










# Response of stratospheric water vapour to warming constrained by satellite observations

Received: 22 July 2022

Accepted: 12 April 2023

Published online: 26 June 2023

 Check for updates

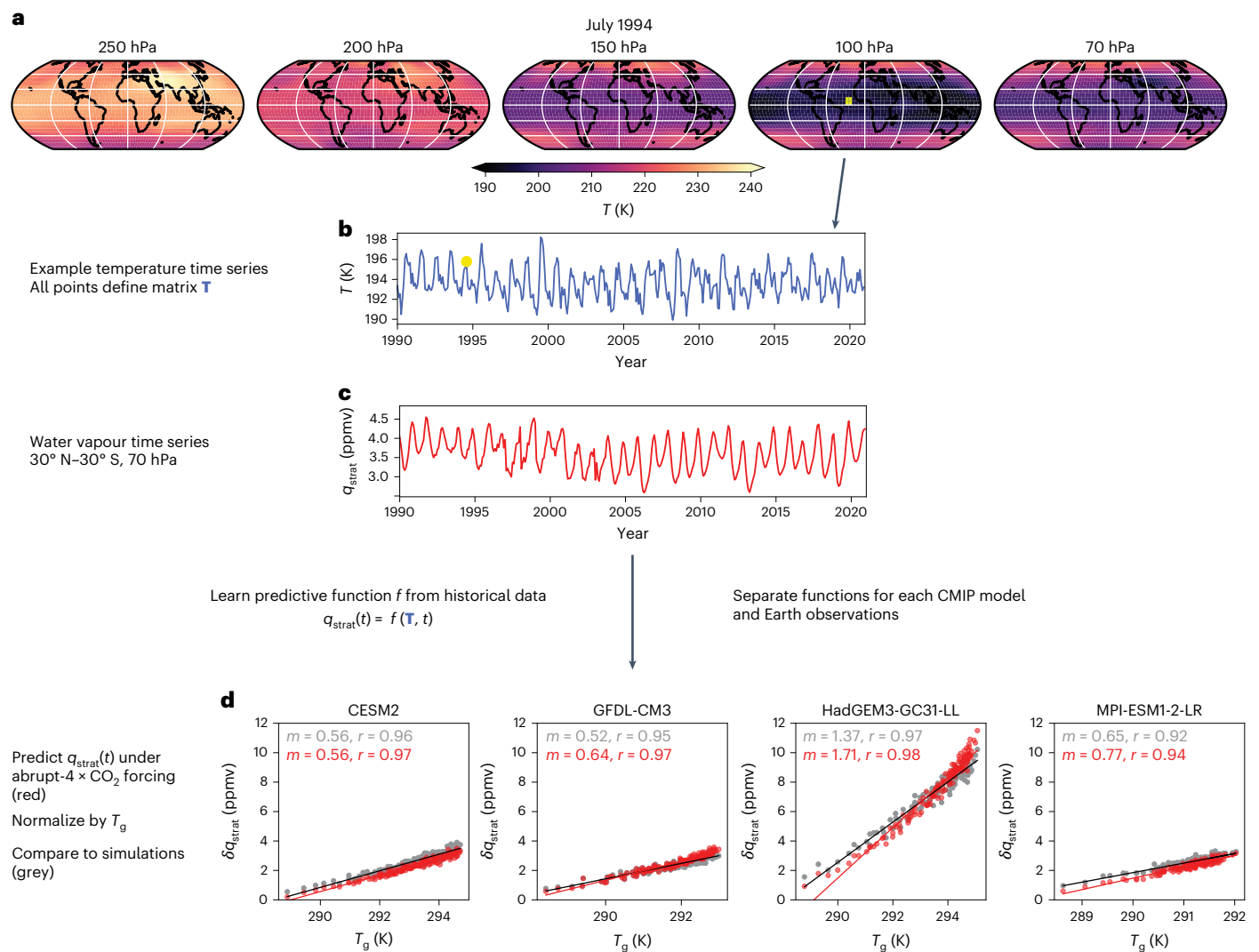
Peer Nowack <sup>1,2,3,4</sup>✉, Paulo Ceppi <sup>2</sup>, Sean M. Davis <sup>5</sup>, Gabriel Chiodo <sup>6</sup>, Will Ball <sup>6,7,8,14</sup>, Mohamadou A. Diallo <sup>9</sup>, Birgit Hassler <sup>10</sup>, Yue Jia <sup>5,11</sup>, James Keeble <sup>12,13</sup> & Manoj Joshi <sup>1</sup>

Future increases in stratospheric water vapour risk amplifying climate change and slowing down the recovery of the ozone layer. However, state-of-the-art climate models strongly disagree on the magnitude of these increases under global warming. Uncertainty primarily arises from the complex processes leading to dehydration of air during its tropical ascent into the stratosphere. Here we derive an observational constraint on this longstanding uncertainty. We use a statistical-learning approach to infer historical co-variations between the atmospheric temperature structure and tropical lower stratospheric water vapour concentrations. For climate models, we demonstrate that these historically constrained relationships are highly predictive of the water vapour response to increased atmospheric carbon dioxide. We obtain an observationally constrained range for stratospheric water vapour changes per degree of global warming of  $0.31 \pm 0.39$  ppmv K<sup>-1</sup>. Across 61 climate models, we find that a large fraction of future model projections are inconsistent with observational evidence. In particular, frequently projected strong increases (>1 ppmv K<sup>-1</sup>) are highly unlikely. Our constraint represents a 50% decrease in the 95th percentile of the climate model uncertainty distribution, which has implications for surface warming, ozone recovery and the tropospheric circulation response under climate change.

The stratosphere is extremely dry. This was first realized by Alan Brewer in his pioneering analysis of balloon measurements in the 1940s, where he reported that the atmospheric *water content is found to fall very rapidly just above the tropopause*<sup>1</sup>. It is now well established that average

stratospheric specific humidity is around 3–5 parts per million volume (ppmv) globally, with substantial daily to decadal variations driven by volcanic eruptions<sup>2,3</sup>, convective overshooting<sup>4</sup>, monsoonal circulations<sup>5</sup> and climate modes such as the El Niño–Southern Oscillation

<sup>1</sup>Climatic Research Unit, School of Environmental Sciences, University of East Anglia, Norwich, UK. <sup>2</sup>Grantham Institute and Department of Physics, Imperial College London, London, UK. <sup>3</sup>Data Science Institute, Imperial College London, London, UK. <sup>4</sup>Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany. <sup>5</sup>NOAA Chemical Sciences Laboratory, Boulder, CO, USA. <sup>6</sup>Institute for Atmospheric and Climate Science, ETH Zurich, Zurich, Switzerland. <sup>7</sup>Department of Geoscience and Remote Sensing, Delft University of Technology, Delft, The Netherlands. <sup>8</sup>Physik alisch-Meteorologisches Observatorium Davos World Radiation Centre, Davos, Switzerland. <sup>9</sup>Institute of Energy and Climate Research, Stratosphere (IEK-7), Forschungszentrum Jülich, Jülich, Germany. <sup>10</sup>Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany. <sup>11</sup>Cooperative Institute for Research in Environmental Sciences (CIRES), University of Colorado Boulder, Boulder, CO, USA. <sup>12</sup>Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, UK. <sup>13</sup>National Centre for Atmospheric Science (NCAS), University of Cambridge, Cambridge, UK. <sup>14</sup>Deceased: Will Ball. ✉e-mail: [peer.nowack@kit.edu](mailto:peer.nowack@kit.edu)



**Fig. 1 | Sketch of the statistical-learning framework.** **a**, Example ERA5 reanalysis temperature ( $T$ ) data from the European Centre for Medium-Range Weather Forecasts (ECMWF) for July 1994 at all five pressure levels used to predict tropical lower SWV ( $q_{\text{strat}}$ ). **b**, Example of a temperature time series for the grid location highlighted in yellow on the 100 hPa map in **a**; the yellow dot indicates the July 1994 value. All  $5^\circ \times 5^\circ$  grid points within  $60^\circ \text{N}$ – $60^\circ \text{S}$  serve as predictors for  $q_{\text{strat}}(t)$ , defining a temperature matrix  $\mathbf{T}(t)$ . **c**, The observational  $q_{\text{strat}}(t)$  time series. Using ridge regression, we learn predictive relationships between  $\mathbf{T}$  covering all five pressure levels and  $q_{\text{strat}}$ , considering simultaneous and lagged temperature data ( $\tau_{\text{max}} = 2$  months). This process is applied to 150 combinations of temperature reanalyses and versions of  $q_{\text{strat}}$  observations from 1990 to 2020 (Methods). The result is an ensemble of predictive functions that are consistent with the observational record. We then learn equivalent functions from data

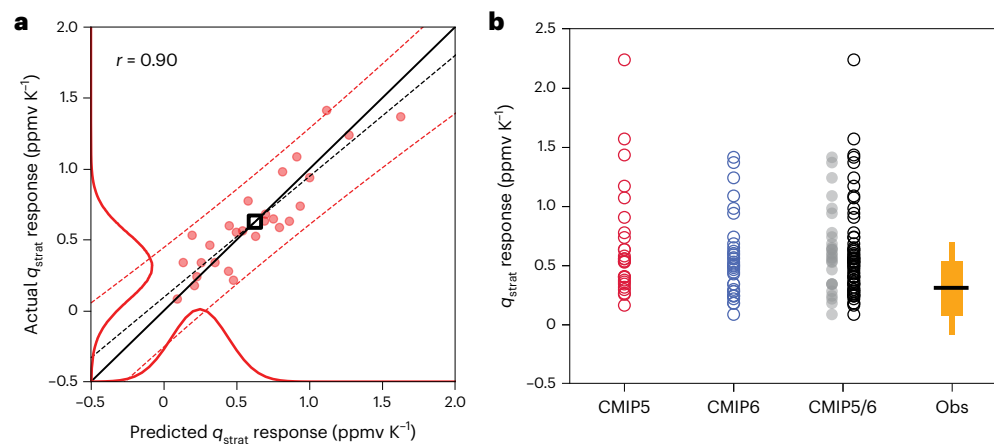
produced by 27 climate models from the CMIP5/6 archives, that is, using data of the same spatial and temporal coverage for  $\mathbf{T}$  and  $q_{\text{strat}}$ . **d**, For these CMIP models, the predictive skill of the functions can be evaluated under strong climate change scenarios. As illustrated here for four CMIP models, this is achieved by comparing the annually averaged predictions (red) with actual abrupt-4 × CO<sub>2</sub> simulation results (grey). From these comparisons, a framework-related uncertainty is estimated characterizing imperfections in the statistical-learning predictions under 4 × CO<sub>2</sub>, which needs to be smaller than the model uncertainty to be constrained. To account for different levels of global warming simulated by each model, the  $q_{\text{strat}}$  response (in ppmv; here shown relative to average 1990–2020 levels) is normalized by the change in global mean surface temperature ( $T_g$ ). Colour-coded inset values for  $m$  represent the linear regression slopes (ppmv K<sup>-1</sup>) and  $r$  the correlation coefficients.

(ENSO)<sup>6,7</sup> and the Quasi-Biennial Oscillation (QBO)<sup>8,9</sup>. Brewer also already suggested that the dryness of the stratosphere can be explained by a large-scale stratospheric overturning circulation nowadays referred to as the Brewer–Dobson circulation (BDC)<sup>10</sup>, where air is freeze dried to very low concentrations as it enters the stratosphere through the cold tropical upper troposphere and lower stratosphere (UTLS).

The freeze-drying process is now better understood than ever, for example, using air parcel trajectory models of the tropical UTLS<sup>5,11–16</sup>. However, despite this qualitative understanding, there is still substantial model uncertainty in projections of future changes in stratospheric water vapour (SWV)<sup>17–21</sup>. The uncertainty is a pressing concern, not only because SWV is a greenhouse gas affecting surface temperature and the

atmospheric circulation<sup>22–26</sup>, but also because of its key role in shaping atmospheric chemistry and stratospheric ozone recovery<sup>27–29</sup>. Changes in the thickness of the ozone layer, in turn, affect the tropospheric photochemical environment, air quality, human health and ecology<sup>30,31</sup>.

Historically, tropical lower SWV observations show—if anything—a slight decrease over the last three decades (Supplementary Fig. 1; refs. 15,32), at least until the recent Hunga–Tonga eruption<sup>33</sup>. In contrast, the majority of climate models show long-term increases in historical simulations (Supplementary Figs. 1 and 2; refs. 17,21). Given substantial model biases in background concentrations and seasonal cycle representations of SWV and UTLS temperatures<sup>17,21</sup>, one might therefore ask if these models can reliably project SWV for future scenarios.



**Fig. 2 | Framework performance and the observational constraint.** **a**, Red circles show abrupt-4 × CO<sub>2</sub> simulation results ('actual') regressed against predicted changes in  $q_{\text{strat}}$ , both normalized by  $T_g$ , for 27 CMIP models. The multi-model mean is indicated as a black square; the one-to-one line in solid black. Dashed lines show the least squares regression fit (black) and the 5% to 95% prediction intervals (red). The probability distributions (red curves) on the axes represent the observational estimates. The distribution on top of the x axis indicates the spread in predictions based on combining functions learned from observations with the CMIP temperature responses. The final probability distribution, defining the observational constraint, is attached to the y axis and additionally accounts for the framework prediction uncertainty.

**b**, The observational constraint ( $n = 4,050$ ) relative to CMIP model uncertainty. Circles show  $T_g$ -normalized changes in  $q_{\text{strat}}$  for 27 CMIP5 models (red), 34 CMIP6 models (blue) and their combination (black). The grey circles indicate the selected 27 models fulfilling the minimum variance criterion compared with SWOOSH used for the framework validation in **a**. The observational constraint (orange; Obs) is illustrated on the right with the horizontal black line indicating the 50th percentile (0.31 ppmv K<sup>-1</sup>). The thin and thick bars denote 90% (−0.09 to 0.69 ppmv K<sup>-1</sup>) and 66% (0.08 to 0.54 ppmv K<sup>-1</sup>) confidence intervals, respectively. The CMIP mean (median) values are 0.67 (0.53) ppmv K<sup>-1</sup> for CMIP5, 0.55 (0.49) ppmv K<sup>-1</sup> for CMIP6, 0.60 (0.52) ppmv K<sup>-1</sup> for the combined set of CMIP5/6 and 0.63 (0.59) ppmv K<sup>-1</sup> for the 27 selected models.

In addition, it is unclear if models that better match observations<sup>17,18</sup> can be trusted more in their projections because past biases do not always translate into future projections<sup>21,34</sup>.

Here we introduce a statistical-learning framework to derive an observational constraint on these uncertain model projections. We estimate high-dimensional regression functions to predict tropical lower SWV from the UTLS temperature structure (Fig. 1), given the aforementioned link between UTLS temperatures and tropical dehydration. UTLS temperatures integrate the effects of a large number of processes affecting air dehydration, either directly or indirectly<sup>8,15,18,35</sup>. Our primary interest therefore is to quantify known relationships between the UTLS temperature structure and tropical lower SWV<sup>9,36</sup> but in a novel way that allows these relationships to hold under strong climate change scenarios. This, in turn, will open up new pathways for estimating observational constraints on future projections. Our analysis will not consider SWV production from changes in stratospheric methane concentrations, which gains importance in the middle to upper stratosphere<sup>37,38</sup>. To minimize such influences on our results, we focus on the tropical lower stratosphere at 70 hPa, that is, water vapour just above the tropical cold-trap region where air dehydration takes place<sup>39</sup>.

## Learning predictive relationships from historical data

We aim to learn predictive functions  $f$ , ultimately characterized by their coefficients  $\Theta$

$$\log(q_{\text{strat}}(t)) = f(\Theta, \mathbf{T}; t, \tau_{\text{max}}) = \sum_i^{\text{lat}} \sum_j^{\text{lon}} \sum_k^p \sum_{\tau}^{\tau_{\text{max}}} \Theta_{ijk,\tau} dT_{ijk}(t - \tau) \quad (1)$$

which predict 30° N–30° S average, monthly and zonal mean SWV (specific humidity, in ppmv) at 70 hPa, mimicking frequently used indices characterizing water vapour entry rates through the tropical cold-trap region<sup>17,18,35</sup>. Hereafter, we will refer to this quantity as  $q_{\text{strat}} \cdot dT_{ijk}$  is the standard-scaled monthly mean temperature (that is, zero meaned and scaled by its own  $\sigma$  over the training period) at 5° × 5° latitude–longitude

grid points indexed by  $(ij)$  within one of  $p = 5$  atmospheric levels (250, 200, 150, 100, 70 hPa) indexed by  $k$ , covering a latitudinal range of 60° N–60° S (Fig. 1a).

$f$  predicts  $q_{\text{strat}}$  at time  $t$  with high skill and we cross-validated its performance for various regression specifications (Extended Data Fig. 1). These tests included the choice of pressure levels, latitude range and number of time lags for  $\mathbf{T}$ . Unsurprisingly, we found that predictive performance improves if we also consider temperature data from the two preceding months (that is,  $\tau_{\text{max}} = 2$  months), being reflective of the slow vertical ascent of air through the tropical tropopause layer<sup>12,35</sup>. We additionally apply logarithmic transformations to the  $q_{\text{strat}}$  data as to approximately account for nonlinearity in the  $T$ – $q_{\text{strat}}$  relationships. A central concern when learning such high-dimensional regression functions from a relatively small number of (observed) monthly samples is to avoid overfitting. To manage this issue, we here use ridge regression<sup>40</sup>, similar to an approach recently applied successfully to constrain the global cloud feedback on climate change<sup>41</sup>.

We then learn different functions  $f$  from sets of  $\mathbf{T}$  and  $q_{\text{strat}}$  time series from both observations and climate models (Fig. 1b,c). As proxies for observations, we use the Stratospheric Water and OzOne Satellite Homogenized (SWOOSH)<sup>42</sup>  $q_{\text{strat}}$  dataset and three reanalysis products for temperature and remove months when  $q_{\text{strat}}$  observations are missing or unreliable (for example, related to the Mount Pinatubo eruption; Methods). The use of multiple reanalysis products and of SWOOSH uncertainty estimates allows us to incorporate the effects of measurement uncertainty in our constraints. For climate model data, we use simulations covering the same historical period or slightly shifted (depending on data availability) from the Coupled Model Intercomparison Project phases 5 and 6 (CMIP5/CMIP6; Methods). We treat each CMIP dataset in the same way as SWOOSH by masking equivalent months. Many models do not achieve realistic amplitudes of  $q_{\text{strat}}$  variability<sup>17,21</sup>, leading to insufficiently clear signals for ridge regression to learn from. We therefore sub-select 27 models that at least approximate the SWOOSH variance (Methods). This selection is still meant to sample the model uncertainty in the  $T_{ijk}$  and  $q_{\text{strat}}$  responses across CMIP models

so that our observational constraint will be based on better estimates for the learned parameters  $\Theta$ .

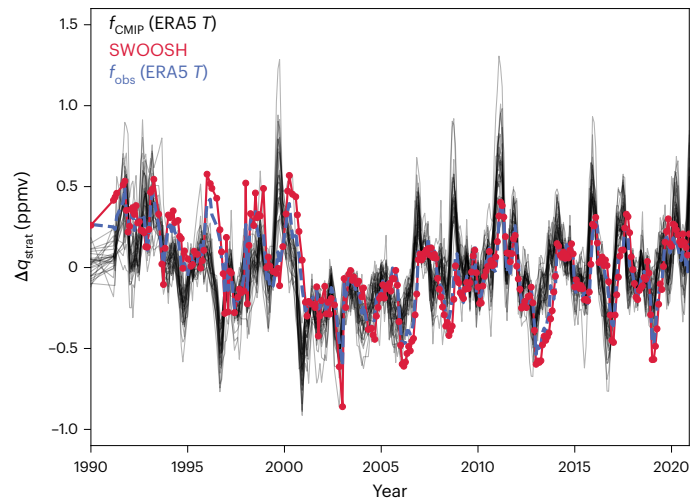
We obtained high predictive skill for each  $f$  on historical time slices not used during training and cross-validation (typically  $r^2$  scores  $> 0.8$ ). However, this is unsurprising given the central role temperature plays in setting present-day  $q_{\text{strat}}$ . More challenging is the aim to use the functions trained on historical data to predict  $q_{\text{strat}}$  responses under increased greenhouse gas forcing, that is, that the past relationships also hold in significantly warmer climates. Such climate-invariant functions open up new pathways to observationally constrain the  $q_{\text{strat}}$  response to climate change. We note that previous studies developed climate index-based multiple linear regression (MLR) methods to analyse SWV variability and trends in observations and models<sup>9,15,18,35,43,44</sup>. On the basis of small sets of indices, these valuable tools can explain large fractions of variance and have been used to infer drivers of SWV changes under climate forcing<sup>18</sup>. However, these MLR methods do not achieve the predictive performance of ridge regression under extrapolation (Supplementary Fig. 3), underlining the value of statistical learning for deriving our observational constraint. We further see advantages in exploiting only well-observed UTLS temperatures as predictors, whereas index-based methods typically require BDC metrics that are not a widely available CMIP output, allowing for evaluation of a greater number of models.

### The observational constraint

We evaluate the extrapolation idea in a perfect-model setting: the 27 functions trained on historical CMIP data are used to predict  $q_{\text{strat}}$  under abrupt- $4 \times \text{CO}_2$  forcing, using the  $\mathbf{T}_{4 \times \text{CO}_2}$  from the corresponding CMIP simulations as predictors. In Fig. 1d, we show four examples of such comparisons between statistical-learning predictions (red) and actual  $4 \times \text{CO}_2$  simulation results (grey; Supplementary Figs. 4 and 5 provide all 27 model results). To enable comparisons across models with very different climate sensitivities, we normalize annually averaged  $q_{\text{strat}}$  trends by the model-specific changes in global mean surface temperature ( $T_g$ ), as is common in climate-feedback analyses, for example, ref. 41, resulting in  $q_{\text{strat}}$  trends per degree of global warming ( $\text{ppmv K}^{-1}$ ). This choice is justified by the close coupling between UTLS temperatures and surface warming (Supplementary Figs. 4 and 5). However, as an alternative viewpoint, we provide equivalent results for a normalization by zonal mean temperatures close to the cold-trap region ( $20^\circ \text{N}$ – $20^\circ \text{S}$ , 100 hPa) in Supplementary Figs. 6–8.

Comparing the predictions to the actual CMIP results, we find excellent agreement across the multi-model ensemble ( $r = 0.90$ ; Fig. 2a), indicating that the historical  $T$ – $q_{\text{strat}}$  relationships also hold well under strong greenhouse gas forcing and opening up a path to an observational constraint in three steps: (1) Given three reanalysis datasets and  $n = 50$  iterations of SWOOSH  $q_{\text{strat}}$  time series with varying added noise patterns (Methods; to estimate sampling and measurement uncertainty), we arrive at  $3 \times 50 = 150$  statistical-learning functions  $f_{\text{obs}}$  consistent with the observational record, each with its own set of coefficients  $\Theta_{\text{obs}}$ . (2) Using these  $f_{\text{obs}}$ , we combine the uncertainty contributions introduced due to spread in the  $\Theta_{\text{obs}}$  and in the CMIP temperature responses  $\mathbf{T}_{4 \times \text{CO}_2}$  (Methods), leading to a probability distribution for the observational prediction (shown along the x axis of Fig. 2a; solid red curve). (3) Finally, this distribution is convolved (Methods) with the framework-intrinsic prediction error evident from the scatter around the one-to-one line in Fig. 2a. The result is the probability distribution characterizing the observational constraint (shown along the y axis).

Figure 2a demonstrates the large spread across the CMIP  $q_{\text{strat}}$  responses ( $0.08$  to  $1.41 \text{ ppmv K}^{-1}$ ), as compared to  $0.31 \pm 0.39 \text{ ppmv K}^{-1}$  (90% confidence interval) for our observational constraint. The large model uncertainty is illustrated even more clearly in Fig. 2b for 27 CMIP5 models (red), 34 CMIP6 models (blue), the combined CMIP5/6 ensemble (black) and for the selected 27 CMIP5/6 models (grey). Ten



**Fig. 3 | CMIP-based predictions of past variability in tropical lower stratospheric water vapour.** Black: monthly mean predictions of deseasonalized  $\Delta q_{\text{strat}}$  anomalies using the 27 CMIP-based functions provided with ERA5 reanalysis temperature data. We also show SWOOSH observational data for the same period (red). The blue line indicates average predictions conducted with the cross-validated functions learned from SWOOSH and ERA5, if ERA5 temperatures are used again as input. These predictions are highly correlated with SWOOSH ( $r^2$  score = 0.90; Pearson's  $r = 0.96$ ). The CMIP-based predictions also correlate well with SWOOSH but typically overestimate the amplitude of the undulations in line with their too-large sensitivities under climate change. Predictions with other reanalysis temperatures provide similar results (Extended Data Fig. 2).

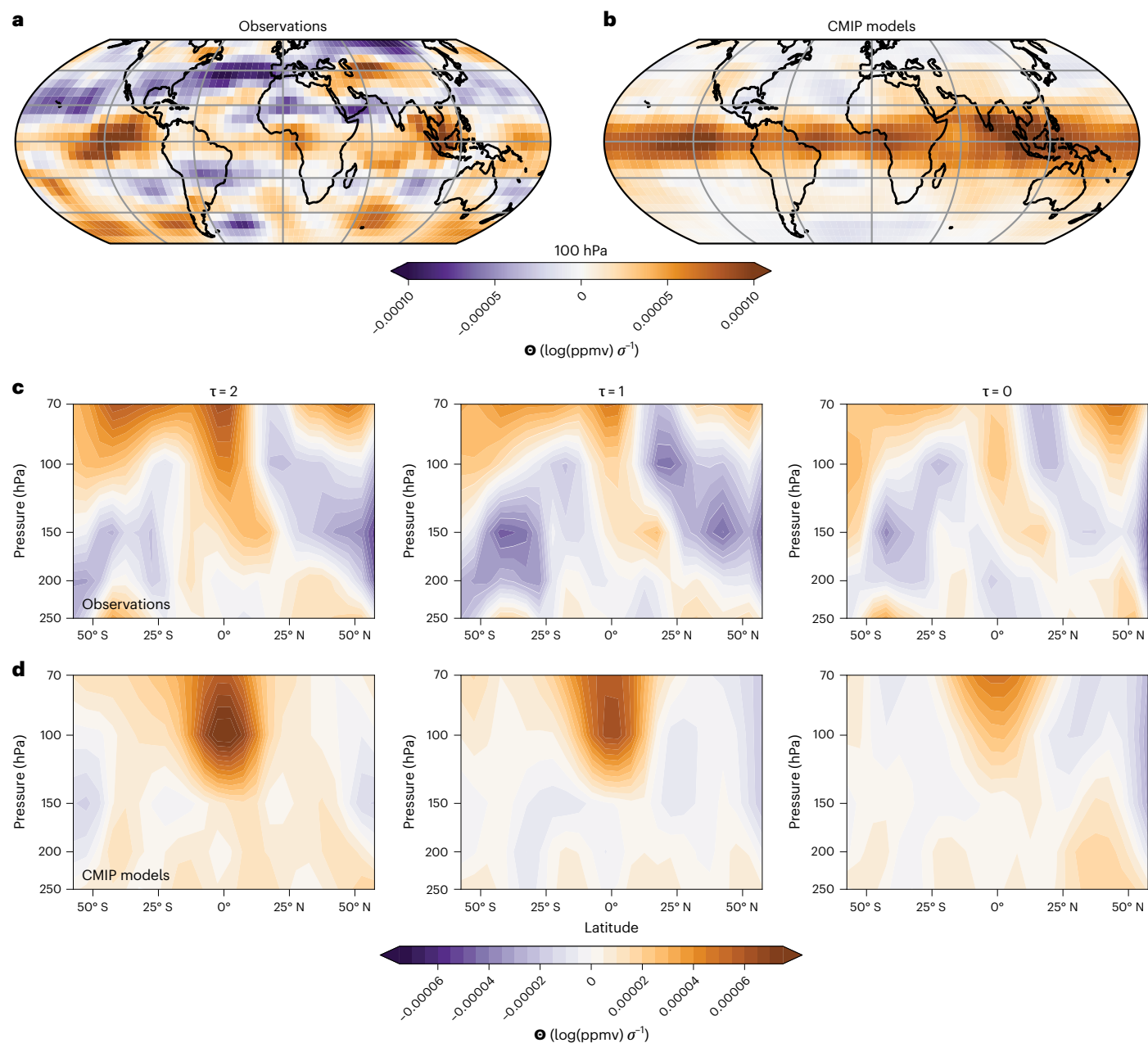
CMIP models simulate trends of approximately  $1.0 \text{ ppmv K}^{-1}$  or larger, that is, well outside the observationally plausible range (orange). While the median across all 61 CMIP models ( $0.52 \text{ ppmv K}^{-1}$ ) is still within typical uncertainty bounds, a substantial number of models are highly likely to overestimate the  $q_{\text{strat}}$  feedback under global warming. Considering typical confidence intervals for our constraint with 0.08 to  $0.54 \text{ ppmv K}^{-1}$  (17% to 83%),  $-0.09$  to  $0.69 \text{ ppmv K}^{-1}$  (5% to 95%) and  $-0.17$  to  $0.77 \text{ ppmv K}^{-1}$  (2.5% to 97.5%), we find that about one-fifth (13) of the models exceed even the 97.5th percentile of the constraint and almost half of the models (27) exceed the 83rd percentile (upper end of the thick orange bar in Fig. 2b).

Notably, the observational constraint includes small negative  $q_{\text{strat}}$  responses not seen in the models. It is probably unsurprising that negative feedbacks cannot be entirely ruled out given historically (rather) negative trends under global warming, limited sample size and possible external interferences not removed by data pre-processing (for example, remaining effects of volcanic eruptions). A negative SWV trend could also be driven by a strong BDC response, which would act to cool the tropical UTLS under  $\text{CO}_2$  forcing<sup>15,18,35</sup>. The fact that we find a robustly positive 50th percentile for the constrained response underlines our hypothesis that the framework does not merely reproduce historical trends but can indeed learn approximately climate-invariant  $T$ – $q_{\text{strat}}$  relationships from internal variability (for example, related to QBO or ENSO) instead of being negatively impacted by it.

### Emulation of the historical record and inference

We now ask if CMIP models are, in principle, able to reproduce observed  $q_{\text{strat}}$  variability. We use the 27 functions trained on CMIP data to emulate the historical record of  $q_{\text{strat}}$  anomalies, given temperature data from the reanalyses as predictors (Fig. 3 and Extended Data Fig. 2). These CMIP-based predictions of the historical record (black) are compared to SWOOSH (red). We find that the observed variations, including the sudden drop in  $q_{\text{strat}}$  in the year 2000 (refs. 25,45), are





**Fig. 4 | Interpretation of the statistical-learning results. a**, Ridge coefficients  $\Theta$  at 100 hPa and for lag  $\tau = 2$  months averaged over all 150 observational functions. **b**, The same for the multi-model mean of the 27 CMIP models. The predictor temperature data were standard scaled for each grid point (units  $\log(\text{ppmv}) \sigma^{-1}$ ). The  $\Theta$  magnitudes are therefore also directly comparable, that is, larger positive

coefficients imply a greater humidifying effect for a typical local temperature increase. **c, d**, Zonal mean  $\Theta$  for observations (**c**) and for the CMIP multi-model mean (**d**). The latitude–height cross sections illustrate an upward propagation of the sources of predictive information over time, reflective of the slow ascent of air through the tropical UTLS.

captured relatively well. The implication is that if provided with realistic UTLS temperature fields, most CMIP models would display the correct  $q_{\text{strat}}$  tendencies. However, we also highlight that the year-to-year variability of the statistical-learning predictions typically takes on amplitudes substantially larger than those observed, in agreement with our key result of overly sensitive  $T$ – $q_{\text{strat}}$  relationships already seen in their abrupt-4xCO<sub>2</sub> responses.

To detect the origin of these overestimated  $T$ – $q_{\text{strat}}$  sensitivities, we highlight the use of the statistical-learning functions for understanding model–observation discrepancies, for example, by visualizing the parameters  $\Theta$  (Fig. 4 and Supplementary Figs. 9–11). This is possible because in ridge regression the absolute magnitude of each coefficient is proportional to its estimated prediction importance, that is,

larger size coefficients imply greater importance. While interpreting these coefficient maps is non-trivial, we point out a few emerging patterns. For  $\Theta_{\text{obs}}$ , independent of the reanalysis dataset used, we, for example, find at 100 hPa (Fig. 4a) and below features suggestive of influences by tropical circulation anomalies, possibly ENSO<sup>79,18</sup>, in particular around the Maritime Continent and above the Eastern Pacific. The largest coefficients occur in a narrow band across the inner tropics (10° N–10° S) from 100 hPa upwards in agreement with the well-understood slow vertical ascent of tropical air masses through this cold-trap region (Fig. 4a,b). Zonal mean latitude–height cross sections of  $\Theta$  show a clear upward progression of predictive information over time (Fig. 4c,d), supporting the view that the functions correctly identify the main characteristics of the underlying coupling between

the large-scale circulation and tropical UTLS dehydration. Crucially, the CMIP multi-model mean  $\Theta$  (Fig. 4b,d) strongly overestimates the inner tropical relationships between  $T$  and  $q_{\text{strat}}$ , underlining the over-sensitivity of CMIP models on average. For CMIP models, the peak positive inner tropical  $\Theta$  (probably representing the tape-recorder signal<sup>36</sup>), additionally maximizes at 100 hPa without further growth with altitude, contrary to the  $\Theta_{\text{obs}}$ . We speculate that this discrepancy could be caused by the low vertical resolution of many CMIP models around the tropical tropopause. The  $\Theta$  maps also uncover a few other intriguing discrepancies, including a pattern of large negative  $\Theta_{\text{obs}}$  at 100 hPa across the North Atlantic, which is part of a general strong positive to negative, tropics to extratropics gradient in  $\Theta_{\text{obs}}$  not reproduced in the CMIP mean. However, similar, or even clearer, patterns do occur in many individual models (Supplementary Figs. 12–15). A reason might be the modulating role of the subtropical jet streams and their induced mixing barriers on tropics–extratropics SWV exchange whose strength will also depend on UTLS temperature gradients. To test such hypotheses, and to distinguish significant patterns in the coefficients from noise, we below recommend future modelling experiments to design systematically perturbed datasets to train ridge regressions on.

### Constraint on the radiative feedback and implications

In conclusion, we have derived an observational constraint for changes in tropical lower SWV per degree global warming of  $0.31 \pm 0.39$  ppmv  $\text{K}^{-1}$  (90% confidence interval). This constraint on current modelling uncertainty has important implications for the stratospheric feedback onto climate change<sup>19,20,26</sup> and for the recovery of the stratospheric ozone layer<sup>27–29</sup>. Indeed, our framework opens up new routes to the process-oriented evaluation of, and observational constraints on, state-of-the-art climate model projections. As such, we recommend its use as a complement to easily interpretable, but only analytically applicable, climate index-based regressions<sup>9,15,18,35,43</sup>.

Our observational constraint is possible only through a highly effective statistical-learning approach to estimate climate-invariant relationships between UTLS temperatures and SWV from the still very limited record of SWV observations. In a perfect-model setting, we have confirmed that these relationships also seem to hold under large  $\text{CO}_2$  forcing and are robust to the presence (or absence) of, for example, historical changes in aerosol and methane-related interferences. Our results reveal a widespread over-sensitivity in CMIP models of tropical lower SWV to changes in UTLS temperatures. In particular, our constraint implies that frequently modelled large increases per degree global warming  $> 1$  ppmv  $\text{K}^{-1}$  are highly unlikely. Strikingly, around a quarter of CMIP models exceed even the upper 95th percentile of our constraint. Given the 90% range of model responses ( $0.18$ – $1.41$  ppmv  $\text{K}^{-1}$ ), our constraint represents a 50% decrease in the 95th percentile of the climate model uncertainty distribution and a narrowing of 37% of the overall 90% range (which includes small negative responses not found in CMIP).

Whereas the 27 models exhibit a wide range of total radiative SWV feedbacks of  $0.091$ – $0.256$   $\text{W m}^{-2} \text{K}^{-1}$  (90% confidence interval, with a median of  $0.18$   $\text{W m}^{-2} \text{K}^{-1}$ ; Methods and Supplementary Table 1), we can apply our constraint on tropical lower SWV to also constrain this uncertainty by 30% to  $0.086$ – $0.201$   $\text{W m}^{-2} \text{K}^{-1}$  (median =  $0.14$   $\text{W m}^{-2} \text{K}^{-1}$ ; Extended Data Fig. 3). This estimate represents an uncertainty reduction of  $0.05$   $\text{W m}^{-2} \text{K}^{-1}$ , which is comparable to the effects of changes in biogenic volatile organic compounds or ozone (table 6.8 in ref. 46) and thus of relevance to policymakers. Our work further opens up new pathways for constraining the effects of changes in SWV on catalytic ozone-depletion cycles<sup>28,29</sup>, Arctic amplification, the North Atlantic Oscillation, the stratospheric circulation and the tropospheric jet streams<sup>24,26</sup>.

Finally, we highlight the urgent need to identify and address model-dependent root causes of the over-sensitivity, such as UTLS

temperature biases<sup>7,17,21</sup>, atmospheric chemical feedbacks<sup>19</sup>, QBO influences on UTLS temperature variability<sup>8</sup> or unrealistic diffusivity of water vapour across the tropical tropopause<sup>47</sup>. Such efforts might benefit from analyses of how a variety of stratospheric (de-)hydration mechanisms<sup>4,39</sup> affect results within our novel observational constraint framework, for example, by learning from data produced in perturbed-physics ensembles or following targeted climate model tuning<sup>48,49</sup>. In particular, processes that might not (yet) be included in climate models pose potential blind spots to our perfect-model validation approach (Fig. 2a). We also highlight that our framework could be extended to address other key uncertainty factors in stratospheric climatology and atmospheric chemistry. A non-exhaustive list includes extratropical SWV trends, especially those in the radiatively important lowermost stratosphere<sup>2,20,50</sup>, and trends in lower stratospheric ozone<sup>19</sup>.

### Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41561-023-01183-6>.

### References

- Brewer, A. W. Evidence for a world circulation provided by the measurements of helium and water vapour distribution in the stratosphere. *Q. J. R. Meteorol. Soc.* **75**, 351–363 (1949).
- Joshi, M. M. & Shine, K. P. A GCM study of volcanic eruptions as a cause of increased stratospheric water vapor. *J. Clim.* **16**, 3525–3534 (2003).
- Sioris, C. E., Malo, A., McLinden, C. A. & D'Amours, R. Direct injection of water vapor into the stratosphere by volcanic eruptions. *Geophys. Res. Lett.* **43**, 7694–7700 (2016).
- Dessler, A. et al. Transport of ice into the stratosphere and the humidification of the stratosphere over the 21st century. *Geophys. Res. Lett.* **43**, 2323–2329 (2016).
- Randel, W. & Park, M. Diagnosing observed stratospheric water vapor relationships to the cold point tropical tropopause. *J. Geophys. Res. Atmos.* **124**, 7018–7033 (2019).
- Scaife, A. A. Can changes in ENSO activity help to explain increasing stratospheric water vapor? *Geophys. Res. Lett.* **30**, 1880 (2003).
- Garfinkel, C. I. et al. Influence of the El Niño–Southern Oscillation on entry stratospheric water vapor in coupled chemistry–ocean CCM1 and CMIP6 models. *Atmos. Chem. Phys.* **21**, 3725–3740 (2021).
- Diallo, M. et al. Response of stratospheric water vapor and ozone to the unusual timing of El Niño and the QBO disruption in 2015–2016. *Atmos. Chem. Phys.* **18**, 13055–13073 (2018).
- Tian, E. W., Su, H., Tian, B. & Jiang, J. H. Interannual variations of water vapor in the tropical upper troposphere and the lower and middle stratosphere and their connections to ENSO and QBO. *Atmos. Chem. Phys.* **19**, 9913–9926 (2019).
- Butchart, N. The Brewer–Dobson circulation. *Rev. Geophys.* **52**, 157–184 (2014).
- Fueglistaler, S. & Haynes, P. H. Control of interannual and longer-term variability of stratospheric water vapor. *J. Geophys. Res. Atmos.* **110**, D24108 (2005).
- Fueglistaler, S. et al. Tropical tropopause layer. *Rev. Geophys.* **47**, RG1004 (2009).
- Riese, M. et al. Impact of uncertainties in atmospheric mixing on simulated UTLS composition and related radiative effects. *J. Geophys. Res. Atmos.* **117**, D16305 (2012).
- Schoeberl, M. R., Dessler, A. E. & Wang, T. Modeling upper tropospheric and lower stratospheric water vapor anomalies. *Atmos. Chem. Phys.* **13**, 7783–7793 (2013).

15. Dessler, A. E. et al. Variations of stratospheric water vapor over the past three decades. *J. Geophys. Res. Atmos.* **119**, 12588–12598 (2014).
16. Rollins, A. W. et al. Observational constraints on the efficiency of dehydration mechanisms in the tropical tropopause layer. *Geophys. Res. Lett.* **43**, 2912–2918 (2016).
17. Gettelman, A. et al. Multimodel assessment of the upper troposphere and lower stratosphere: tropics and global trends. *J. Geophys. Res. Atmos.* **115**, D00M08 (2010).
18. Smalley, K. M. et al. Contribution of different processes to changes in tropical lower-stratospheric water vapor in chemistry-climate models. *Atmos. Chem. Phys.* **17**, 8031–8044 (2017).
19. Nowack, P. J., Abraham, N. L., Braesicke, P. & Pyle, J. A. The impact of stratospheric ozone feedbacks on climate sensitivity estimates. *J. Geophys. Res. Atmos.* **123**, 4630–4641 (2018).
20. Banerjee, A. et al. Stratospheric water vapor: an important climate feedback. *Clim. Dyn.* **53**, 1697–1710 (2019).
21. Keeble, J. et al. Evaluating stratospheric ozone and water vapour changes in CMIP6 models from 1850 to 2100. *Atmos. Chem. Phys.* **21**, 5015–5061 (2021).
22. Shindell, D. T. Climate and ozone response to increased stratospheric water vapor. *Geophys. Res. Lett.* **28**, 1551–1554 (2001).
23. Forster, P. M. de F. & Shine, K. P. Assessing the climate impact of trends in stratospheric water vapor. *Geophys. Res. Lett.* **29**, 1086 (2002).
24. Joshi, M. M., Charlton, A. J. & Scaife, A. A. On the influence of stratospheric water vapor changes on the tropospheric circulation. *Geophys. Res. Lett.* **33**, L09806 (2006).
25. Solomon, S. et al. Contributions of stratospheric water vapor to decadal changes in the rate of global warming. *Science* **327**, 1219–1223 (2010).
26. Li, F. & Newman, P. Stratospheric water vapor feedback and its climate impacts in the coupled atmosphere–ocean Goddard Earth Observing System Chemistry–Climate Model. *Clim. Dyn.* **55**, 1585–1595 (2020).
27. Dvortsov, V. L. & Solomon, S. Response of the stratospheric temperatures and ozone to past and future increases in stratospheric humidity. *J. Geophys. Res. Atmos.* **106**, 7505–7514 (2001).
28. Stenke, A. & Grewe, V. Simulation of stratospheric water vapor trends: impact on stratospheric ozone chemistry. *Atmos. Chem. Phys.* **5**, 1257–1272 (2005).
29. Rosenlof, K. H. Changes in water vapor and aerosols and their relation to stratospheric ozone. *C. R. Geosci.* **350**, 376–383 (2018).
30. Madronich, S. et al. Changes in air quality and tropospheric composition due to depletion of stratospheric ozone and interactions with changing climate: implications for human and environmental health. *Photochem. Photobiol. Sci.* **14**, 149–169 (2015).
31. Nowack, P. J., Abraham, N. L., Braesicke, P. & Pyle, J. A. Stratospheric ozone changes under solar geoengineering: implications for UV exposure and air quality. *Atmos. Chem. Phys.* **16**, 4191–4203 (2016).
32. Hegglin, M. I. et al. Vertical structure of stratospheric water vapour trends derived from merged satellite data. *Nat. Geosci.* **7**, 768–776 (2014).
33. Millán, L. et al. The Hunga Tonga–Hunga Ha’apai hydration of the stratosphere. *Geophys. Res. Lett.* **49**, e2022GL099381 (2022).
34. Nowack, P., Runge, J., Eyring, V. & Haigh, J. D. Causal networks for climate model evaluation and constrained projections. *Nat. Commun.* **11**, 1415 (2020).
35. Dessler, A. E., Schoeberl, M. R., Wang, T., Davis, S. M. & Rosenlof, K. H. Stratospheric water vapor feedback. *Proc. Natl Acad. Sci. USA* **110**, 18087–18091 (2013).
36. Mote, P. W. et al. An atmospheric tape recorder: the imprint of tropical tropopause temperatures on stratospheric water vapor. *J. Geophys. Res. Atmos.* **101**, 3989–4006 (1996).
37. le Texier, H., Solomon, S. & Garcia, R. R. The role of molecular hydrogen and methane oxidation in the water vapour budget of the stratosphere. *Q. J. R. Meteorol. Soc.* **114**, 281–295 (1988).
38. Yu, W., Garcia, R., Yue, J., Russell, J. & Mlynczak, M. Variability of water vapor in the tropical middle atmosphere observed from satellites and interpreted using SD-WACCM simulations. *J. Geophys. Res. Atmos.* **127**, e2022JD036714 (2022).
39. Randel, W. J. & Jensen, E. J. Physical processes in the tropical tropopause layer and their roles in a changing climate. *Nat. Geosci.* **6**, 169–176 (2013).
40. Hoerl, A. E. & Kennard, R. W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67 (1970).
41. Ceppi, P. & Nowack, P. Observational evidence that cloud feedback amplifies global warming. *Proc. Natl Acad. Sci. USA* **118**, e2026290118 (2021).
42. Davis, S. M. et al. The Stratospheric Water and Ozone Satellite Homogenized (SWOOSH) database: a long-term database for climate studies. *Earth Syst. Sci. Data* **8**, 461–490 (2016).
43. Ye, H., Dessler, A. E. & Yu, W. Effects of convective ice evaporation on interannual variability of tropical tropopause layer water vapor. *Atmos. Chem. Phys.* **18**, 4425–4437 (2018).
44. Ziskin Ziv, S., Garfinkel, C. I., Davis, S. & Banerjee, A. The roles of the quasi-biennial oscillation and El Niño for entry stratospheric water vapor in observations and coupled chemistry–ocean CCM1 and CMIP6 models. *Atmos. Chem. Phys.* **22**, 7523–7538 (2022).
45. Randel, W. J., Wu, F., Vömel, H., Nedoluha, G. E. & Forster, P. Decreases in stratospheric water vapor after 2001: links to changes in the tropical tropopause and the Brewer–Dobson circulation. *J. Geophys. Res. Atmos.* **111**, D12312 (2006).
46. Szopa, S. et al. Short-lived climate forcings, in *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (eds Masson-Delmotte et al.) Ch. 6 (Cambridge Univ. Press, 2021); [https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC\\_AR6\\_WGI\\_Chapter06.pdf](https://www.ipcc.ch/report/ar6/wg1/downloads/report/IPCC_AR6_WGI_Chapter06.pdf)
47. Hardiman, S. C. et al. Processes controlling tropical tropopause temperature and stratospheric water vapor in climate models. *J. Clim.* **28**, 6516–6535 (2015).
48. Joshi, M. M., Webb, M. J., Maycock, A. C. & Collins, M. Stratospheric water vapour and high climate sensitivity in a version of the HadSM3 climate model. *Atmos. Chem. Phys.* **10**, 7161–7167 (2010).
49. Hourdin, F. et al. The art and science of climate model tuning. *Bull. Am. Meteorol. Soc.* **98**, 589–602 (2017).
50. Forster, P. M. de F. & Shine, K. P. Stratospheric water vapour changes as a possible contributor to observed stratospheric cooling. *Geophys. Res. Lett.* **26**, 3309–3312 (1999).

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023



## Methods

### Water vapour observations and their uncertainty estimates

For SWV observations, we use the global Stratospheric Water and OzOne Satellite Homogenized (SWOOSH)<sup>42</sup> dataset, which includes vertically resolved water vapour data from a subset of the limb-profiling satellite instruments operating since the 1980s. SWOOSH is designed to accurately reproduce monthly average variability present in the underlying data. We select the variable *combinedanomfillh2oq* at 68 hPa, which is an anomaly-filled zonal mean specific humidity field in parts per million volume (ppmv) at 10° latitude resolution. We spatially weight (cosine-weighting for latitudes) and average (latitude–longitude) the field for points within 30° N to 30° S to obtain a representation of tropical lower SWV.

For our analysis, we consider SWOOSH v2.7 data covering the period from January 1984 to including December 2020. The version incorporates recent improvements in Earth Observing System Aura Microwave Limb Sounder (MLS) data<sup>51</sup>. However, before the availability of Aura MLS satellite data (September 2004–present), the SWOOSH dataset has a high number of missing data in the tropics. For the anomaly-filled version of SWOOSH, missing data were filled using a procedure that made 2D latitude–time linear interpolations for each month on deseasonalized anomalies using information from adjacent grid cells for which data existed; the seasonal cycle was added back on after filling in the latitude–time plane at each pressure level<sup>42</sup>. However, the interpolation was not evaluated for potential biases introduced by the procedure so that the use of the filled product for our statistical-learning approach introduces an additional uncertainty factor.

Here we make a first-order estimate of additional biases for the pre-MLS period by using the MLS period where sampling is high and effectively unbiased by latitude and time and masking these data as for the pre-MLS period. For each month for which at least one sample exists in the 30° S and 30° N latitude band in the pre-MLS period, we identify the month of that year in all Aura MLS data, mask the data as for the month of interest and estimate the bias introduced by integrating, weighting by latitude over 30° S to 30° N where data exist, and by comparing with the ‘true’ unbiased MLS/SWOOSH value. From this, we estimate a mean and standard deviation of the bias assuming Gaussianity.

For our final uncertainty calculations, we dropped all months from the SWOOSH dataset for which there was not at least one sample measured within 30° S to 30° N, which reduces the total number of samples (months) considered from 444 to 315. In effect, this also removes all SWOOSH data before January 1990. For the remaining 315 months, we estimate the uncertainty introduced by the biases of the anomaly-filling method by sampling (in addition to the uncertainty provided by SWOOSH and assuming a normal distribution) from the standard deviation in the MLS bias estimates outlined above and adding this to a random normal sample from the standard error of the SWOOSH data itself, that is,  $\sigma/\sqrt{N}$ , latitude-weighted by the sum of the squared errors. Here  $N$  is the overall measured number of samples per monthly data point considered<sup>42</sup>. Adding these randomly drawn estimates for each month to the original filled SWOOSH time series yields sets of ‘sampled time series’. Here we use  $n = 50$  such randomly drawn time series to estimate the effects of the sampling biases on our overall uncertainty estimates. We find that the effect of sampling biases are small to negligible for the overall uncertainty estimation but we still include these error estimates in our uncertainty analysis for completeness.

### Temperature data

To approximate observations for UTLS temperatures, we use three different reanalysis datasets for temperature at 250, 200, 150, 100 and 70 hPa over the same time period: ERA5 (ref. 52), MERRA-2 (ref. 53) and JRA-55 (ref. 54). For MERRA-2, we do not include the year 2020 as the corresponding data could not be found in the archive used at the time of writing (Data Availability). For learning the observational constraint

functions, we combine each of the three reanalyses once with each of the  $n = 50$  SWOOSH randomly drawn time series, resulting in 150 functions overall. From these 150 functions, we derive a first observational uncertainty estimate on predictions under  $4 \times \text{CO}_2$  forcing by providing each function once with the modelled (and standard-scaled) monthly mean temperature profiles found under  $4 \times \text{CO}_2$  for the 27 selected CMIP models (steps also described in the main text).

### CMIP data

We consider climate model data from both the CMIP5 (ref. 55) and CMIP6 (ref. 56) archives. In total, this amounted to 61 models for which we found data for all required scenarios:

- 27 CMIP5 models: ACCESS1-0, ACCESS1-3, BCC-CSM1-1, BCC-CSM1-1-m, BNU-ESM, CanESM2, CCSM4, CNRM-CM5, CSIRO-Mk3-6-0, EC-EARTH, FGOALS-g2, GFDL-CM3, GFDL-ESM2G, GFDL-ESM2M, GISS-E2-H, GISS-E2-R, HadGEM2-ES, INM-CM4, IPSL-CM5A-MR, IPSL-CM5B-LR, MIROC5, MIROC-ESM, MPI-ESM-LR, MPI-ESM-MR, MPI-ESM-P, MRI-CGCM3, NorESM1-M.
- 34 CMIP6 models: ACCESS-CM2, ACCESS-ESM1-5, AWI-CM1-1-MR, BCC-CSM2-MR, BCC-ESM1, CAMS-CSM1-0, CanESM5, CESM2, CESM2-WACCM, CNRM-CM6-1, CNRM-ESM2-1, E3SM-1-0, EC-Earth3-Veg, FGOALS-f3-L, FGOALS-g3, GFDL-CM4, GFDL-ESM4, GISS-E2-1-G, GISS-E2-1-H, HadGEM3-GC31-LL, HadGEM3-GC31-MM, INM-CM4-8, INM-CM5-0, IPSL-CM6A-LR, MIROC6, MIROC-ES2L, MPI-ESM1-2-HR, MPI-ESM1-2-LR, MRI-ESM2-0, NESM3, NorESM2-LM, NorESM2-MM, SAMO-UNICON, UKESM1-0-LL.

An overview of all CMIP models, including individual model references and  $T_g$ -normalized  $q_{\text{strat}}$  feedback values, is provided in Supplementary Table 2. Equivalent results for normalization by 20° N–20° S temperature at 100 hPa are tabulated in Supplementary Table 3. For each model, we used variable output for 30° S–30° N average zonal mean specific humidity (hus) at 70 hPa and air temperature (ta) at 250, 200, 150, 100 and 70 hPa. To train the ridge regressions, we combined atmosphere–ocean-coupled historical simulations from 1 January 1984 onwards with Representative Concentration Pathway 4.5 (RCP4.5)/Shared Socioeconomic Pathway 3–7.0 (SSP3–7.0) scenarios. The future RCP scenarios were selected as to maximize the number of models for which we could match the observed period within either CMIP archive given that scenario differences across the period 2005 (end of historical simulations for CMIP5) to 2020 (end of observed period used here) are negligible for our calculations. The same two variables plus surface air temperature (tas) were extracted for the same set of models for the abrupt- $4 \times \text{CO}_2$  simulations. In all cases, we use only the first available ensemble member for each model.

It is well known that tropical UTLS water vapour variability is not represented well in many atmospheric models, both in terms of the timing and amplitude of the seasonal cycle and/or variations relative to it<sup>17,21</sup> (Supplementary Figs. 1 and 2). A concern of particular importance for the statistical-learning process employed here are cases where variability is substantially underestimated, because this will reduce the ability of ridge regression to learn meaningful  $T-q_{\text{strat}}$  relationships, especially if the goal is to extrapolate potentially very large abrupt- $4 \times \text{CO}_2$  responses<sup>41,57,58</sup>. We therefore include only CMIP models that represent at least 95% of the observed variance found for SWOOSH across the 315 potential training samples in our calculations (Fig. 2a) in the main text. These 27 models are:

- Six CMIP5 models: ACCESS1-0, ACCESS1-3, GFDL-CM3, MPI-ESM-LR, MPI-ESM-MR, MPI-ESM-P.
- 21 CMIP6 models: ACCESS-CM2, ACCESS-ESM1-5, AWI-CM1-1-MR, CAMS-CSM1-0, CanESM5, CESM2, CESM2-WACCM, FGOALS-f3-L, GISS-E2-1-G, GISS-E2-1-H, HadGEM3-GC31-LL,



HadGEM3-GC31-MM, INM-CM4-8, INM-CM5-0, MPI-ESM1-2-HR, MPI-ESM1-2-LR, MRI-ESM2-0, NESM3, NorESM2-LM, NorESM2-MM, UKESM1-0-LL.

Analyses equivalent to the one shown in Fig. 2 but for other choices of percentage of observed variance thresholds (0%, 50%, 80%, 90%) are provided in Supplementary Fig. 16.

For a few of the selected models, the relevant SWOOSH period (January 1990 to December 2020) could not be matched with a consistent set of simulations. Instead, we sampled equivalent months from their historical simulations only. For CMIP5, this concerns MPI-ESM-P for which we considered the period 1968–2004, amounting to the same number of samples (note that, for example, the period 1984 to 1990 is excluded according to the data mask derived from SWOOSH). For CMIP6, the following selected models are affected: CESM2, FGOALS-f3-L, GISS-E2-1-H, HadGEM3-GC31-LL, HadGEM3-GC31-MM, INM-CM4-8, INM-CM5-0, MPI-ESM1-2-HR, MPI-ESM1-2-LR, NESM3, NorESM2-LM, NorESM2-MM. For these CMIP6 models, we instead used data covering the period 1977–2013. To keep consistency with the SWOOSH record as close as possible, we applied the mask representing SWOOSH data gaps to each model dataset from 1984 onwards, which includes masking of the period immediately following the Mt. Pinatubo eruption in 1991, which could otherwise have been considered an unusual event in the CMIP data not characterized by SWOOSH<sup>42</sup>.

### Statistical-learning framework

For each CMIP model and SWOOSH/reanalysis pair of specific humidity and temperature data, we train a predictive function  $f$  (see equation (1)). The exclusion of lags or the addition of time lags longer than  $\tau_{\max} = 2$  months do not further improve the performance (Extended Data Fig. 1). To quasi-linearize the  $T$ - $q_{\text{strat}}$  relationships, we apply the natural logarithm to the specific humidity data, which also improves the overall predictive performance of the learned functions, in particular, under extrapolation (Extended Data Fig. 1c).

Here we use temperatures within 60° N–60° S at each of the five atmospheric pressure levels as predictors. Our set-up is constrained by our empirical results that extending the area of predictors to the polar regions did neither improve the predictive performance on historical test data nor on the abrupt-4 × CO<sub>2</sub> simulations (Extended Data Fig. 1d,e). However, in particular for observations, we received the best cross-validation results on historical data when using 60° N–60° S instead of only tropical (30° N–30° S) temperatures. In a classic statistical-learning set-up of training, cross-validation and separate testing, we therefore chose the best performing configuration for the historical cross-validation data also for the abrupt-4 × CO<sub>2</sub> ‘test’ scenario. We also explored the sensitivity of the extrapolation results to a longer training period (Extended Data Fig. 1f) and to the number of pressure levels at which temperature is considered as predictor (seven/three/one in Extended Data Fig. 1g,h,i). As another simplification and to speed up the learning process, we interpolated the temperature data for each CMIP model and reanalysis dataset to a common 5° × 5° (latitude × longitude) grid. This coarser spatial resolution also allows us to homogenize the predictor resolution for all temperature datasets, which is necessary to later combine different sets of temperature predictors and ridge coefficients  $\Theta$  for the observational constraint.

To estimate the coefficients  $\Theta$ , we use ridge regression<sup>40</sup>, which here minimizes the cost function

$$J_{\text{ridge}}(\Theta) = \sum_t \left( \log(q_{\text{strat},t}) - \sum_{i,j,k,\tau} \Theta_{ijk,\tau} dT_{ijk}(t-\tau) \right)^2 + \alpha \sum_{i,j,k,\tau} \Theta_{ijk,\tau}^2 \quad (2)$$

over 315 monthly mean samples indexed by  $t$ . The total number  $M$  of temperature predictors is 25,920 (5 levels × 24 latitudes × 72 longitudes × 3 months for maximum lag  $\tau_{\max} = 2$ ). This large number

of predictors, especially given the limited length of the observational record, would lead to overfitting using multiple linear regression (MLR). Next, to avoid overfitting, ridge regression is also known for its good performance in managing ill-posed problems with many col-linear predictors<sup>41,57</sup>. Note that the first term in equation (2) is the MLR least squares error, which, as discussed, tends to overfit the data given large  $M$ . Ridge regression addresses overfitting through the second  $l^2$ -norm regularization term, which penalizes large absolute values for  $\Theta$ , modulated by the choice for the regularization parameter  $\alpha$ . To approximate optimal  $\alpha$ , we use fivefold cross-validation searching over  $\alpha \in [0.0001, 0.0003, 0.1, \dots, 1 \times 10^9]$  and evaluate according to the  $r^2$  scores (coefficients of determination; ref. 58 provides a detailed explanation) as defined by Python’s scikit-learn package<sup>59</sup> across the historical validation sets. This general search range for  $\alpha$  was determined incrementally following tests showing that larger and smaller values for  $\alpha$  would never be selected during cross validation. As mentioned above, we standardize temperature time series at each grid point to zero mean and unit standard deviation (over the historical period) to ensure that they are considered equally and so that the absolute magnitudes of the resulting sensitivities are reflective of their relative physical importance<sup>57</sup>. When combining  $\Theta$  derived from SWOOSH/reanalysis pairs with CMIP temperature responses under 4 × CO<sub>2</sub>, we therefore re-scale the temperature fields according to the grid point  $\sigma$  values of the reanalysis dataset to represent the relative amplitude of the CMIP modelled temperature anomalies consistently. Due to the standard scaling of temperatures and our focus on the SWV response per degree warming, we do not carry over baseline model biases in mean values of temperature and humidity into our observational constraint.

### Calculation of framework-related uncertainty

We follow a similar approach to Ceppi and Nowack<sup>41</sup>, in which the uncertainty in the constraint is calculated in several steps. First, we obtain a probability distribution of the observational prediction ( $x$  axis of Fig. 2a; solid red curve) by combining the uncertainties in  $\Theta_{\text{obs}}$ , denoted  $\sigma_{\Theta}$ , with those due to the different CMIP 4 × CO<sub>2</sub> temperature responses,  $\sigma_T$ . For this, we first linearly combine all of the 150 estimates of  $\Theta_{\text{obs}}$  with each of the 27 CMIP  $T_{4 \times \text{CO}_2}$  fields, leading to 4,050 observationally constrained  $T_g$ -normalized  $q_{\text{strat}}$  predictions. To obtain  $\sigma_{\Theta}$ , we first take the multi-model mean over all predictions made using the same set of observed coefficients and subsequently calculate the standard deviation of these 150 samples. We follow the same procedure for  $\sigma_T$  but now averaging estimates involving the same UTLS temperature response, calculating the standard deviation of the resulting 27 estimates. These uncertainties are then combined in quadrature,  $\sigma_p = \sqrt{\sigma_{\Theta}^2 + \sigma_T^2}$ , to yield the uncertainty for the observational prediction  $q_{\text{strat},p}$ .

Next, this observational prediction uncertainty is convolved with the prediction error, calculated via standard least squares regression formulas<sup>60</sup>, whose 5–95% interval are represented by dashed red curves in Fig. 2a. This yields a probability distribution for the actual normalized  $q_{\text{strat}}$  response  $q_{\text{strat},a}$  on the  $y$  axis of Fig. 2a:

$$P(q_{\text{strat},a}) = \int_{-\infty}^{+\infty} P(q_{\text{strat},a}|q_{\text{strat},p})P(q_{\text{strat},p}) dq_{\text{strat},p} \quad (3)$$

where the conditional probability  $P(q_{\text{strat},a}|q_{\text{strat},p})$  represents the prediction error.  $P(q_{\text{strat},a})$  is calculated numerically by Monte Carlo sampling, with a sample size of  $10^7$ , and we apply a Gaussian kernel smoother to the result with a standard deviation of 0.01 ppmv K<sup>-1</sup> to obtain the final probability distribution.

### Stratospheric water vapour feedback calculation

We approximate the implications of our tropical lower SWV constraint for also constraining the overall SWV radiative climate-feedback

parameter<sup>20,46</sup>. This is justified by our empirical finding that the  $q_{\text{strat}}$  metric is highly correlated with SWV feedback parameters estimated from radiative transfer calculations (Pearson's  $r = 0.85$ ; Extended Data Fig. 3). For this purpose, we combined the feedback parameters for the six selected CMIP5 models calculated by Banerjee et al.<sup>20</sup> and ran additional calculations for the 21 selected CMIP6 models. We then followed the same regression approach as taken for the observational constraint in Fig. 2a but replacing the variable along the y axis by the SWV feedback parameters. With this procedure, we obtain an observationally constrained 90% confidence interval for the feedback parameter of  $0.086\text{--}0.201\text{ W m}^{-2}\text{ K}^{-1}$ , equalling an uncertainty reduction of  $0.05\text{ W m}^{-2}\text{ K}^{-1}$  over the 90% confidence interval ( $0.091\text{--}0.256\text{ W m}^{-2}\text{ K}^{-1}$ ) for the 27 CMIP models. We computed the SWV feedback for the CMIP6 models using the Parallel Offline Radiative Transfer programme<sup>61</sup> and following the procedure outlined in Banerjee et al. Briefly, for each model we computed the SWV change between the abrupt-4 × CO<sub>2</sub> and pre-industrial control simulations, based on the last 50 years of each simulation. We input these water vapour fields into the Parallel Offline Radiative Transfer programme to compute the stratospherically adjusted net tropopause radiative flux change for each model individually using the fixed dynamical heating approximation. Then, the SWV feedback (in  $\text{W m}^{-2}\text{ K}^{-1}$ ) is computed by dividing the tropopause radiative flux change by the global mean surface temperature change (again, averaged over the last 50 years of each simulation).

## Data availability

All observational, reanalysis and climate model datasets used in this study are publicly available. SWOOSH data can be found at <https://csl.noaa.gov/groups/csl8/swoosh/>. CMIP data were obtained from the UK Center for Environmental Data Analysis portal (<https://esgf-index1.ceda.ac.uk/search/cmip6-ceda/>). MERRA-2 data were obtained from the Collaborative REAnalysis Technical Environment (CRE-ATE) project (<https://esgf-node.llnl.gov/search/create-ip/>). JRA-55 data were downloaded from the National Center for Atmospheric Research/University Corporation for Atmospheric Research Data Archive (<https://rda.ucar.edu/datasets/ds628.1/>). ERA5 data were downloaded from the Copernicus Climate Data Store (<https://doi.org/10.24381/cds.fi7050d7>). In addition, pre-processed versions of the data used to run the calculations and source data to produce the figures in the manuscript are archived on figshare (<https://doi.org/10.6084/m9.figshare.22335712>).

## Code availability

Python Jupyter notebooks used for the data analysis and production of figures are available on the Github webpage of P.N. at [https://github.com/peernow/SWV\\_Nature\\_Geoscience](https://github.com/peernow/SWV_Nature_Geoscience).

## References

- Livesey, N. J. et al. Investigation and amelioration of long-term instrumental drifts in water vapor and nitrous oxide measurements from the Aura Microwave Limb Sounder (MLS) and their implications for studies of variability and trends. *Atmos. Chem. Phys.* **21**, 15409–15430 (2021).
- Hersbach, H. et al. The ERA5 global reanalysis. *Q. J. R. Meteorol. Soc.* **146**, 1999–2049 (2020).
- Gelaro, R. et al. The modern-era retrospective analysis for research and applications, version 2 (MERRA-2). *J. Clim.* **30**, 5419–5454 (2017).
- Kobayashi, S. et al. The JRA-55 reanalysis: general specifications and basic characteristics. *J. Meteorol. Soc. Jpn* **93**, 5–48 (2015).
- Taylor, K. E., Stouffer, R. J. & Meehl, G. A. An overview of CMIP5 and the experiment design. *Bull. Am. Meteorol. Soc.* **93**, 485–498 (2012).
- Eyring, V. et al. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci. Model Dev.* **9**, 1937–1958 (2016).
- Nowack, P. et al. Using machine learning to build temperature-based ozone parameterizations for climate sensitivity simulations. *Environ. Res. Lett.* **13**, 104016 (2018).
- Nowack, P., Konstantinovskiy, L., Gardiner, H. & Cant, J. Machine learning calibration of low-cost NO<sub>2</sub> and PM<sub>10</sub> sensors: non-linear algorithms and their impact on site transferability. *Atmos. Meas. Tech.* **14**, 5637–5655 (2021).
- Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
- Wilks, D. S. *Statistical Methods in the Atmospheric Sciences* (Academic Press, 2006).
- Conley, A. J., Lamarque, J. F., Vitt, F., Collins, W. D. & Kiehl, J. PORT, a CESM tool for the diagnosis of radiative forcing. *Geosci. Model Dev.* **6**, 469–476 (2013).

## Acknowledgements

P.N. and P.C. were supported through Imperial College Research Fellowships and the UK Natural Environment Research Council (NERC) grant number NE/V012045/1. P.C. was additionally supported by NERC grant NE/T006250/1. G.C. was supported by the Swiss National Science Foundation through the Ambizione grant number PZ00P2\_180043. M.A.D. was funded by the Deutsche Forschungsgemeinschaft (DFG), individual research grant number DI2618/1-1. B.H. was supported by the European Research Council (ERC) Synergy grant 'Understanding and modelling the Earth System with Machine Learning (USMILE)' under the Horizon 2020 research and innovation programme (grant agreement number 855187) and by the Helmholtz Society project 'Advanced Earth System Model Evaluation for CMIP' (Eval4CMIP). J.K. was supported by the UK Met Office CSSP-China programme through the POZSUM project and by the NERC-funded InHALE project (NE/X003574/1). P.N. used JASMIN, the UK collaborative data analysis facility, and the High Performance Computing Cluster supported by the Research and Specialist Computing Support service at the University of East Anglia. We acknowledge the World Climate Research Programme (WCRP), which through its Working Group on Coupled Modeling, coordinated and promoted CMIP6. We thank the climate modelling groups for producing and making available their model output, the Earth System Grid Federation (ESGF) for archiving the data and providing access and the funding agencies that support CMIP6 and ESGF. We dedicate this paper to our coauthor, colleague and friend Will Ball, who passed away in April 2022. He brought this group together, ultimately resulting in this publication.

## Author contributions

P.N. conducted the analysis and wrote the paper in discussion with P.C. The pre-processing of and uncertainty quantification for SWOOSH data was designed by W.B. in discussion with S.M.D. and P.N. W.B. also wrote the corresponding Methods section about SWOOSH, which was edited by P.N., S.M.D. and B.H. The CMIP6 radiative transfer calculations were run by S.M.D. and Y.J., supported by G.C. P.N., G.C., B.H. and M.A.D. fetched and pre-processed data used to train the statistical-learning functions. P.C., S.M.D., G.C., M.A.D., B.H., J.K. and M.J. provided feedback on the paper draft, leading to important improvements and additions such as the calculation of the constraint on the SWV feedback parameter. P.N. suggested the study.

## Competing interests

The authors declare no competing interests.

## Additional information

**Extended data** is available for this paper at <https://doi.org/10.1038/s41561-023-01183-6>.

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41561-023-01183-6>.

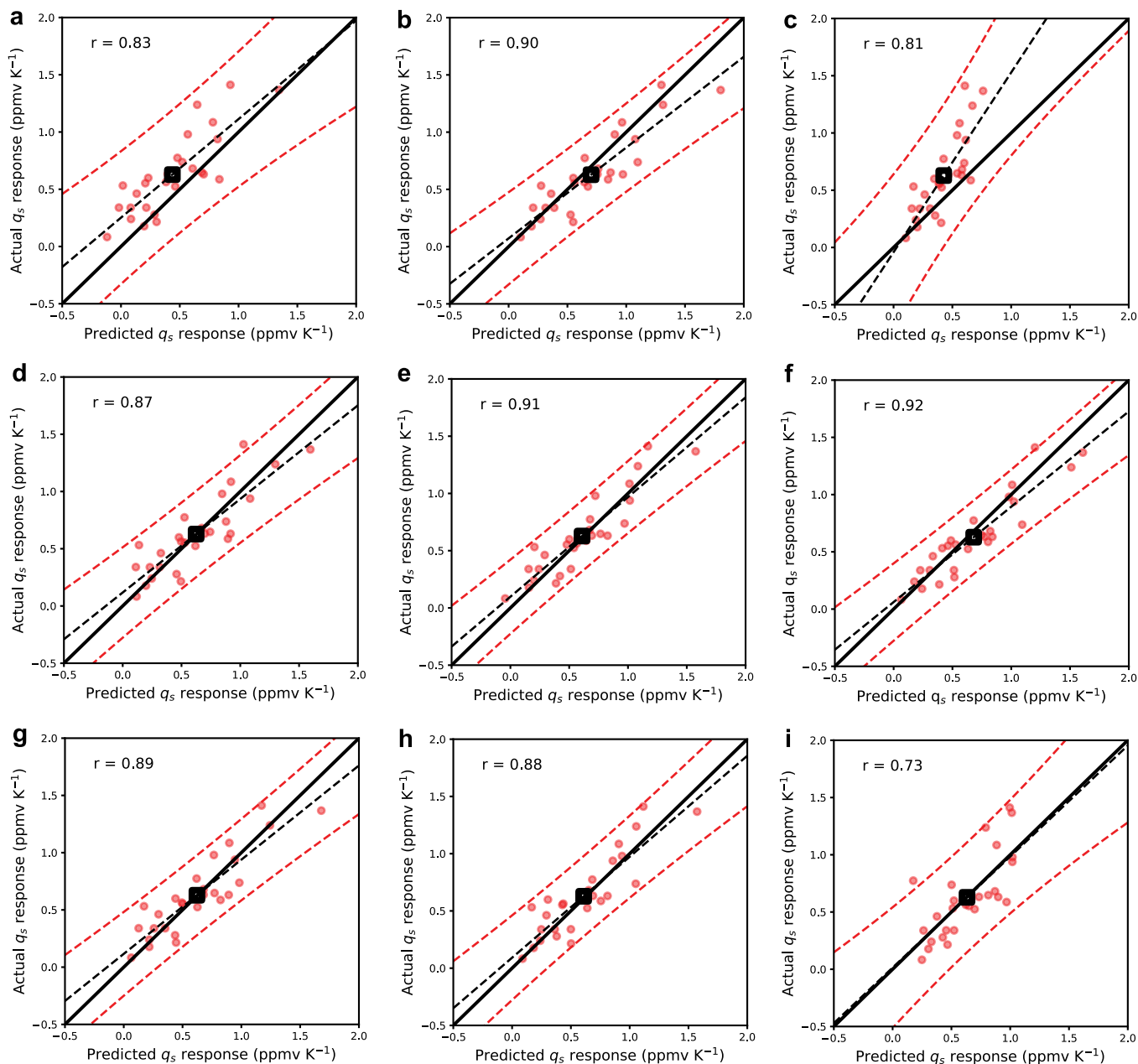
**Correspondence and requests for materials** should be addressed to Peer Nowack.

**Peer review information** *Nature Geoscience* thanks Mark Schoeberl and the other, anonymous, reviewer(s) for their contribution to the

peer review of this work. Primary Handling Editor: Tom Richardson, in collaboration with the *Nature Geoscience* team.

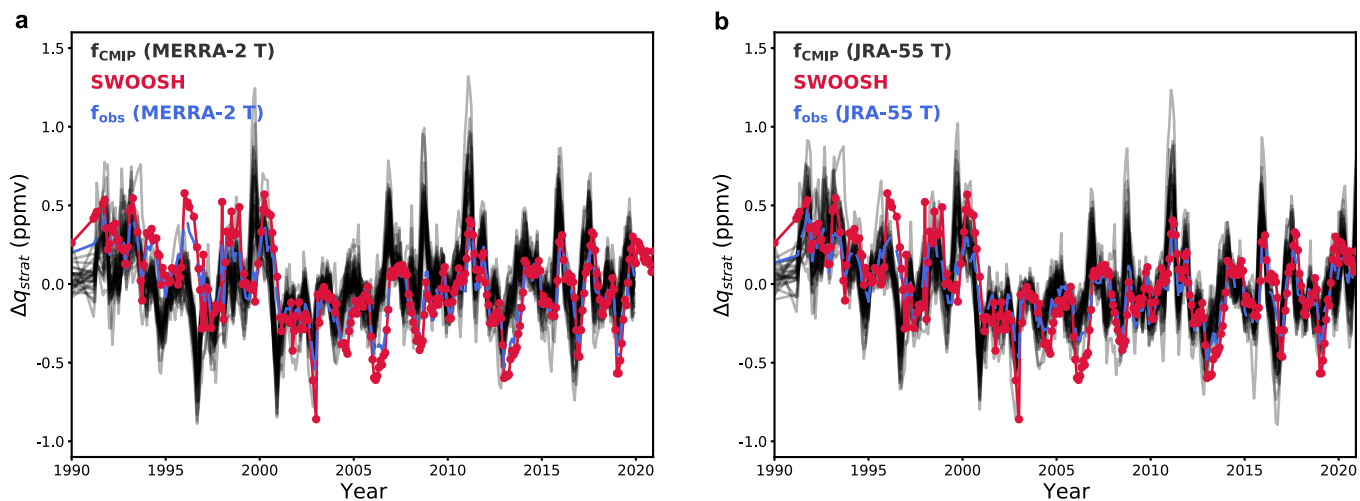
**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).





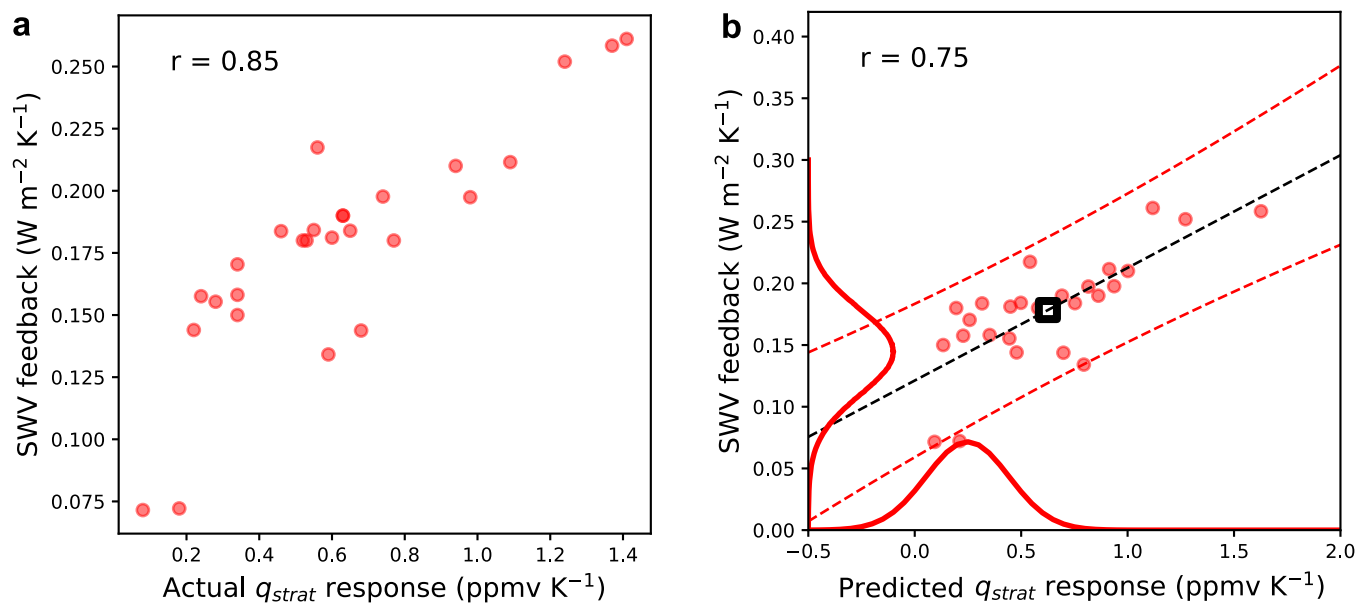
**Extended Data Fig. 1 | Framework performance depending on regression settings.** As Fig. 2a, that is red circles show abrupt-4xCO<sub>2</sub> simulation results ('actual') regressed against predicted changes in  $q_{\text{strat}}$  (here abbreviated as  $q_s$ ), both normalized by  $T_g$ , for 27 CMIP models. The multi-model-mean is indicated as a black square; the one-to-one line in solid black. Dashed lines show the least squares regression fit (black) and the 5 to 95% prediction intervals (red). The one-at-a-time differences are that in **a** no lagged temperature data was considered as predictors; in **b** one additional time lag ( $\tau_{\text{max}} = 3$ ) was considered;

in **c** we did not take the natural logarithm of  $q_{\text{strat}}$ ; in **d** temperature predictors at all latitudes were considered; in **e** temperature predictors only within 30°N - 30°S were considered; and in **f** 444 samples (months covering all years from 1984 to 2020) were used for training the CMIP functions, instead of the 315 months used in the main paper. In **g**, temperature data at seven pressure levels (300, 250, 200, 150, 100, 70, 50 hPa) were considered as predictors, whereas in **h** only three levels (200, 150, 100 hPa) and in **i** temperature only at 100 hPa was considered.



**Extended Data Fig. 2 | CMIP-based predictions of past variability in tropical lower stratospheric water vapour using two other reanalysis temperature datasets.** Black: monthly mean predictions of past  $\Delta q_{\text{strat}}$  anomalies (relative to the respective seasonal cycles), using the CMIP-based functions provided with **a** MERRA-2 and **b** JRA-55 temperature data. We also show SWOOSH observational data for the same period (red), with the dots indicating the timing of the 315 months used in our calculations. The same months were selected from MERRA-2/JRA-55 for the CMIP-based  $q_{\text{strat}}$  predictions. The blue dashed line indicates the

averaged predictions using the cross-validated ridge functions learned from the 50 combinations of SWOOSH and MERRA-2/JRA-55 data, if MERRA-2/JRA-55 is used again as the consistent input. The comparison with the SWOOSH time series (red) itself underlines that these ridge regressions represent a large fraction of the SWV variance, as evident from high  $r^2$  scores of 0.89 for **a** and 0.79 for **b** (see for example ref. 58 for a detailed explanation of this time series performance metric) and Pearson correlation coefficients of 0.96 and 0.90, respectively.



**Extended Data Fig. 3 | Constraint on the stratospheric water vapour feedback parameter.** **a** Correlations of the radiative feedback parameters for the 27 models also used in Fig. 2a in the main text against the  $q_{\text{strat}}$  metric, yielding a

high correlation. **b** As Fig. 2a, but again with the radiative feedback parameters instead, leading to an observational constraint. The final distribution is shown along the y-axis.