2D total precipitation, mean sea level pressure, air surface temperature, wind stress, and surface thermal radiation; 2D sea surface temperature and salinity; and sea ice concentration. These fields are averaged monthly (240 time steps over 20 years) .
5. ***Calculation of standard metrics.*** Then, the model fields are compared to the same reference datasets as those used in Reichler and Kim (2008), which consist of observational and reanalysis datasets. For each field, a grid cell area-weighted average of the model departure from the corresponding reference is evaluated and then normalized by the variance of that field in the reference dataset. Thus, for each field, one number is retained that describes the mismatch between that field in the ensemble and the reference. Five such numbers are available for each ensemble, since each ensemble uses five ensemble members.
6. ***Statistical testing.*** For each index we compare two five-member ensembles and determine whether the two ensembles are statistically indistinguishable from one another. Since no prior assumption can be made on the underlying statistical distribution of the samples, we use a two-sample Kolmogorov-Smirnov (KS) test. KS tests are non-parametric, which makes them suitable for our application. A Monte Carlo analysis reveals that for a prescribed level of significance of 5%, the power of the two-sample KS test exceeds 80% (a standard in research) when the means from the two samples are separated by at least 2 standard deviations (Figure 2). Stated otherwise, there is a non-negligible probability (>20%) that small but actual differences (less than 2 standard deviations) are not detected by the test. In this test, the null hypothesis is that the two samples are statistically indistinguishable from each other with a confidence of 95%. This means that under this null hypothesis of no difference, significant differences are expected to occur 5 % of the time. Because the tests are done on global metrics, there is a possibility of false negatives when the same performance is obtained from spatially varying biases. However, the use of multiple variables for the assessment of maps aims at minimizing this issue.