

# 1 Causal networks for climate model evaluation and 2 constrained projections

3 Peer Nowack<sup>1,2,3,4\*</sup>, Jakob Runge<sup>5,1</sup>, Veronika Eyring<sup>6,7</sup>, Joanna D. Haigh<sup>1,2</sup>

4 <sup>1</sup>Grantham Institute, Imperial College London, London, SW7 2AZ, UK.

5 <sup>2</sup>Department of Physics, Faculty of Natural Sciences, Imperial College London, London, SW7 2AZ, UK.

6 <sup>3</sup>Data Science Institute, Imperial College London, London, SW7 2AZ, UK.

7 <sup>4</sup>School of Environmental Sciences, University of East Anglia, Norwich, NR4 7TJ, UK.

8 <sup>5</sup>Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institute of Data Science, Jena, 07745, Germany.

9 <sup>6</sup>Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institute of Atmospheric Physics, Oberpfaffenhofen,  
10 82234, Germany.

11 <sup>7</sup>University of Bremen, Institute of Environmental Physics, Bremen, 28359, Germany.

12 Corresponding author: Peer Nowack (p.nowack@uea.ac.uk)

13 *Keywords:* Climate models, Earth observations, atmospheric dynamics, causal discovery, network  
14 algorithms, model evaluation and intercomparison, machine learning, CMIP5, climate change,  
15 precipitation patterns.

## 16 **Abstract**

17 **Global climate models are central tools for understanding past and future climate change.**  
18 **The assessment of model skill, in turn, can benefit from modern data science approaches.**  
19 **Here we apply causal discovery algorithms to sea level pressure data from a large set of**  
20 **climate model simulations and, as a proxy for observations, meteorological reanalyses. We**  
21 **demonstrate how the resulting causal networks (fingerprints) offer an objective pathway for**  
22 **process-oriented model evaluation. Models with fingerprints closer to observations better**  
23 **reproduce important precipitation patterns over highly populated areas such as the Indian**  
24 **subcontinent, Africa, East Asia, Europe and North America. We further identify expected**  
25 **model-interdependencies due to shared development backgrounds. Finally, our network**  
26 **metrics provide stronger relationships for constraining precipitation projections under**  
27 **climate change as compared to traditional evaluation metrics for storm tracks or precipitation**  
28 **itself. Such emergent relationships highlight the potential of causal networks to constrain**  
29 **longstanding uncertainties in climate change projections.**

## 30 **Introduction**

31 State-of-the-art climate and Earth system models represent an enormous scientific achievement and  
32 are central tools to understand past climates as well as to project future climate change. More than  
33 forty modelling centres worldwide undertake climate model development<sup>1-3</sup> and have rapidly  
34 elevated their level of sophistication. Nowadays, many models simulate not only fundamental  
35 physical laws of fluid motion, energy and momentum conservation but also include interactive carbon  
36 cycle, aerosol and atmospheric chemistry schemes, or resolve the entire stratosphere<sup>4-10</sup>. However,  
37 while all climate models are based on the same physical principles, there are development-specific  
38 choices that lead to significant model differences, in particular related to subgrid-scale  
39 parameterizations of clouds, convection and aerosols<sup>11-13</sup>. These contribute to persistent  
40 discrepancies between models and observations as well as among model projections, for example  
41 regarding precipitation changes<sup>1,14,15</sup>. Multi-model evaluation and intercomparison is often based on  
42 the mean and variance of aggregate quantities such as temperature, or spectral properties and  
43 (auto-)correlation measures<sup>16-18</sup>. One issue with such metrics is that models can be right for the  
44 wrong reasons due to offsetting biases<sup>11,12,16</sup>.

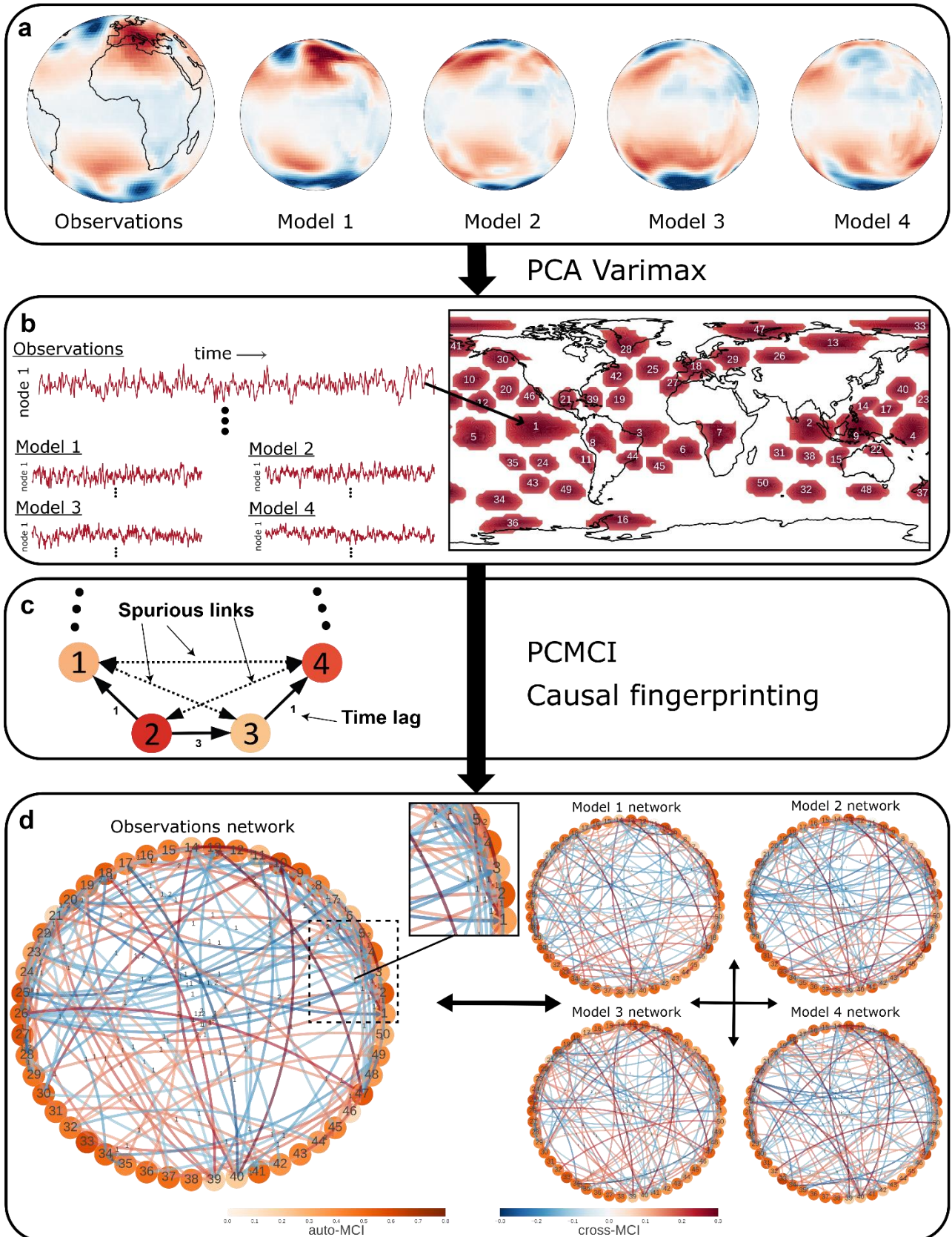
45 Here we introduce causal model evaluation (CME) as a type of process-oriented model  
46 evaluation<sup>11,18-20</sup>. CME deploys recently developed causal discovery methods<sup>21-23</sup> adapted for  
47 applications to climate data<sup>23-27</sup>. Within the CME framework, we evaluate the ability of models from  
48 the Coupled Model Intercomparison Project Phase 5 (CMIP5) to simulate atmospheric dynamical  
49 interactions classically measured as lagged correlations between climate variables at remote  
50 locations<sup>28-31</sup>. Causal discovery algorithms go beyond correlation-based measures by systematically  
51 excluding common driver effects and indirect links<sup>22,26,32,33</sup>. We show that characteristic causal  
52 fingerprints can be learned from climate datasets, which are robust among ensemble members of  
53 the same model and, for example, can identify shared model development backgrounds. Fingerprints  
54 closer to observations are also associated with smaller precipitation biases in climate models.  
55 Finally, we highlight the potential of our approach to offer a pathway to reducing uncertainties in  
56 climate change projections, as well as to understand differences between models and observations.

## 57 **Results**

58 **Causal model evaluation framework.** To characterize the network of global dynamical interactions,  
59 we use a causal discovery algorithm to reconstruct directed, time-lagged interdependency networks  
60 from global climate datasets. Figure 1 provides an overview of the individual steps of the CME  
61 framework (see Methods for details).

62 The selection of components defining the network nodes will typically be guided by expert  
63 knowledge in conjunction with dimension reduction techniques. Here we use components obtained  
64 through Varimax-rotated principal component analysis<sup>34,35</sup> (PCA) applied to sea level pressure  
65 anomaly data (Figure 1a; Methods). For sea level pressure data, PCA-Varimax components can be  
66 interpreted as major modes of climate variability<sup>25,28,36,37</sup>. Due to the seasonal character of interaction  
67 pathways<sup>28,38</sup>, we construct individual components, and in the next step networks, for the four  
68 meteorological seasons: December, January, February (DJF); March, April, May (MAM); June, July,  
69 August (JJA); September, October, November (SON). We select fifty components for each season  
70 (Methods) whose geographic locations for DJF are indicated in Figure 1b (for all seasons see  
71 Supplementary Fig. 1). PCA-Varimax can identify the major modes of variability<sup>37</sup>, for example  
72 related to the El Niño Southern Oscillation (ENSO) in the East, West and Central Pacific<sup>39</sup>  
73 (components 1,4,5 in Figure 1b).

74 We calculate interactions among these nodes as causal networks from the associated  
75 component time series (Figure 1b). For this step, we use the PCMCI algorithm by Runge et al.<sup>23,26</sup>,  
76 which is particularly suited for high-dimensional and auto-correlated climate data (Methods). In  
77 contrast to pure correlation measures, causal discovery methods are built to remove spurious links  
78 due to common drivers and indirect pathways from the networks (Figure 1c)<sup>22,26</sup>. The resulting  
79 networks contain information on the direction and associated time lags of potential causal links,  
80 characterizing the pathways of the global interaction network. PCMCI has been tested extensively  
81 to successfully recover important interactions in the climate system such as the tropical Walker  
82 circulation and predictors of polar vortex states<sup>23,24,26,27</sup>. Note that, in these network structures, some  
83 established interactions measured traditionally as direct correlations between climate modes can  
84 follow a more complex pathway of indirect links. We illustrate this for the coupling between ENSO  
85 and the Pacific-South American (PSA) pattern<sup>29,40</sup> in Supplementary Figure 2.



86  
87  
88 **Fig. 1 | Sketch of the causal model evaluation framework.** **a**, Gridded Earth system data, here daily-mean sea level pressure from the  
89 NCAR-NCEP reanalysis (approximating observations)<sup>41</sup>, is dimension-reduced using PCA-Varimax to **b**, a set of regionally confined  
90 climate modes of variability. The same transformation is subsequently applied to climate model data (Methods). Core component regions  
91 (in this case for the season December-January-February) are indicated in red. Each component is associated with a time series and  
92 serves as one of the network nodes. Here, the component time series are afterwards 3-day-averaged. **c**, PCMCI estimates directed lagged  
93 links among these nodes giving rise to **d**, dataset-characteristic causal fingerprints, which can be used for model evaluation and  
94 intercomparison. Node colours in **d** indicate the level of autocorrelation (auto-MCI) as the self-links of each component and link colours  
95 the interdependency strength (cross-MCI). Link-associated time lags (unit=3 days) are indicated by small labels. Only the around two  
96 hundred most significant links for the reanalysis and for data from four <sup>2</sup>climate models are shown. Links with lag zero, for which directions  
cannot be easily causally resolved, are not shown.

97 The resulting causal networks effectively represent characteristic causal fingerprints<sup>42,43</sup> for  
98 each sea level pressure dataset (Figure 1d), which can be compared using network metrics<sup>25</sup>. Each  
99 network consists of hundreds of links. Generally, we conduct pair-wise comparisons of all possible  
100 links in a network A to a network B, taking A as the reference network. For example, we test if a link  
101 from component 4 (West Pacific ENSO) to component 1 (East Pacific ENSO) found in observations  
102 is also detected in climate model datasets. We use a modified asymmetric  $F_1$ -score (Methods) as  
103 the harmonic mean of precision (fraction of links in B that also occur in A) and recall (fraction of links  
104 in A that are detected in B).  $F_1$ -scores vary between 0 and 1 (perfect network match). The network  
105 comparison results depend on the number of links considered to be statistically significant (Methods).  
106 However, we tested that all conclusions based on the 400-500 most significant links per network  
107 included here are robust to a large range of possible network link densities from a hundred to more  
108 than a thousand links (Supplementary Figs. 3-6; Supplementary Table 1).

109 **Application to pre-industrial simulations.** Pre-industrial simulations are well suited for the CME  
110 of atmospheric dynamical interactions due to the many years simulated by each model in the  
111 absence of transient effects caused by anthropogenic forcings<sup>1-3</sup>. Specifically, we applied the CME  
112 framework to 210 years of global DJF sea level pressure data from each of in total twenty CMIP5  
113 models at a 3-day time resolution (Methods; Figure 2). In our algorithm settings, we include  
114 interactions on a time-scale of up to 30 days ( $\tau_{\max}=10$ ; Methods). We split each 210-year dataset into  
115 three 70-year intervals (ensemble members) to study multi-decadal variations<sup>44,45</sup>. As a result, we  
116 obtain nine possible network comparisons for each pair of models and six distinct comparisons  
117 between ensemble members of the same model.  $F_1$ -scores for these model intercomparisons are  
118 shown in Figure 2. Three major features highlight the skill of the CME framework.

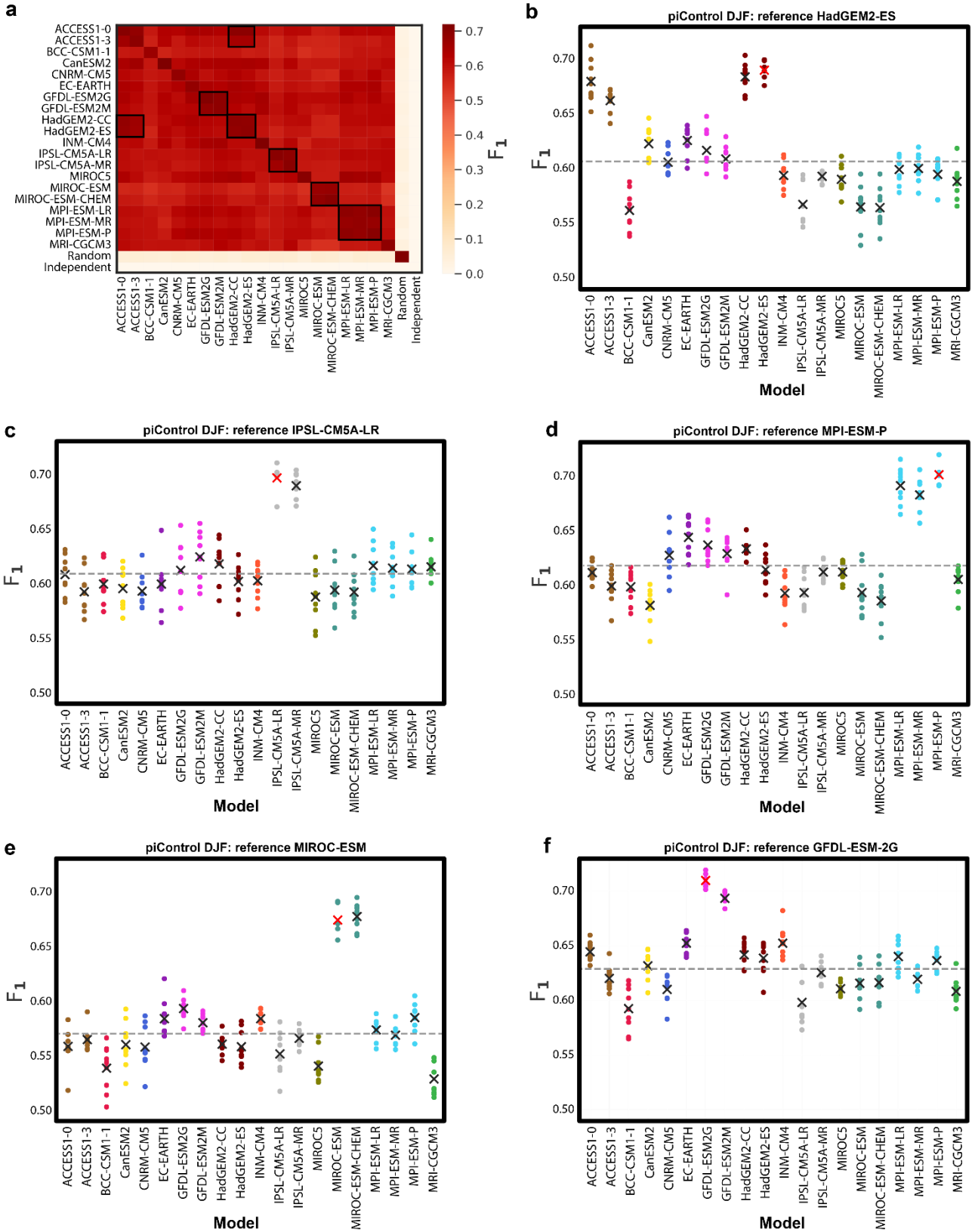
119 Firstly, each model can be recognized individually purely based on its causal fingerprint.  
120 Networks estimated from different ensemble members of the same model are more consistent than  
121 networks estimated from two different models as evident from the high  $F_1$ -scores on the diagonal of  
122 the matrix in Figure 2a (dark red). Each row in Figure 2a denotes the model used as the reference  
123 against which each column is compared.

124 Secondly, models with shared development background can be detected. Many climate  
125 models share software, resulting in important interdependencies among them<sup>12,46-50</sup>. CME can detect

126 such shared backgrounds (highlighted by black squares in Figure 2a). For example, CME identifies  
127 the models HadGEM2-ES, HadGEM2-CC, ACCESS1-0 and ACCESS1-3 as similar, which are all  
128 versions of the HadGEM model family<sup>51,52</sup> developed by the UK Met Office. There is a clear  
129 separation between these four and the remaining models, see Figure 2b showing all scores when  
130 HadGEM2-ES networks are taken as the reference. The different models developed by the Institute  
131 Pierre Simon Laplace (IPSL), the Max-Planck Society (MPI) and the Geophysical Fluid Dynamics  
132 Laboratory (GFDL) are also each recognized as subgroups (Figures 2c-e). For the Japanese MIROC  
133 models, two out of three are detected as a subgroup (MIROC-ESM, MIROC-ESM-CHEM), whereas  
134 MIROC5 is even less similar than the multi-model average (gray line in Figure 2f). We conclude that  
135 CME can detect similar models, a condition often but, as shown here, not always synonymous with  
136 models developed under the same research umbrella. This demonstrates the significant potential of  
137 using CME to assess model interdependencies based on causal networks.

138 Thirdly, climate models are recognized to share a physical ground truth. We further compared  
139 all twenty models with two artificial reference cases: Random and Independent (last two  
140 rows/columns in Figure 2a; Methods). For Random, we created fifty randomly coupled and auto-  
141 correlated noise time series, i.e. there are links in the system, but these do not follow any Earth  
142 system physics. As evident from Figure 2a, the corresponding networks are self-consistent (diagonal  
143 entry) but achieve very low  $F_T$ -scores when compared to the actual climate models. For Independent,  
144 we created auto-correlated time series without any significant coupling among them so that any  
145 detected links occur randomly in the system (false positives). CME expectedly finds low scores  
146 throughout for this case.

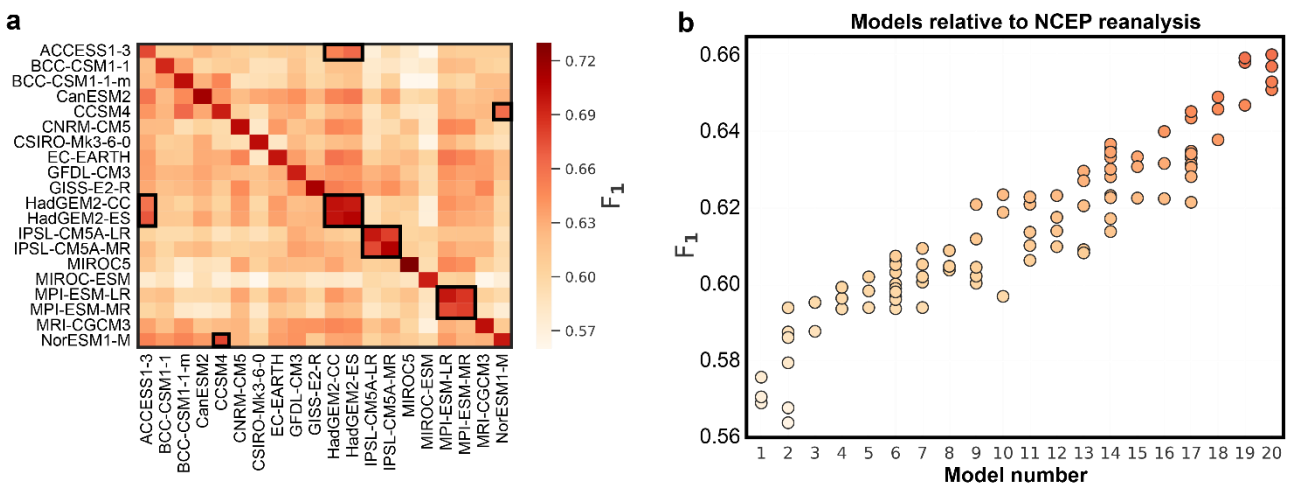
147 **Causal model evaluation of historical simulations.** Motivated by CME's skill to recognize models  
148 with shared development background, we next evaluate the CMIP5 models with NCAR-NCEP  
149 reanalysis data<sup>41</sup> as a proxy for recent observations. We calculate fingerprints from twenty CMIP5  
150 simulations covering approximately the historical period from 1st January 1948 to 31st December  
151 2017 (Methods). For better statistical estimates, we only included models for which at least three  
152 ensemble members were available (Supplementary Table 2). To additionally investigate the role of  
153 seasonal variability, we carried out separate analyses for DJF, MAM, JJA and SON. However, all



154  
 155 **Fig. 2 | Pre-industrial network intercomparison scores.** **a**, Matrix of average  $F_1$ -scores for pair-wise network comparisons between  
 156 ensemble members of twenty climate models (labelled following CMIP5 nomenclature in capital letters) using data for December-January-  
 157 February (DJF) and two surrogate models (Random, Independent). Rows are reference models, columns are the models which are  
 158 compared to these references. Higher scores imply better agreement between networks, i.e. that two models are more similar in terms of  
 159 their causal fingerprint. **b-f**, Scatter plots showing each individual network comparison score, with different models taken as reference (as  
 160 labelled in the sub-figure titles) that the other models (labelled on the x-axis using capital letters) are compared to. Black crosses (red for  
 161 the reference) mark average results also shown in **a**. Gray dashed lines mark the average score excluding the reference itself. Our causal  
 162 model evaluation approach detects the expected similarities between certain model groups as shown in **b-f**, which are additionally  
 163 indicated by inset black squares in **a**. Source data are provided as a Source Data file.

164 seasons yielded very similar results (Supplementary Figs. 3-6) and we focus the discussion on  
 165 annual  $F_1$ -scores averaged over all seasons (Methods).

166 We find effectively the same model subgroups as before (inset boxes in Figure 3a). Due to  
 167 the slightly different setup, there is an additional subgroup related to the climate model CCSM4  
 168 (Supplementary Fig. 7). Taking the NCAR-NCEP reanalysis network as the reference, we obtain an  
 169 estimate of how well individual models capture the observed causal fingerprint (Figure 3b; the  
 170 models are ordered by average  $F_1$ -score). The result is a continuum rather than a clear-cut  
 171 differentiation between a better and a worse group of models. However, models do exhibit  
 172 significantly different causal fingerprints ( $p$ -value<sup>53</sup>  $< 10^{-9}$ ). We conducted the same analysis using a  
 173 shorter ERA-Interim reanalysis dataset<sup>54</sup> to estimate the reference network and obtained almost the  
 174 same model order (Supplementary Fig. 8, Supplementary Table 1).



175 **Fig. 3 | Historical network comparisons.** **a**, As Figure 2a, but for climate model simulations spanning approximately the historical period  
 176 from 1st January 1948 to 31st December 2017 for which twenty CMIP5 models with up to ten different ensemble members are available.  
 177 **b**, Ordered  $F_1$ -scores when the causal fingerprint learned from NCAR-NCEP reanalysis data is taken as the reference. Differences in **b**  
 178 are highly statistically significant, with  $p$ -values  $< 9 \times 10^{-10}$  for a non-parametric Kruskal-Wallis-test and  $p < 5 \times 10^{-30}$  for a standard one-way  
 179 ANOVA F-Test. The model key for **b** is provided in Supplementary Table 1. We note that similar model rankings have been found regionally  
 180 for precipitation, e.g. for China<sup>55</sup>. Individual network scores (marker colours) in **b** follow the colour code from **a**. Source data are provided  
 181 as a Source Data file.  
 182

183 **Implications for precipitation modelling.** Atmospheric dynamical interactions as imprinted here  
 184 on the sea level pressure field are well-known drivers of precipitation anomalies in many world  
 185 regions<sup>28,29</sup>. Therefore, we test for relationships between the reanalysis-referenced  $F_1$ -scores of  
 186 CME and Taylor  $S$ -scores<sup>55,56</sup> for precipitation rates, which measure grid-cell-wise errors in  
 187 conjunction with overall discrepancies in precipitation variability across a spatial domain. To calculate  
 188 the  $S$ -scores, which also range from 0 to 1, we use historical Climatic Research Unit (CRU)<sup>57</sup> land



189 surface precipitation data from the University of East Anglia, averaged over the years 1948-2017  
190 (Methods).

191 We find that better fingerprints are associated with smaller land precipitation biases ( $F_I$ - and  
192 S-scores are positively correlated; Figure 4a). This is true globally (correlation coefficient  $R=0.7$ ) as  
193 well as in many world regions known to be influenced by (remote) dynamical interactions, in  
194 particular North America ( $R=0.7$ ), East Asia ( $R=0.6$ ), Africa ( $R=0.5$ ) and South Asia ( $R=0.5$ ). These  
195 results also hold if we disregard models belonging to the same subgroups as marked in Figure 3a.  
196 There are some regional exceptions (e.g. Australia, Indonesia) where we find no significant  
197 correlations. A possible explanation is predominant regional factors<sup>17,39</sup> rendering a global network  
198 metric less suitable. In addition, regional correlations are sometimes dependent on the number of  
199 links included in the networks. For example, we find generally higher (lower) correlations for  
200 Europe/North America (Africa) if weaker links are included (excluded), likely because tropical  
201 connections have on average stronger dependencies (Supplementary Figs. 9-13).

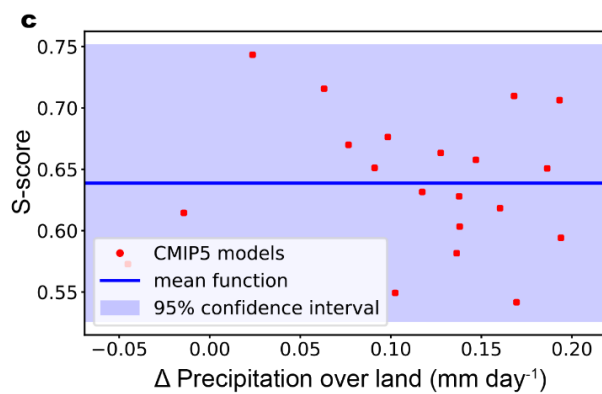
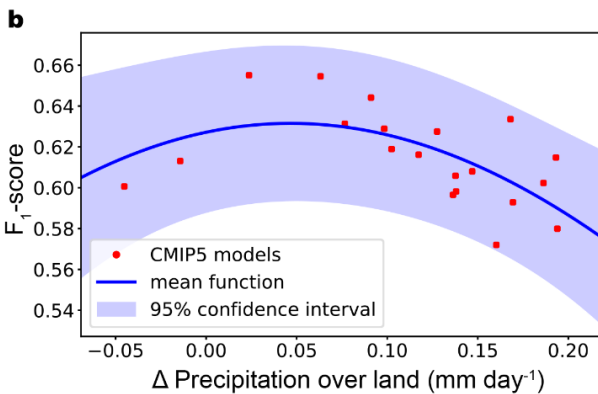
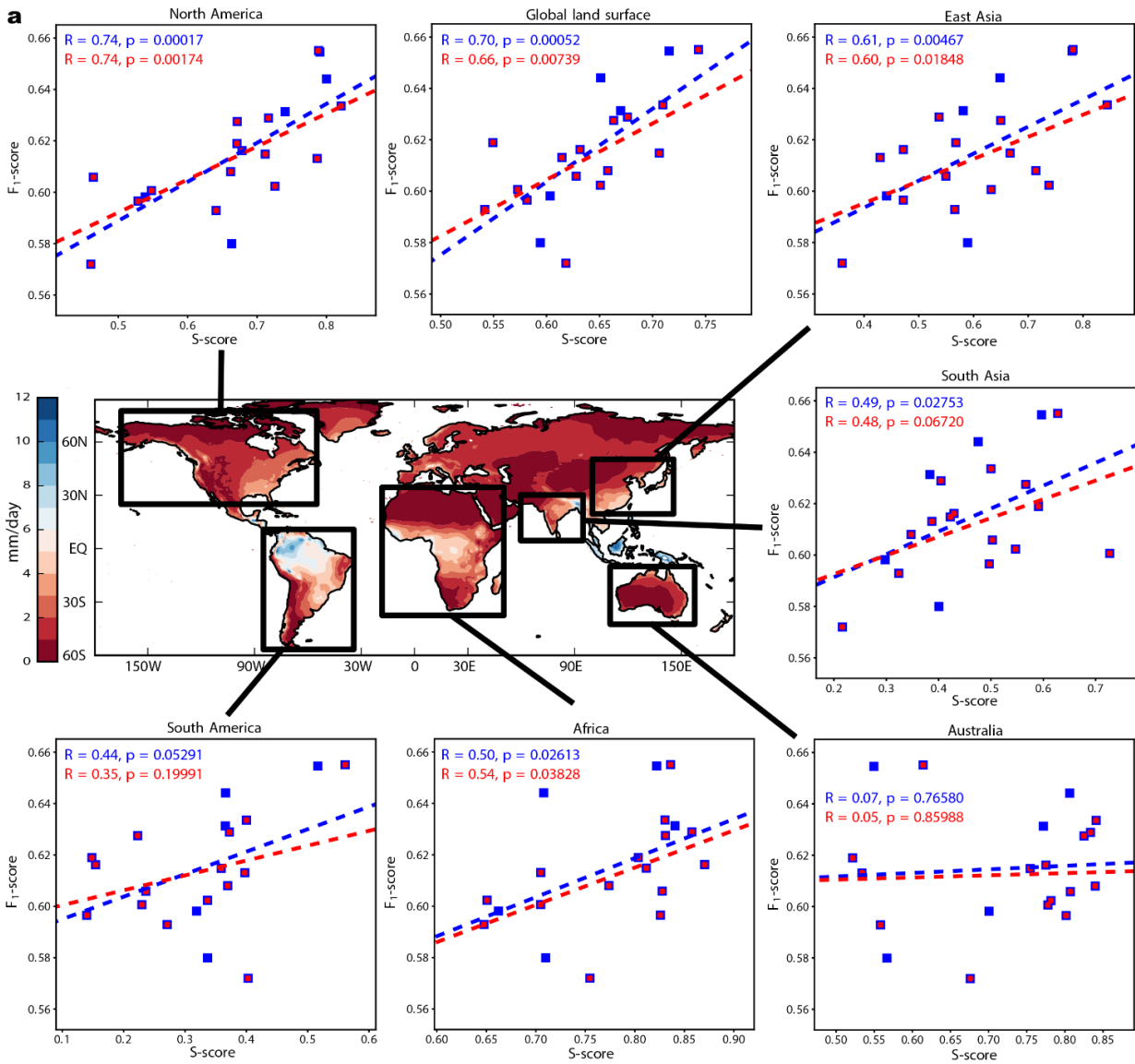
202 An interesting question is how to interpret the relationship between precipitation and the  
203 causal network skill scores from a physical point of view. Notably, the causal networks are, especially  
204 at stringent significance thresholds, dominated by interactions on a timescale of less than one week  
205 (lag  $\tau \leq 2$ ; Figure 1d). This timescale is broadly equivalent to dynamical interactions related to storm  
206 tracks<sup>58</sup>. Simple metrics have been used before to quantify the skill of climate models to capture  
207 storm tracks, e.g. pattern correlations in standard deviations of 2-6-days bandpass-filtered daily  
208 mean sea level pressure data<sup>59</sup>. Indeed, Taylor S-scores for precipitation are also positively  
209 correlated with such simpler metrics (Supplementary Figs. 18-20), which altogether indicates that a  
210 large part of the links in the causal networks represent dynamical interactions related to storm tracks.  
211 This result is in agreement with earlier work by Ebert-Uphoff and Deng<sup>32,33</sup> who constructed networks  
212 from DJF and JJA NCEP-NCAR reanalysis geopotential height data, as well as from equivalent data  
213 from a single climate model. In their network analyses, they also found storm tracks to be a key  
214 driver of network connectivity (see Methods for a comparison of our network methodologies).

215 Having highlighted the importance of storm tracks, we also point out that the simpler pattern  
216 correlation storm track metrics generally show smaller and less significant correlations with the  
217 precipitation S-scores on a global as well as on regional scales than our  $F_I$ -network scores. This

218 underlines that our causal networks identify additional relationships which further improve the  
219 correlations with precipitation. Longer time-scale dynamical interactions, for example triggered by  
220 the ENSO and its zonal couplings as well as its effects on the extratropics are prime candidates for  
221 explaining some of the higher skill related to our causal network scores.

222 Finally, we find strong indications that our causal metrics could aid in constraining uncertainty  
223 in precipitation projections under climate change. As mentioned above, past model skill in a quantity  
224 does not automatically imply skill for future projections as models can be right for the wrong reasons.  
225 The networks we use here infer rather complex dynamical coupling relationships from sea level  
226 pressure data that are effectively impossible to calibrate against current observations, different from,  
227 for example, quantities such as global surface temperature<sup>11</sup>. Causal discovery methods could thus  
228 provide more robust insights by identifying dynamical coupling mechanisms arising from underlying  
229 physical processes that are more likely to hold also under future climate change scenarios (see  
230 Discussion). It is therefore interesting to consider our complex causal information quantity in terms  
231 of constraining future precipitation projections. Indeed, we find no relationship between the past  
232 global precipitation skill  $S$ -scores and future precipitation rate changes in the CMIP5 projections, but  
233 there appears to be an approximately parabolic relationship between projected CMIP5 global land  
234 precipitation rate changes attained by the period 2050-2100 (relative to 1860-1910; Supplementary  
235 Fig. 16) and  $F_T$ -scores from historical runs (Figures 4b/c). This implies intermediate model range  
236 land precipitation changes of around 0.0-0.1 mm/day according to the causal fingerprint scores, as  
237 opposed to the most extreme negative and positive changes. We also note that simpler dynamical  
238 metrics, e.g. based on sea level pressure Taylor  $S$ -Scores, or the aforementioned storm track skill  
239 scores, and using the same non-parametric Gaussian Process regression (Figure 4b/c; Methods),  
240 do also not yield such emergent relationships (Figure 4b/c, Supplementary Figs. 17-20).

241 Any method resting on the assumption that past model skill in a certain metric can be related  
242 to projected future changes necessarily suffers from certain restrictions. Firstly, there could be  
243 processes that are not at all (or not well) represented in climate models today, which might become  
244 important in the future. However, this is true for any emergent relationship based on model evaluation  
245 against past observations. Secondly, not all relevant processes might be well-captured through the  
246 chosen metric. Our metric here is focused on dynamical processes (although it might, at least



247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263

**4 | Historical network scores and precipitation.** **a**, Centre map: Climatic Research Unit (CRU) annual mean precipitation rate climatology<sup>57</sup> in  $\text{mm day}^{-1}$ . Surrounding: linear correlations between the  $F_1$ -scores for the CMIP5 models (with the NCAR-NCEP reanalysis as the reference case) and regional precipitation bias scores (S-scores). Higher S-scores are equivalent to a better representation of annual mean precipitation in a given model. Correlations are shown for six world regions and for the global land surface (excluding Antarctica), as labelled. Blue denotes data for all models; red the case where five models from causally similar sub-groups are excluded (IPSL-CM5A-LR, ACCESS1-3, HadGEM2-CC, NorESM1-M, MPI-ESM-LR). **b** Relationship between  $F_1$ -scores and land precipitation changes projected by the CMIP5 models. The latter are calculated as the difference between the periods 1860-1910 and 2050-2100 under the RCP8.5 scenario. The relationship exhibits an approximately parabolic structure, as evident from a Gaussian Process fit to the data (log-marginal likelihood: 44.15; Methods). Past model precipitation skill as measured through the global S-score does not provide a strong relationship (**c**; log-marginal likelihood=26.35). This result is robust to the use of a different reanalysis, the number of links included in the network, and can also be demonstrated to be statistically significant in a direct parabolic fit (Supplementary Figs. 14, 15). This implies that precipitation rate changes (Supplementary Fig. 16) can be constrained using the  $F_1$ -scores (best estimate is around 0.0-0.1  $\text{mm day}^{-1}$ ), whereas past model skill for the same variable does not provide such a constraint; in line with previous demonstrations that past model biases in simple metrics are not necessarily indicative of future model projections<sup>12,62</sup>. Other simple dynamical metrics we tested generally provided lower correlation scores with historical precipitation modelling skill and also did not provide the same emergent relationship for future projections (Supplementary Figs. 17-19). Source data are provided as a Source Data file.

264 indirectly, capture the effects of some thermodynamical processes<sup>14,60</sup>), whereas, for example, future  
265 changes in soil moisture are probably primarily thermodynamically driven. Future changes in soil  
266 moisture, in turn, could regionally modulate future changes in land precipitation<sup>61</sup>. Finally, the  
267 possibilities for future projections are also constrained by the models participating in CMIP5.  
268 Therefore, we can only constrain the relationship within the given data boundaries, and it should be  
269 further verified across other scenarios and ensembles (such as CMIP6). Similar model evaluation  
270 exercises, also concerning variables other than precipitation and atmospheric dynamical  
271 interactions, could test for similar emergent relationships in the ever-expanding data made available  
272 through observations and climate modelling projects. Such studies might flexibly combine the  
273 blueprint of the method outlined here with other dimension reduction techniques and/or causal  
274 discovery algorithms<sup>32,33</sup>.

## 275 **Discussion**

276 We have highlighted causal model evaluation (CME) as a framework to evaluate state-of-the-art  
277 climate models. Based on data-driven causal fingerprints, CME is able to detect models with shared  
278 development backgrounds. By considering a large set of climate models simultaneously, we find that  
279 climate models with more realistic dynamical causal fingerprints also have smaller precipitation  
280 biases globally, and over highly populated areas such as North America, India and China. More  
281 realistic fingerprints appear to also have implications for projected future changes in land surface  
282 precipitation. Causal network analyses could therefore be a promising tool to constrain climate  
283 change projections. The underlying premise is that physical processes (e.g., convection, cloud  
284 formation, the large-scale circulation) lead to dynamical coupling mechanisms in Earth's  
285 atmosphere. CME aims at statistically representing these couplings in the form causal networks,  
286 which in turn are, as we show here, indicative of modelling skill in precipitation. It appears intuitive  
287 that modelling skill as captured through our causal fingerprint scores is therefore also relevant for  
288 modelling future changes in precipitation, at least so far as the physical processes relevant for  
289 present-day precipitation remain important in future climates.

290 Our work builds on several previous causal network studies in climate science, which were  
291 typically focused on network algorithm applications to individual climate modelling or reanalysis  
292 datasets, or on the evaluation of dynamical interactions within individual climate models (e.g. refs.

293 <sup>27,32,33,63</sup>). Our results also add to work on global patterns of precipitation co-organization<sup>64</sup>,  
 294 suggesting atmospheric dynamical interactions as a key driver of important regional climate model  
 295 errors. We see great scope in using our framework to better understand differences between models  
 296 and observations, or among climate models, especially regarding causal interdependencies<sup>26</sup>.  
 297 Finally, we hope that our work will stimulate the use of novel model evaluation metrics. Causal  
 298 discovery algorithms have the potential to be at the forefront of this effort as they are able to detect  
 299 central features of Earth system dynamics such as the direction and time-lag associated with a global  
 300 teleconnection, opening the door for more in-depth causal interpretation studies<sup>26</sup>. CME could be  
 301 used to evaluate many other model systems, or could help tracking the impact of model development  
 302 over time. Ideally, CME will increasingly complement current evaluation approaches<sup>65</sup> and tools<sup>66</sup>,  
 303 and will help constraining uncertainties in climate change projections<sup>67,68</sup>, also for climate variables  
 304 other than global land surface precipitation (Supplementary Fig. 21). The ever expanding use and  
 305 development of machine learning techniques in the scientific community<sup>63,69–72</sup>, as well as the  
 306 upcoming CMIP6<sup>3</sup>, will greatly accelerate this movement. As such we consider our work as an  
 307 important stepping-stone for a range of machine learning and other data-driven methods aimed at  
 308 improving the state-of-the-art of climate modelling and complex system understanding.

## 309 **Methods**

310  **$F_1$  scores for network comparisons.** The network comparisons are purely based on the existence  
 311 or non-existence of links in a network relative to a given reference network, assuming a certain  
 312 statistical significance threshold in the PCMCI method ( $\alpha$ -level). The resulting true links are typically  
 313 only a small fraction (3-10%; depending on the  $\alpha$ -level) of all possible lagged connections ( $N*(N-$   
 314  $1)*\tau_{max}=24,500$ ) so that the binary (link vs. no link) network comparison becomes an imbalanced  
 315 classification problem. The  $F_1$ -score is a widely used, however necessarily imperfect<sup>73</sup>, metric for  
 316 such problems. It balances the statistical precision ( $P$ ) and recall ( $R$ ). It is defined by

$$317 \quad F_1 = \frac{2*P*R}{P+R} \quad (1)$$

318 With precision and recall defined by

$$319 \quad P = \frac{TP}{TP+FP} \quad (2)$$

$$320 \quad R = \frac{TP}{TP+FN} \quad (3)$$

321 Where  $FP$  ( $FN$ ) is the number of falsely detected links (not detected links) relative to the reference  
 322 model and  $TP$  the number of true positive detected links. We further modified the definition of the  $F_1$ -

323 score slightly to account for the sign of dependence (positive or negative) and the networks' discrete  
 324 time-step nature and the expected natural variance in the precise timing of connections: assuming  
 325 a link exists in the reference network A, we tested if a matching link with the same sign of  
 326 dependence exists in network B (with the same causal direction) in a time interval of up to  $\pm 2$  time  
 327 lags; equivalent to a time precision of about  $\pm$  one week (six days). If a link was found at a time lag  
 328 not identical with the reference case, the sign of dependence was tested at the original time step. If  
 329 also found identical, the link was considered to exist in both networks. Due to this relaxation of the  
 330 time-lag constraint, pair-wise network comparison scores do depend on which network is considered  
 331 as the reference case. As a result, the scores for pair-wise network comparisons shown in Figures  
 332 2a and 3a are not symmetric (cross-diagonal entries are not identical) leading to a larger number of  
 333 possible comparisons.  $F_1$ -scores can be calculated for each season, e.g. DJF as shown in Figure 2.  
 334 For the historical networks (Figure 3), an average  $F_1$ -score was calculated from the individual scores  
 335 for each of the four seasons as

$$336 \quad F_1 = \frac{F_{1,DJF} + F_{1,MAM} + F_{1,JJA} + F_{1,SON}}{4} \quad (4)$$

337 **S scores for measuring precipitation modelling skill.** First suggested by Taylor<sup>56</sup>, the S-score  
 338 measures how well a model captures the behaviour of a given climate variable (e.g. temperature,  
 339 precipitation) over a specific spatial domain relative to an observational dataset. It is defined by

$$340 \quad S = \frac{(1+R)^4}{4\left(SDR + \frac{1}{SDR}\right)^2} \quad (5)$$

341 where  $R$  is the pattern correlation coefficient between the models and observations and  $SDR$  is the  
 342 ratio of spatial standard deviations between models and observations<sup>55,56</sup>. The calculation of  $R$  and  
 343  $SDR$  incorporate grid cell area specific weighting with weights  $w$

$$344 \quad R = \frac{\frac{1}{W} \sum_{i=1}^n w_i \left(x_i - \frac{1}{W} \sum_{j=1}^n w_j x_j\right) \left(y_i - \frac{1}{W} \sum_{j=1}^n w_j y_j\right)}{\sigma_{\text{model}} \sigma_{\text{ref}}} \quad (6)$$

345 where  $x_i$  and  $y_i$  are values for the same quantity (e.g. precipitation rate;  $\text{mm day}^{-1}$ ) in a given grid cell  
 346  $i$  in the two datasets to be compared,  $n$  is the number of grid cells, and  $W$  is the sum of area weights

$$347 \quad W = \sum_{j=1}^n w_j \quad (7)$$

348 The spatially-weighted standard deviations  $\sigma$  (that is  $\sigma_{\text{model}}$  and  $\sigma_{\text{ref}}$ ) and the final  $SDR$  term are  
 349 calculated through

$$350 \quad \sigma^2 = \frac{1}{W} \sum_{i=1}^n w_i \left(x_i - \frac{1}{W} \sum_{j=1}^n w_j x_j\right)^2 \quad (8)$$

$$351 \quad SDR = \frac{\sigma_{\text{model}}}{\sigma_{\text{ref}}} \quad (9)$$

352 The S-score thus considers both the pattern similarity over the spatial domain with regard to a given  
 353 quantity as well as their amplitude ratios, as both the spatial coherence and magnitude range of a  
 354 variable is important for measuring model skill<sup>56</sup>.

355 **PCA Varimax.** The dimension reduction step (Figure 1b) serves as a data-driven method to extract  
356 large-scale patterns of regional sea level pressure variability that in many cases resemble well-  
357 known climatological processes such as the ENSO or the North Atlantic Oscillation (NAO). To extract  
358 climatological processes, we here choose truncated principal component analysis, followed by a  
359 Varimax rotation (PCA-Varimax)<sup>34,35</sup>. Principal components, often referred to as empirical orthogonal  
360 functions (EOFs) in climate science and meteorology, are frequently used to identify orthogonal,  
361 uncorrelated global modes of climate variability<sup>25,28,36,37</sup>. To remove noisy components, we then  
362 truncate and keep only the first 100 leading components in terms of their explained variance. The  
363 additional Varimax rotation on these leading components then maximizes the sum of the variances  
364 of the squared weights so that the loading of weights at different grid locations will be either large or  
365 very small. It has been shown that this leads to more physically consistent representations of actual  
366 climate modes, mainly because the Varimax rotation allows spatial patterns associated with the  
367 components to become more localised and their time series of weights to be correlated, as is the  
368 case for actual physical modes<sup>25,36,37</sup>. Principal components without rotation consecutively maximize  
369 variance and therefore often mix contributions of physically defined modes such ENSO, Pacific  
370 Decadal Oscillation (PDO), or the NAO, whose time-behaviour is not orthogonal, making patterns  
371 more difficult to interpret. We here estimated the spatial pattern (loading) of the Varimax components  
372 from 70-year (1948-2017) daily sea level pressure anomalies of the NCAR-NCEP reanalysis  
373 dataset<sup>41</sup> and then used these weights to also consistently extract the Varimax component time  
374 series from the CMIP5 sea level pressure simulations. The motivation behind using sea level  
375 pressure as the variable underlying the networks is that it is a standard variable to characterize large-  
376 scale atmospheric dynamics and corresponding variability, e.g. in climate modes or weather  
377 patterns. Therefore, it is also available in virtually any reanalysis dataset or model data archive,  
378 which allowed us to work with the largest possible number of ensemble members for the CMIP5  
379 analysis. The components obtained for the four meteorological seasons for the NCEP data can be  
380 found in Supplementary Figs. 22-421. For the subsequent causal discovery method, we further  
381 filtered weights in terms of their spatial separability and their frequency spectra, leading to a total of  
382 fifty components for each season. For example, we typically excluded components that exhibited a  
383 sudden change in behaviour when entering the satellite era (1979-), which resulted in unresolved  
384 frequency spectra (e.g. DJF components 18, 36, 38, 41 provided as Supplementary Figs. 40, 58, 60,  
385 and 63). Such apparently unphysical component time series changes were in particular found in  
386 Asia, Africa and the Middle East and could therefore be related to a lack of historical data coverage  
387 feeding into the reanalysis in those regions. To further control for the importance of choosing a  
388 certain set of components for the overall results and conclusions, we sometimes included some of  
389 these components for certain seasons (e.g. component 7 for DJF), but we did not find any noticeable  
390 sensitivity of the relative  $F_1$ -scores to this selection process. A side effect of this selection process,  
391 however, remains a reduced network coverage in those areas. Overall, we found that the global  
392 network metrics were effectively insensitive to the choice of nodes and their geographical  
393 distribution. This is also evident from the relative insensitivity of the model rankings to the specific

394 season (Supplementary Figures 1, 3-6 and Supplementary Table 1). The indices of the fifty  
395 components chosen for each season are provided at the beginning of each section in Supplementary  
396 section 2. The component time series were averaged to 3-day-means before the application of  
397 PCMCI. This time-aggregation presents a compromise to resolve short-term interactions in our  
398 intercomparison (a few days), while limiting the increase in dimensionality due to additional time lags  
399 (here 10 time lags for  $\tau_{\max}=30$ ).

400 **PCMCI causal discovery method.** PCMCI is a time series causal discovery method further  
401 described in ref. <sup>23</sup>. Commonly, causal discovery for time series is conducted with Granger causality  
402 which is based on fitting a multivariate autoregressive time series model of a variable  $Y$  on its own  
403 past, the past of a potential driver  $X$ , and all the remaining variables' past (up to some maximum  
404 time delay  $\tau_{\max}$ ). Then  $X$  Granger-causes  $Y$  if any of the coefficients corresponding to different time  
405 lags of  $X$  is non-zero (typically tested by an F-test). As analyzed in ref. <sup>23</sup>, Granger causality, due to  
406 a too high model complexity given finite sample size, has low detection power for causal links (true  
407 positive rate) if too many variables are used and for strong autocorrelation, both of which are relevant  
408 in our analysis. PCMCI avoids conditioning on all variables by an efficient condition-selection step  
409 (PC) that iteratively performs conditional independence tests to identify the typically few relevant  
410 necessary conditions. In a second step, this much smaller set of conditions is used in the momentary  
411 conditional independence (MCI) test that alleviates the problem of strong autocorrelation. In general,  
412 both the PC and MCI step can be implemented with linear or nonlinear conditional independence  
413 tests. Here we focus on the linear case and utilize partial correlation (ParCorr). A causal  
414 interpretation rests on a number of standard assumptions of causal discovery as discussed in ref.  
415 22, such as the Causal Markov assumption, Faithfulness, and stationarity of the causal network over  
416 the time sample considered. The free parameter of PCMCI is the maximum time delay  $\tau_{\max}$ , here  
417 chosen to include atmospheric timescales over which we expect dependencies to be stationary. The  
418 pruning hyper-parameter  $pc-\alpha$  in the PC condition-selection step is optimized using the Akaike  
419 information criterion (among  $pc-\alpha = 0.05, 0.1, 0.2, 0.3, 0.4, 0.5$ ). PCMCI yields a  $p$ -value (based on  
420 a two-sided  $t$ -test) for every pair of components at different lags. We defined links in the networks  
421 using a strict significance level of  $10^{-4}$  in the main paper. However, very similar results are found for  
422 other more relaxed or even stricter significance levels; as demonstrated extensively in the  
423 Supplementary Material.

424 **Other network construction methods.** As discussed in the main text, causal networks have been  
425 used several times before in the climate context. Two of the most prominent cases of such studies  
426 are those described in refs. <sup>32,33</sup>, where Ebert-Uphoff and Deng also discuss remote impacts and  
427 information pathways as well as the role of storm tracks as important drivers of network connectivity.  
428 Their work is further a good demonstration of other possible ways to construct causal networks, the  
429 effect of which might be an interesting topic for future studies. For example, their network approach  
430 was carried out on a grid-cell-wise level rather than using PCA Varimax components. The latter are  
431 designed to capture distinct regional climatological processes while an analysis at the grid-cell level



432 is more granular which, however, carries the challenges of higher dimensionality, will have a strong  
433 redundancy among neighbouring grid cells, and grid-level metrics will require handling varying  
434 spatial resolution among datasets. Furthermore, the original PC causal discovery algorithm used in  
435 their work is less suited for the time series case than PCMCI<sup>23</sup>. They also used another  
436 meteorological variable (500 hPa geopotential height) to construct their networks and compared  
437 aggregate network metrics rather than comparing networks on a link-by-link basis.

438 **CMIP5 data.** For the network constructions, we used daily mean sea level pressure data from the  
439 CMIP5 data archive, as stored by the British Atmospheric Data Centre (BADC). An overview of all  
440 models and simulations used is given in Supplementary Table 2. The twenty models used for the  
441 pre-industrial networks are as labelled in Figure 2a. The twenty models used for the historical and  
442 RCP8.5 reference case are as labelled in Figure 3a. Typically, we used the final 210 years of each  
443 pre-industrial simulation, assuming that these years represent the most equilibrated state of each  
444 model. For historical and RCP8.5 simulations, we used at least three ensemble members which  
445 typically covered 70 years between 1<sup>st</sup> January 1936 and 31<sup>st</sup> December 2017. Relaxing the left time  
446 boundary by up to twelve years relative to the reanalysis data time period allowed us to include more  
447 models, as some modelling centres ran more historical than RCP8.5 simulations. If sufficient data  
448 was available for both the historical and RCP8.5 simulation, the two simulations were merged on 1<sup>st</sup>  
449 January 2006; the day after historical simulations ended in most cases. All data (including the  
450 reanalysis datasets) was linearly de-trended on a grid cell basis and seasonally anomalized by  
451 removing the long-term daily mean. Note that sea level pressure data is effectively stationary even  
452 under historically forced climatic conditions so that the de-trending is a prudent step to remove any  
453 potentially occurring small trends to a good approximate degree. Of course, we cannot fully account  
454 for the very long time-scales that may be associated with some climate processes<sup>74</sup> beyond the time-  
455 scale covered by each individual dataset. Each model dataset was bi-linearly interpolated to a 2.5°  
456 latitude x 2.5° longitude grid in order to extract the component time series based on the Varimax  
457 loading weights computed from the NCAR-NCEP<sup>41</sup> reanalysis data.

458 **Precipitation data.** As observational reference, we used the land surface CRU TS v4.03 dataset  
459 from the University of East Anglia<sup>57</sup>, which does not cover Antarctica. CMIP5 precipitation data was  
460 taken from single ensemble members (Supplementary Table 2) of the historical and RCP8.5  
461 simulations, as described above. As for the sea level pressure data, all precipitation data was bi-  
462 linearly interpolated to the NCAR-NCEP spatial grid prior to the intercomparison. Climate change-  
463 induced differences shown in Figures 4b,c were calculated by subtracting the model-specific land  
464 surface (using an ocean and Antarctica mask equivalent to the one of the CRU dataset) average  
465 precipitation rate for the period 1860-1910 (covered by all models) from the same measure for the  
466 years 2050-2100.

467 **Random and Independent data.** The datasets for the Random and Independent case in Figure 2a  
468 were created with Gaussian noise driven multivariate autoregressive models of the same number of  
469 variables as in the original data. For the Independent case only the lag-1 autocorrelation coefficients

470 are non-zero and set to a value of 0.7. Hence, all variables are independent, but due to finite sample  
471 effects, the estimated networks with PCMCI will still contain some cross-links. For the Random case,  
472 we created a random network with a link density of 5%, randomly connecting two components at  
473 lag-1 with a coefficient of 0.1, in addition to autocorrelation coefficients with a value of 0.7 for each  
474 component. Like for the original data, we simulated three datasets (covering 70-year periods of the  
475 210 years) with the same sample size as the original data.

476 **Gaussian Process regression.** To estimate the nonlinear dependency between  $F_I/S$ -scores and  
477 land precipitation changes (Figures 4b,c and Supplementary Figure 14), we used Gaussian  
478 Processes (GP) as a widely used Bayesian non-parametric regression approach<sup>75</sup>. We implemented  
479 the GP with a standard radial basis function kernel with an added white noise kernel and optimized  
480 the hyperparameters using the log-marginal likelihood. The resulting fit line is approximately  
481 parabolic when using the  $F_I$ -score. In Supplementary Figure 15 we also directly fit a parabolic  
482 function  $y=a+bx+cx^2$ .

483 **Data availability.** All raw sea level pressure, surface temperature and precipitation rate data is  
484 publicly available. CMIP5 data is available through the Lawrence Livermore Laboratory  
485 (<https://pcmdi.llnl.gov/mips/cmip5/availability.html>) and many other sources such as the British  
486 Atmospheric Data Centre (BADC, <http://www.badc.rl.ac.uk/>) as variables 'psl', 'tas' and 'pr', see  
487 Supplementary Table 2 for an overview of all selected simulations. CRU precipitation rate data is  
488 publicly available through e.g. <https://crudata.uea.ac.uk/cru/data/hrg/>; as is the NCAR-NCEP  
489 reanalysis through <https://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html>. ERA-  
490 Interim data is accessible via [https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-  
491 interim](https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era-interim). The source data underlying Figures 2a–f, 3a/b, and 4a-c are provided as a Source Data file.

492 **Code Availability.** Tigramite source code is available through  
493 <https://github.com/jakobrunge/tigramite>. Example Jupyter-notebooks and Python code used to carry  
494 out the Varimax and PCMCI analysis here will be made available through  
495 [https://github.com/peernow/CME\\_NCOMMS\\_2020](https://github.com/peernow/CME_NCOMMS_2020).

## 496 **References**

- 497 1. Stocker, T. F. *et al.* *Climate Change 2013: the Physical Science Basis. Contribution of working group*  
498 *I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, Geneva,*  
499 *Switzerland.* (Cambridge University Press, 2013).
- 500 2. Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull.*  
501 *Am. Meteorol. Soc.* **93**, 485–498 (2012).
- 502 3. Eyring, V. *et al.* Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6)  
503 experimental design and organization. *Geosci. Model Dev.* **9**, 1937–1958 (2016).
- 504 4. Rea, G., Riccio, A., Fierli, F., Cairo, F. & Cagnazzo, C. Stratosphere-resolving CMIP5 models  
505 simulate different changes in the Southern Hemisphere. *Clim. Dyn.* **50**, 2239–2255 (2018).
- 506 5. Friedlingstein, P. *et al.* Uncertainties in CMIP5 Climate Projections due to Carbon Cycle Feedbacks.

- 507 *J. Clim.* **27**, 511–526 (2013).
- 508 6. Nowack, P. J. *et al.* A large ozone-circulation feedback and its implications for global warming  
509 assessments. *Nat. Clim. Chang.* **5**, 41–45 (2015).
- 510 7. Nowack, P. J., Abraham, N. L., Braesicke, P. & Pyle, J. A. The impact of stratospheric ozone  
511 feedbacks on climate sensitivity estimates. *J. Geophys. Res. Atmos.* **123**, 4630–4641 (2018).
- 512 8. Shindell, D. T. & Faluvegi, G. Climate response to regional radiative forcing during the twentieth  
513 century. *Nat. Geosci.* **2**, 294–300 (2009).
- 514 9. Bastos, A. *et al.* European land CO<sub>2</sub> sink influenced by NAO and East-Atlantic Pattern coupling. *Nat.*  
515 *Commun.* **7**, 10315 (2016).
- 516 10. Bell, C. J., Gray, L. J., Charlton-Perez, A. J., Joshi, M. M. & Scaife, A. A. Stratospheric  
517 communication of El Niño teleconnections to European winter. *J. Clim.* **22**, 4083–4096 (2009).
- 518 11. Hourdin, F. *et al.* The art and science of climate model tuning. *Bull. Am. Meteorol. Soc.* **98**, 589–602  
519 (2017).
- 520 12. Knutti, R. The end of model democracy? *Clim. Change* **102**, 395–404 (2010).
- 521 13. Sherwood, S. C., Bony, S. & Dufresne, J.-L. Spread in model climate sensitivity traced to  
522 atmospheric convective mixing. *Nature* **505**, 37–42 (2014).
- 523 14. Shepherd, T. G. Atmospheric circulation as a source of uncertainty in climate change projections.  
524 *Nat. Geosci.* **7**, 703–708 (2014).
- 525 15. Knutti, R. & Sedláček, J. Robustness and uncertainties in the new CMIP5 climate model projections.  
526 *Nat. Clim. Chang.* **3**, 369–373 (2013).
- 527 16. Bellenger, H., Guilyardi, E., Leloup, J., Lengaigne, M. & Vialard, J. ENSO representation in climate  
528 models: from CMIP3 to CMIP5. *Clim. Dyn.* **42**, 1999–2018 (2013).
- 529 17. Langenbrunner, B. & Neelin, J. D. Analyzing ENSO teleconnections in CMIP models as a measure of  
530 model fidelity in simulating precipitation. *J. Clim.* **26**, 4431–4446 (2013).
- 531 18. Wenzel, S., Eyring, V., Gerber, E. P. & Karpechko, A. Y. Constraining future summer austral jet  
532 stream positions in the CMIP5 ensemble by process-oriented multiple diagnostic regression. *J. Clim.*  
533 **29**, 673–687 (2016).
- 534 19. Eyring, V. *et al.* Taking climate model evaluation to the next level. *Nat. Clim. Chang.* **9**, 102–110  
535 (2019).
- 536 20. Eyring, V. *et al.* A strategy for process-oriented validation of coupled chemistry-climate models. *Bull.*  
537 *Am. Meteorol. Soc.* **86**, 1117–1133 (2005).
- 538 21. Spirtes, P. Introduction to Causal Inference Approaches. *J. Mach. Learn. Res.* **11**, 1643–1662 (2010).
- 539 22. Runge, J. Causal network reconstruction from time series: From theoretical assumptions to practical  
540 estimation. *Chaos An Interdiscip. J. Nonlinear Sci.* **28**, 075310 (2018).
- 541 23. Runge, J., Nowack, P. J., Kretschmer, M., Flaxman, S. & Sejdinovic, D. Detecting and quantifying  
542 causal associations in large nonlinear time series datasets. *Sci. Adv.* **5**, 1–46 (2019).
- 543 24. Kretschmer, M., Coumou, D., Donges, J. F. & Runge, J. Using Causal Effect Networks to analyze  
544 different Arctic drivers of mid-latitude winter circulation. *J. Clim.* **29**, 4069–4081 (2016).
- 545 25. Runge, J. *et al.* Identifying causal gateways and mediators in complex spatio-temporal systems. *Nat.*

- 546 *Commun.* **6**, 8502 (2015).
- 547 26. Runge, J. *et al.* Inferring causation from time series in Earth system sciences. *Nat. Commun.* **10**, 2553  
548 (2019).
- 549 27. Kretschmer, M., Runge, J. & Coumou, D. Early prediction of extreme stratospheric polar vortex  
550 states based on causal precursors. *Geophys. Res. Lett.* **44**, 8592–8600 (2017).
- 551 28. Trenberth, K. E. *et al.* Progress during TOGA in understanding and modeling global teleconnections  
552 associated with tropical sea surface temperatures. *J. Geophys. Res.* **103**, 14291–14324 (1998).
- 553 29. Yeh, S. W. *et al.* ENSO Atmospheric Teleconnections and Their Response to Greenhouse Gas  
554 Forcing. *Rev. Geophys.* **56**, 185–206 (2018).
- 555 30. Bjerknes, J. Atmospheric teleconnections from the equatorial Pacific. *Mon. Weather Rev.* **97**, 163–  
556 172 (1969).
- 557 31. Braesicke, P., Morgenstern, O. & Pyle, J. Might dimming the sun change atmospheric ENSO  
558 teleconnections as we know them? *Atmos. Sci. Lett.* **12**, 184–188 (2011).
- 559 32. Deng, Y. & Ebert-Uphoff, I. Weakening of atmospheric information flow in a warming climate in the  
560 Community Climate System Model. *Geophys. Res. Lett.* **41**, 193–200 (2014).
- 561 33. Ebert-Uphoff, I. & Deng, Y. A new type of climate network based on probabilistic graphical models:  
562 Results of boreal winter versus summer. *Geophys. Res. Lett.* **39**, 1–7 (2012).
- 563 34. Kaiser, H. F. The varimax criterion for varimax rotation in factor analysis. *Psychometrika* **23**, 187–  
564 204 (1958).
- 565 35. Vautard, R. & Ghil, M. Singular spectrum analysis in nonlinear dynamics, with applications to  
566 paleoclimatic time series. *Phys. D Nonlinear Phenom.* **35**, 395–424 (1989).
- 567 36. Hannachi, A., Jolliffe, I. T. & Stephenson, D. B. Empirical orthogonal functions and related  
568 techniques in atmospheric science: A review. *Int. J. Climatol.* **27**, 1119–1152 (2007).
- 569 37. Vejmelka, M. *et al.* Non-random correlation structures and dimensionality reduction in multivariate  
570 climate data. *Clim. Dyn.* **44**, 2663–2682 (2015).
- 571 38. Stan, C. *et al.* Review of Tropical-Extratropical Teleconnections on Intraseasonal Time Scales. *Rev.*  
572 *Geophys.* **55**, 902–937 (2017).
- 573 39. Nowack, P. J., Braesicke, P., Abraham, N. L. & Pyle, J. A. On the role of ozone feedback in the  
574 ENSO amplitude response under global warming. *Geophys. Res. Lett.* **44**, 3858–3866 (2017).
- 575 40. Karoly, D. J. Southern Hemisphere Circulation Features Associated with El Niño-Southern  
576 Oscillation Events. *Journal of Climate* **2**, 1239–1252 (1989).
- 577 41. Kalnay, E. *et al.* The NCEP NCAR 40-Year Reanalysis Project. *Bull. Am. Meteorol. Soc.* **77**, 437–  
578 472 (1996).
- 579 42. Hegerl, G. C. *et al.* Detecting greenhouse-gas-induced climate change with an optimal fingerprint  
580 method. *Journal of Climate* **9**, 2281–2306 (1996).
- 581 43. Hegerl, G., Zwiers, F. & Tebaldi, C. Patterns of change: whose fingerprint is seen in global warming?  
582 *Environ. Res. Lett.* **6**, 044025 (2011).
- 583 44. Batehup, R., McGregor, S. & Gallant, A. J. E. The influence of non-stationary teleconnections on  
584 palaeoclimate reconstructions of ENSO variance using a pseudoproxy framework. *Clim. Past* **11**,

- 585 1733–1749 (2015).
- 586 45. Ashcroft, L., Gergis, J. & Karoly, D. J. Long-term stationarity of El Niño–Southern Oscillation  
587 teleconnections in southeastern Australia. *Clim. Dyn.* **46**, 2991–3006 (2016).
- 588 46. Knutti, R. *et al.* A climate model projection weighting scheme accounting for performance and  
589 interdependence. *Geophys. Res. Lett.* **44**, 1909–1918 (2017).
- 590 47. Sanderson, B. M., Knutti, R. & Caldwell, P. Addressing interdependency in a multimodel ensemble  
591 by interpolation of model properties. *J. Clim.* **28**, 5150–5170 (2015).
- 592 48. Sanderson, B. M., Wehner, M. & Knutti, R. Skill and independence weighting for multi-model  
593 assessments. *Geosci. Model Dev.* **10**, 2379–2395 (2017).
- 594 49. Bishop, C. H. & Abramowitz, G. Climate model dependence and the replicate Earth paradigm. *Clim.*  
595 *Dyn.* **41**, 885–900 (2013).
- 596 50. Abramowitz, G. & Bishop, C. H. Climate model dependence and the ensemble dependence  
597 transformation of CMIP projections. *J. Clim.* **28**, 2332–2348 (2015).
- 598 51. Jones, C. D. *et al.* The HadGEM2-ES implementation of CMIP5 centennial simulations. *Geosci.*  
599 *Model Dev.* **4**, 543–570 (2011).
- 600 52. Collins, W. J. *et al.* Development and evaluation of an Earth-System model – HadGEM2. *Geosci.*  
601 *Model Dev.* **4**, 1051–1075 (2011).
- 602 53. Kruskal, W. H. & Wallis, W. A. Use of Ranks in One-Criterion Variance Analysis. *J. Am. Stat.*  
603 *Assoc.* **47**, 583–621 (1952).
- 604 54. Dee, D. P. *et al.* The ERA-Interim reanalysis: Configuration and performance of the data assimilation  
605 system. *Q. J. R. Meteorol. Soc.* **137**, 553–597 (2011).
- 606 55. Chen, L. & Frauenfeld, O. W. A comprehensive evaluation of precipitation simulations over China  
607 based on CMIP5 multimodel ensemble projections. *J. Geophys. Res. Atmos.* **119**, 5767–5786 (2014).
- 608 56. Taylor, K. E. Summarizing multiple aspects of model performance in a single diagram. *J. Geophys.*  
609 *Res.* **106**, 7183–7192 (2001).
- 610 57. Harris, I., Jones, P. D., Osborn, T. J. & Lister, D. H. Updated high-resolution grids of monthly  
611 climatic observations - the CRU TS3.10 Dataset. *Int. J. Climatol.* **34**, 623–642 (2014).
- 612 58. Blackmon, M. L. A climatological spectral study of the 500 mb geopotential height of the Northern  
613 Hemisphere. *Journal of the Atmospheric Sciences* **33**, 1607–1623 (1976).
- 614 59. Ulbrich, U. *et al.* Changing Northern Hemisphere storm tracks in an ensemble of IPCC climate  
615 change simulations. *J. Clim.* **21**, 1669–1679 (2008).
- 616 60. Byrne, M. P. & O’Gorman, P. A. Trends in continental temperature and humidity directly linked to  
617 ocean warming. *Proc. Natl. Acad. Sci.* **115**, 4863–4868 (2018).
- 618 61. Seneviratne, S. I. *et al.* Impact of soil moisture-climate feedbacks on CMIP5 projections: First results  
619 from the GLACE-CMIP5 experiment. *Geophys. Res. Lett.* **40**, 5212–5217 (2013).
- 620 62. Rowell, D. P., Senior, C. A., Vellinga, M. & Graham, R. J. Can climate projection uncertainty be  
621 constrained over Africa using metrics of contemporary performance? *Clim. Change* **134**, 621–633  
622 (2016).
- 623 63. Falasca, F., Bracco, A., Nenes, A. & Fountalis, I. Dimensionality reduction and network inference for

- 624 climate data using  $\delta$ -MAPS: application to the CESM Large Ensemble sea surface temperature. *J.*  
625 *Adv. Model. Earth Syst.* **11**, 1–37 (2019).
- 626 64. Boers, N. *et al.* Complex networks reveal global pattern of extreme-rainfall teleconnections. *Nature*  
627 **566**, 373–377 (2019).
- 628 65. Flato *et al.*, G. Evaluation of Climate Models. in *Climate Change 2013: The Physical Science Basis.*  
629 *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on*  
630 *Climate Change* 741–866 (Cambridge University Press, 2013).
- 631 66. Eyring, V. *et al.* ESMValTool (v1.0)-a community diagnostic and performance metrics tool for  
632 routine evaluation of Earth system models in CMIP. *Geosci. Model Dev.* **9**, 1747–1802 (2016).
- 633 67. Alex Hall, Cox, P., Huntingford, C. & Klein, S. Progressing emergent constraints on future climate  
634 change. *Nat. Clim. Chang.* **9**, 269–278 (2019).
- 635 68. Collins, M. *et al.* Challenges and opportunities for improved understanding of regional climate  
636 dynamics. *Nat. Clim. Chang.* **8**, 101–108 (2018).
- 637 69. Nowack, P. *et al.* Using machine learning to build temperature-based ozone parameterizations for  
638 climate sensitivity simulations. *Environ. Res. Lett.* **13**, 104016 (2018).
- 639 70. Reichstein, M. *et al.* Deep learning and process understanding for data-driven Earth system science.  
640 *Nature* **566**, 195–204 (2019).
- 641 71. Ebert-Uphoff, I. & Deng, Y. Causal discovery for climate research using graphical models. *J. Clim.*  
642 **25**, 5648–5665 (2012).
- 643 72. Monteleoni, C. *et al.* Climate Informatics. in *Computational Intelligent Data Analysis for Sustainable*  
644 *Development* (eds. Yu, T., Chawla, N. & Simoff, S.) 81–126 (Chapman and Hall/CRC, 2013).
- 645 73. Bódai, T. Predictability of threshold exceedances in dynamical systems. *Phys. D Nonlinear Phenom.*  
646 **313**, 37–50 (2015).
- 647 74. Herein, M., Drótos, G., Bódai, T., Lunkeit, F. & Lucarini, V. Reconsidering the relationship of the El  
648 Niño-Southern Oscillation and the Indian monsoon using ensembles in Earth system models. *Preprint*  
649 *at: <https://arxiv.org/abs/1803.08909>* (2019).
- 650 75. Rasmussen, C. E. & Williams, C. K. I. *Gaussian Processes for Machine Learning.* (MIT Press,  
651 2006).

652

653 **Acknowledgements.** P.J.N. is supported through an Imperial College Research Fellowship. J.R.  
654 was supported by a Fellowship from the James S. McDonnell Foundation. We acknowledge the  
655 World Climate Research Programme's Working Group on Coupled Modelling, which is responsible  
656 for CMIP, and we thank the climate modeling groups (listed in Supplementary Table 2 of this paper)  
657 for producing and making available their model output. For CMIP5 the U.S. Department of Energy's  
658 Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led  
659 development of software infrastructure in partnership with the Global Organization for Earth System  
660 Science Portals. For plotting, we used Matplotlib, a 2D graphics environment for the Python  
661 programming language developed by J. D. Hunter. For causal discovery we used the Tigramite

662 package (version 4.1) available from <https://github.com/jakobrunge/tigramite>. We thank James King  
663 (University of Oxford) for helpful discussions.

664 **Author contributions.** P.N. and J.R. together suggested and designed the study. P.N. led the  
665 scientific analysis and paper writing in collaboration with J.R. All authors (i.e. P.N., J.R., V.E. and  
666 J.D.H) contributed to the scientific interpretation of the results and to the paper writing.

667 **Competing interests.** The authors declare no competing interests.