The DOI for this manuscript is doi: 10.1175/JCLI-D-17-0631.1

Caldwell, P., M. Zelinka, and S. Klein, 2018: Evaluating Emergent Constraints on Equilibrium Climate Sensitivity. J. Climate. doi:10.1175/JCLI-D-17-0631.1, in press.

Evaluating Emergent Constraints on Equilibrium Climate Sensitivity

Peter M. Caldwell*, Mark D. Zelinka, and Stephen A. Klein

Lawrence Livermore National Lab, Livermore CA

**Corresponding author address:* Peter M. Caldwell, Lawrence Livermore National Laboratory, 7000 East Ave, L-103 Livermore, CA 94550

E-mail: caldwell19@llnl.gov

ABSTRACT

Emergent constraints are quantities which are observable from current measurements and have skill predicting future climate. This study explores 19 previously-proposed emergent constraints related to equilibrium climate sensitivity (ECS, the global-average equilibrium surface temperature response to CO_2 doubling). Several constraints are shown to be closely related, emphasizing the importance for careful understanding of proposed constraints. A new method is presented for decomposing correlation between an emergent constraint and ECS into terms related to physical processes and geographical regions. Using this decomposition, one can determine whether the processes and regions explaining correlation with ECS correspond to the physical explanation offered for the constraint. Shortwave cloud feedback is generally found to be the dominant contributor to correlations with ECS because it is the largest source of inter-model spread in ECS. In all cases, correlation results from interaction between a variety of terms, reflecting the complex nature of ECS and the fact that feedback terms and forcing are themselves correlated with each other. For 4 of the 19 constraints, the originally-proposed explanation for correlation is borne out by our analysis. These 4 constraints all predict relatively high climate sensitivity. The credibility of 6 other constraints is called into question due to correlation with ECS coming mainly from unexpected sources and/or lack of robustness to changes in ensembles. Another 6 constraints lack a testable explanation and hence cannot be confirmed. The fact that this study casts doubt upon more constraints than it confirms highlights the need for caution when identifying emergent constraints from small ensembles.

1. Introduction

How much will our greenhouse-gas emissions warm our planet? This is a defining question of our time. The magnitude of this warming is usually characterized in terms of the equilibrium climate sensitivity (ECS), which is the global-average surface temperature response to doubling CO_2 from pre-industrial conditions and letting the planet return to equilibrium. Because the planetary response to future changes in atmospheric composition is difficult to determine based on observations of past and current climate (Collins et al. 2013), ECS is often estimated using global climate models (GCMs). Despite its importance, predictions of ECS from different GCMs vary by a factor of 2 (Flato et al. 2013) and inter-model spread in ECS has not decreased substantially over time (Charney and Coauthors 1979; Knutti and Hegerl 2008; Andrews et al. 2012; Knutti et al. 2017). Unsurprisingly, this continued uncertainty has led to a desire to identify models which are more trustworthy. A natural way to do this is to assume that models which more accurately reproduce the current climate are more likely to capture its changes correctly. Unfortunately, models which perform well for some metrics may perform poorly for others (Gleckler et al. 2008), climate predictions from skillful models do not always agree (Waugh and Eyring 2008), and ability to reproduce current climate does not necessarily imply predictive skill. Thus another popular approach (which is the focus of this paper) is to identify quantities in the current climate which have skill at predicting future changes in GCMs. Strength of correlation between predictor and predictand across an ensemble of GCMs is typically used to measure the explanatory power of a potential relationship. Observed values of current-climate predictors can then be used to choose which GCM predictions are most credible. These current-climate predictors are commonly called *emergent constraints*.

One problem with emergent constraints is that large inter-model correlations between current-climate and future-climate quantities are expected by chance in multi-model databases (Masson and Knutti 2013; Caldwell et al. 2014). As a result, emergent constraints without a solid physical basis should be viewed with skepticism. Unfortunately, most emergent constraints in the published literature lack a satisfying physical explanation. This is understandable because the climate system is complex and difficult to distill into simple physical relationships. Identifying these *potential* emergent constraints is an important and natural first step towards uncovering *real* constraints. Since the majority of recently-proposed emergent constraints imply more severe sensitivity to greenhouse gases (Klein and Hall 2015), evaluating the credibility of predictions from emergent constraints has significant societal importance.

The goal of this paper is to evaluate the credibility and independence of previously-published emergent constraints. Our sources of data are described in Sect. 2 and the constraints we test are introduced in Sect. 3. Sect. 4 provides a short primer on statistical significance of correlations before the independence of these emergent constraints is investigated in Sect. 5. In Sect. 6, a new method for decomposing correlation between ECS and an emergent constraint is introduced and used to understand the physical mechanisms underpinning the success of each tested constraint. Discussion and conclusions follow in Sect. 7.

2. Data

Model output used in this paper comes from Phase 3 and 5 of the Coupled Model Intercomparison Project (hereafter CMIP3 and CMIP5). These intercomparisons have been instrumental in making output from a variety of world-class GCMs readily available to the public (Meehl et al. 2007; Taylor et al. 2012). Effective radiative forcing values (which include not just the direct effect of CO_2 doubling, but also the impact of all responses on timescales faster than the global-average surface temperature response) for CMIP3 models are taken from Table 1 of Dufresne and Bony (2008). CMIP3 process-level feedback values are taken from Table 1 of Soden and Held (2006). For CMIP 5 models, both forcing and feedback information are taken from Table 1 of Caldwell et al. (2016). ECS is then calculated from forcing F and net feedback values λ using the equilibrated top-of-atmosphere (TOA) response to a radiative forcing perturbation:

$$\mathrm{ECS} = \frac{-F}{\lambda}. \tag{1}$$

Cloud feedbacks in the tables used for CMIP3 and CMIP5 models were computed using the adjusted cloud radiative forcing technique (Soden et al. 2004, 2008; Shell et al. 2008). Feedback terms unrelated to clouds were computed by converting the relevant physical quantities into TOA radiative perturbations using radiative kernels (Held and Soden 2000; Soden et al. 2008; Shell et al. 2008). For CMIP3, kernels were simply multiplied by the net change in the relevant physical quantity from Intergovernmental Panel on Climate Change Special Report on Emissions Scenarios (SRES) A1B simulations and normalized by global-average surface warming to obtain feedback values. For CMIP5 data, Soden et al. (2008) kernels were used to compute radiative perturbations (with respect to contemporaneous pre-industrial control climatologies) for each year of the 150 year long abrupt4xCO2 simulations. These values were then linearly regressed against corresponding global-averaged changes in surface temperature ΔT_s and feedback values are taken as the best-fit slope. This linear regression method was pioneered by Gregory et al. (2004). Forcing for CMIP5 models is computed by applying the Gregory method to net TOA radiative imbalances. CMIP3 forcings are computed following the method of Forster and Taylor (2006), which involved computing net feedback from simulations where only CO_2 was changed using the Gregory method, then using this information in conjunction with ΔT_s and TOA radiative imbalance to derive effective forcing in SRES A1B simulations.

For both CMIP3 and CMIP5 ensembles, the data used in this study are computed without running experiments to equilibrium. Armour et al. (2013) and Rose et al. (2014) showed that the strength of the net feedback depends on the background climate state. In particular, ECS estimates tend to increase as model runs are extended (Williams et al. 2008; Winton et al. 2010; Andrews et al. 2012; Andrews et al. 2015). In Supplementary Figure 1 we test the impact of temporal variation in net feedback by repeating some of our analysis using just the first 20 years of each 4xCO_2 run, by using just years 21-150, and by using all years between 1 and 150. This figure shows that changes in simulation period have little effect on our results. Because net feedback is likely to continue changing beyond the 150 years evaluated here, our ECS estimates are probably best described as 'effective climate sensitivities' which are underestimates of the true ECS. In spite of the approximate nature of these values, the difference between equilibrium and effective climate sensitivity is probably a second-order effect (as suggested by Fig. 2 from Andrews et al. 2015) and simulations that would allow us to compute something more akin to 'true' ECS (e.g. coupled 2xCO_2 simulations extending thousands of years) are not available for most CMIP5 models.

ECS values from CMIP3 simulations run to equilibrium with fixed deep-ocean heat transports and a shallow 'slab' ocean layer are available from Table 8.2 of Randall et al. (2007); these slab ocean ECS values are somewhat different (correlation between slab and SRES A1B ECS values is 0.63) but switching datasets does not change any of our conclusions. We use SRES A1B values for F and ECS to maintain consistency with process-level feedback values, which are not available for slab runs.

CMIP3 and CMIP5 data differ in several important ways. First, water vapor-feedback for CMIP3 data was computed as the TOA radiative impact of change in specific humidity while for CMIP5 data water-vapor feedback was computed as the TOA radiative impact of *relative* humidity (RH) change (as advocated by Held and Shell 2012). This change in definition requires compensatory changes in Planck and lapse-rate feedbacks. Using fixed-RH feedbacks has little impact on inter-model differences of the Planck feedback (which are small regardless of how they are calculated) but reduces the strong anti-correlation between water vapor and lapse rate found in earlier studies. Additionally, CMIP3 calculations are done on runs where both greenhouse gases and aerosols are varying in time, while CMIP5 simulations test only the impact of greenhouse gas changes. These differences in treatment of CMIP3 and CMIP5 model output force us to consider CMIP3 and CMIP5 models separately in our decomposition. For further details about how feedbacks, forcing, and ECS are calculated for each ensemble, consult the original data sources cited above.

3. Survey of Potential Emergent Constraints Studied

In this section we provide a short overview of each of the 19 proposed emergent constraints analyzed in this paper. For each constraint, we provide:

1. a description of the constraint (also summarized in Table 1 for quick reference)
2. the proposed explanation for why this constraint is a good predictor of ECS
3. an a priori expectation of the sign and magnitude of correlation between the predictor and ECS
4. an initial evaluation of each constraint based on previous literature and correlations computed for this study (summarized in Table 2)

Throughout the paper, each constraint is identified by the last name of the first author on the first study proposing it and constraints are described below in the order they were published.

Most constraint data used here come directly from the studies introducing that constraint. Because not all models used in these previous studies provide information necessary for our decomposition, we also provide correlations in Table 2 computed using the subset of models which provide all data we need. Correlation with a subset of models provides a weak sense of the robustness of our conclusions; testing on new ensembles would provide a more rigorous test. Because the first 5 studies we consider were published before CMIP5 data was available, we are able to test them against data they weren't trained on by computing these constraints ourselves. Constraints that persist across ensembles are unlikely to occur by random chance, though it is worth mentioning that models used in CMIP5 are modified versions of models used in earlier intercomparisons (Pennell and Reichler 2011; Knutti 2010; Knutti et al. 2013, and references therein), so successive CMIP ensembles are not themselves completely independent. It is also worth noting that a real constraint may be present in one ensemble but not in another if the models used in those two ensembles were structurally different. For example, an emergent constraint might be detected in CMIP5 but not CMIP3 if it resulted from a process which was added for the first time in CMIP5 models. Alternatively, a constraint might appear in CMIP3 but not CMIP5 if all developers worked to make sure their models satisfied a constraint identified in CMIP3, thus getting rid of all spread in that predictor in CMIP5. While both of these scenarios are possible in theory, it is hard to imagine how model changes between CMIP3 and CMIP5 would affect any of the 19 constraints considered. As a result, we use reproducibility of a constraint across ensembles as a measure of their credibility.

This study gathers together more previously-proposed constraints than any single previous study, but it is not itself exhaustive. Other studies were omitted because we weren't aware of them while writing this paper, because they have already been shown to not be robust to changes in ensemble (e.g. Klocke et al. 2011), because they propose more constraints than our analysis can handle (Huber et al. 2011), or because computing them for CMIP5 models was too technically challenging given our available time (Shukla et al. 2006; Webb et al. 2015). Our scope is also limited by our focus on ECS, which precludes studies focused on other aspects of the climate system (e.g. Hall and Qu 2006; Cox et al. 2013). Defining emergent constraints relative to specific feedbacks rather than to a more integrative quantity like ECS would perhaps be preferable because it makes articulating a clear physical explanation for emergent relationships easier (Klein and Hall 2015). Furthermore, because the climate system is so complex, it is hard to believe that a single physical mechanism exists which can explain most of the inter-model spread in climate sensitivity (and therefore have very large correlation with ECS). Nonetheless, constraints on ECS are worth pursuing because they have the most value at reducing climate change uncertainty. Constraints on an individual feedback may be easier to find, but their practical utility is limited if that feedback does not project strongly onto ECS. We include Qu et al. (2013) in our study even though it wasn't previously tested on ECS because its mechanism (tropical low clouds) is known to be important for ECS. We also tested the constraints proposed in Gordon and Klein (2014) and McCoy et al. (2016), which both target high-latitude clouds, but ultimately omitted them from this study because they were poorly correlated with ECS; we take this to mean that only constraints on tropical clouds have a strong impact on ECS.

a. Covey

Covey et al. (2000) and Knutti et al. (2006) suggest that the strength of the hemisphere-averaged seasonal cycle of surface temperature may be a good proxy for the sensitivity of the planet to greenhouse gas changes because both are climate responses to radiative forcing changes. Models with a larger seasonal cycle are therefore theorized to respond have a stronger response to CO_2 increase. Because surface air temperature is controlled by many factors, some (like ocean circulation) occurring on timescales longer than a single season, this constraint is likely to be relatively weak.

Because Covey et al. (2000) used data from the CMIP1 archive and Knutti et al. (2006) used data from an ensemble of simulations using a single GCM with perturbed tuning parameters, we compute our own Covey values for CMIP3 and CMIP5 ensembles. For each model, we compute the Covey value by taking the northern-hemisphere average of the climatological surface temperature difference between January and July minus a similar quantity defined over the southern hemisphere. Climatological averages are computed using all available data from 20c3m and historical simulations (for CMIP3 and CMIP5 models, respectively). As in Covey et al. (2000), no attempt was made to correct for drift. As in all computations performed for this paper, computed values are the average over all available ensemble members. Sufficient data (including information to compute surface temperature, ECS, and F and λ components for our decomposition) was available for 12 CMIP3 models and 27 CMIP5 models. Covey et al. (2000) found a correlation of +0.4 between ECS and their constraint for 17 CMIP1 models; we find correlations of -0.36 and +0.35 for CMIP3 and CMIP5 data (respectively). Lack of consistency between ensembles suggests that the Covey constraint may not be robust, but the size of each sample is small (a problem with all statistical studies based on the CMIP archive) and the correlation we are seeking is weak, so false negatives are possible. As noted in Fasullo et al. (2015), perturbed physics ensembles (which typically have many more samples) may be more appropriate for teasing out small correlations like this. Unfortunately, relationships from perturbed physics ensembles often do not generalize to other collections of models (Sanderson 2011; Klocke et al. 2011; Masson and Knutti 2013).

b. Volodin

Volodin (2008) found a strong correlation in CMIP3 models between ECS and the gradient in total cloudiness between the tropics (taken to be between 28°N to 28°S latitude) and southern midlatitudes (between 36°S to 56°S) for years 1980-2000. He hypothesized that cloud response to climate change may be governed by the same mechanisms that cause cloud fraction to decrease with increasing sea surface temperature (SST) as one moves equatorward. This means that models with stronger (more negative) latitudinal cloudiness gradients will have higher ECS. Volodin's logic seems dubious because latitudinal variations in cloudiness are affected not only by local SST but also by the large-scale circulation. Nonetheless, when we compute Volodin values for the CMIP5 archive, we find that strong negative correlation is maintained (Table 2). Because the Volodin constraint wasn't trained on the CMIP5 dataset, this is a strong test of constraint robustness. A modern variant on the Volodin approach is described in the Siler section below.

c. Trenberth

The southern-hemisphere averaged TOA energy balance between 1990-2000 was found to be correlated with ECS in CMIP3 models by Trenberth and Fasullo (2010). Their explanation is that models tend to predict increased cloudiness (negative cloud feedback) over the southern ocean in a warmer climate, but that is only possible because these models strongly underpredict the extremely high observed cloud fraction in this area. Models with more realistic clouds (and hence less positive TOA radiative imbalance) are expected to have less cloud increase in this area and correspondingly higher ECS. When we calculate Trenberth values for CMIP5 data and compute the resulting correlation with ECS, we get a negligibly small value. Grise et al. (2015) performed a similar calculation and arrived at the same conclusion. Upon further investigation, Grise and coauthors found that correlation between southern-hemisphere TOA radiation and ECS in CMIP3 models came as much from subtropical stratocumulus/trade-cumulus areas as from the southern ocean. Further, connection between the southern ocean and ECS was found to only occur in models with excessively-reflective present-day subtropical clouds (which includes most CMIP3 models but only half of the CMIP5 models). Connection between southern-ocean and subtropical clouds seems to be an artifact of tuning (Grise et al. 2015; Kay et al. 2016; McCoy et al. 2016). Because southern-ocean TOA radiation biases were not found to be well-correlated with ECS in the full set of CMIP5 models and because the physical explanation for such a correlation is unclear, Grise et al. (2015) conclude that southern-ocean TOA biases are not a valid emergent constraint. They conclude instead that southern-hemisphere TOA radiation is correlated with ECS primarily through stratocumulus-to-trade-cumulus transition regions, which have greater scope for cloud reduction when they are more extensive in the current-climate.

d. Fasullo M and D

In Fasullo and Trenberth (2012), the authors correlated May-Aug. zonal-mean present-day RH from 1980-2000 against ECS for CMIP3 models and identified the two regions of largest correlation. One of these regions (denoted D) lies in the sub-tropical mid-tropospheric dry zone between approximately 20°S to 8.5°S and 440 to 350 mb. The other region (denoted M) lies in the moist convective region between 1.5°S and 10°N latitude and 740 mb to 570 mb. The physical mechanisms governing these correlations are unclear, so it is impossible to make an a priori prediction of the sign or magnitude of these correlations. Because correlation with ECS was only computed for CMIP3 models in the Fasullo paper, we compute our own values of the Fasullo metrics for the 9 CMIP3 models and 23 CMIP5 models with sufficient data. Our correlations of M and D with ECS are also very similar to Fasullo and Trenberth (2012) values for CMIP3 data but have very weak magnitude when applied to CMIP5 data. This surprising result can be confirmed and understood by comparing Fig. 3 and Fig. S4 from Fasullo and Trenberth (2012). These figures show the correlation between ECS-like quantities and climatological- and zonal-average RH as a function of latitude and height for CMIP3 and CMIP5 models, respectively. While it is true that the general structure of these plots look similar, the M zone of positive correlation has completely disappeared in the CMIP5 plot and the region of negative correlation in the subtropics has shifted towards the surface and has weakened relative to Fasullo's D region. Correlations in Fasullo and Trenberth (2012) Fig. S4 over the M and D boxes as defined in that paper are consistent with the values reported in our Table 2. Thus while patterns of RH over the entire tropics (as advocated by Su et al. 2014, which is described later) may end up being a useful predictor of climate change, the specific regions identified by Fasullo are almost certainly spurious.

e. Qu

Qu et al. (2013) show that global-warming induced changes in low cloud cover (LCC) in subtropical stratocumulus regions can be predicted according to

$$\Delta \mathrm{LCC} = \frac{\partial \mathrm{LCC}}{\partial \mathrm{SST}} \Delta \mathrm{SST} + \frac{\partial \mathrm{LCC}}{\partial \mathrm{EIS}} \Delta \mathrm{EIS} \tag{2}$$

where EIS is estimated inversion strength, Δ is the climate change signal, and $\partial\mathrm{LCC}/\partial\mathrm{SST}$ and $\partial\mathrm{LCC}/\partial\mathrm{EIS}$ are computed from current-climate interannual variability using bivariate linear regression. This can be related to ECS by dropping the second term in Eq. 2 (because Qu et al. (2013) found it to be less important) and dividing through by SST:

$$\frac{\Delta \mathrm{LCC}}{\Delta \mathrm{SST}} \approx \frac{\partial \mathrm{LCC}}{\partial \mathrm{SST}} \tag{3}$$

If ΔSST in stratocumulus regions is taken as a proxy for ΔT_S, ΔLCC is used as a proxy for cloud feedback, and cloud feedback is itself used as a proxy for ECS, $\partial\mathrm{LCC}/\partial\mathrm{SST}$ from current-climate variability could be a good emergent constraint for ECS. Each link in our chain of logic is imperfect, but subtropical stratocumulus are known to be a key factors for climate sensitivity (Bony and Dufresne 2005) so we consider this potential constraint worth testing. We do this using $\partial\mathrm{LCC}/\partial\mathrm{SST}$ values for CMIP3 and CMIP5 taken directly from Qu et al. (2013). Because more positive $\partial\mathrm{LCC}/\partial\mathrm{SST}$ means more shortwave reflection to space and hence smaller climate sensitivity, we expect (and find in Table 2) that ECS is negatively correlated with the Qu constraint.

Bretherton and Blossey (2014) provide a physical explanation for the Qu result based on large-eddy simulations (LES): warmer temperatures increase BL cloud-layer humidity fluxes for a given liquid water path, which increases cloud-top entrainment drying and hence reduces BL cloud mass and fraction. Because this mechanism operates on timescales much shorter than the variability sampled by Qu et al. (2013), short- and long-term behavior should be identical where this mechanism is dominant. Proving that the LES-based Bretherton and Blossey mechanism also explains the timescale invariance found in much coarser/cruder GCM simulations analyzed by Qu et al. (2013) is important future work.

f. Klein TCA and ctp-tau

Klein et al. (2013) provide metrics of model skill at reproducing present-day total cloud amount (TCA) and combined cloud top pressure and optical depth (ctp-tau) which are strongly correlated with cloud feedback λ_{Cld} and particularly shortwave cloud feedback $\lambda_{\mathrm{SW\,Cld}}$ in Cloud Feedback Model Intercomparison Project (CFMIP) phases 1 and 2 models (which correspond roughly to CMIP3 and CMIP5 models, respectively). Strangely, while correlations with cloud feedback are high for CFMIP1 and CFMIP2 ensembles individually, combining ensembles results in much worse correlation. Although not mentioned in Klein et al. (2013), Klein TCA and ctp-tau measures were found to be strongly correlated with ECS in CFMIP2 models but not CFMIP1 models. Inconsistent results in different ensembles suggest these constraints may be spurious (as noted by Klein et al. 2013). Klein constraints are only available for CFMIP models because they require cloud simulator output. We only have data from 9 CMIP5 models for these constraints because not only are we limited to CFMIP models, but we are further limited to models which provided sufficient data for computing feedback and forcing terms to the CMIP archives. Small sample size limits the reliability of results using the Klein constraints. Since no physical explanation for predictive skill by this constraint has been provided, we have no a priori guess as to the sign of this correlation. For the models available, both Klein metrics are among the strongest emergent constraints on ECS tested.

g. Su

Su et al. (2014) shows that changes in tropical clouds can be predicted by changes in the Hadley circulation in which they are embedded. They find that the quality of a model's representation of the present-day Hadley circulation is a good predictor of its ECS value. While it makes sense that cloud (and hence ECS) changes would follow Hadley cell changes, the linkage between a model's representation of the present-day Hadley circulation and its future change is unclear. In particular, Fig. 1 of Su et al. (2014) suggests that the relationship between the mean state and future changes in the Hadley cycle is complicated. This missing piece precludes an a priori prediction for the strength or sign of the Su constraint. The Su constraint is computed by calculating zonal average profiles from the surface to 100 mb of cloud fraction and RH between 45°S to 40°N for both model output and observations, then calculating measures of model quality by either taking the slope of the regression between modeled and observed profiles for each latitude and averaging over latitudes or by computing the spatial correlation between modeled and observed values. Metrics defined with respect to RH or cloud fraction and using the slope or spatial correlation to calculate error provide similar skill and emergent constraint decomposition information, so we use the regression slope of the RH metric (chosen because it has greatest skill) for the remainder of this paper. Despite the fact that Su et al. (2014) only reports results for CMIP5 data, we do not compute this constraint for CMIP3 data because the calculation is complicated and requires observational data which we do not have readily available.

The Su constraint is very similar to an earlier proposal in Volodin (2008), who noted that the error in zonally-averaged RH over certain regions in the tropical mid-troposphere and BL is well correlated with ECS in CMIP3 models. We do not analyze the Volodin RH constraint here because its methodology is unclear and it involves observational datasets we don't have available. If the region of calculation for the Volodin constraint is functionally equivalent to that used by Su and the observations used in both studies are compatible, then the Volodin and Su studies may be taken together as evidence that the Su constraint is valid in both CMIP3 and CMIP5 datasets.

h. Sherwood D, S, and LTMI

Sherwood et al. (2014) provide 3 indices of lower-tropospheric mixing in the current climate which are correlated with ECS. Because direct measures of lower-tropospheric mixing are not available for most models in the CMIP archive, these indices are somewhat indirect. The first index (called S) is meant to measure mixing between the BL and the lower troposphere in the convective parameterizations active in the ascending branch of the tropical overturning circulation. It is calculated as the average of the vertical gradients between 700 and 850 mb of RH and temperature (normalized to receive equal weight and signed so smaller gradients make S more positive) averaged over the West Pacific warm pool. Because temperature and moisture typically decrease with height and mixing moves heat and moisture upward in this region, S becomes more positive as mixing between the BL and lower free troposphere increases. The second index (called D) is framed in terms of vertical differences in resolved-scale vertical velocity with height, so it captures resolved-scale mixing. It measures the fraction of BL air in ascending columns in the tropical east Pacific and tropical Atlantic that leaves the column in the mid-troposphere rather than in the upper-troposphere. The third index, called the Lower-Tropospheric Mixing Index (LTMI) is simply the sum of S and D.

There are several pieces to the physical explanation for correlation between ECS and S, D, or LTMI. First, global-average precipitation and evaporation must be equal on multi-year timescales (because the atmosphere's ability to stockpile moisture is very limited) and are expected to increase by about 2% for each degree C of T_S warming (Held and Soden 2006). Next, most of this precipitation comes from deep convection (taken here to mean both parameterized and resolved-scale events which reach from the BL to the upper troposphere), while shallow circulations (such as those captured by S, D, and LTMI) tend to contribute little to surface precipitation. As a result, deep convective precipitation is constrained by the global water and energy budget to balance surface latent heat flux changes, but shallow convection is not. BL ventilation by shallow convection can be computed as the product of the shallow-convective ventilation rate and the BL specific humidity. BL specific humidity is expected to increase at ~7-8%/K following the Clausius-Clapeyron relationship, so if the shallow convective mixing rate stays constant as the climate warms (which remains to be proven), BL ventilation will increase in the future with corresponding reductions in BL clouds. In this case, models with larger S, D, and LTMI should have stronger reductions in future BL cloudiness and correspondingly larger ECS. This yields an a priori expectation that the correlation between the Sherwood constraints and ECS is positive. Sherwood values in Table 2 match this expectation. Kamae et al. (2016) found that LTMI explained low cloud feedback but not ECS in a perturbed physics ensemble. This provides some suggestion that the Sherwood constraints may not be robust to change in ensemble.

i. Brient Cloud Shallowness

Brient et al. (2015) builds upon Sherwood et al. (2014) by noting that while strengthening of shallow convective drying acts to decrease BL clouds as the planet warms, reductions in BL turbulent moisture flux are also important. Brient and coauthors argue that inter-model spread in both of these quantities are needed to fully explain future changes in shallow convective cloudiness. They use the fraction of clouds below 850 mb which are also below 950 mb in current-climate tropical (30°S to 30°N) weakly-subsiding (pressure velocity between 10-30 mb day^{-1}) ocean regions as a proxy for these effects. Models with higher values of this shallowness index in the current climate have stronger influence by convective drying relative to turbulent moistening and are thus expected to have larger reductions in future clouds. While Brient et al. (2015) provides a more complete explanation for cloud changes in shallow-convective areas, it explains about half as much ECS variance as Sherwood LTMI in our study (0.38 vs 0.65, see Table 2).

j. Zhai

Zhai et al. (2015) found that the seasonal response of boundary layer cloud fraction to sea surface temperature (SST) in subsidence regions over the ocean between 20°and 40°latitude in both hemispheres is a strong predictor of ECS in a combination of CMIP3 and CMIP5 models. This constraint is very similar to that of Qu et al. (2013), but uses regions less focused on stratocumulus and generally further poleward, targets seasonal instead of interannual variability, and does not remove the component of cloud response due to EIS changes before computing $\partial\mathrm{LCC}/\partial\mathrm{SST}$. Nonetheless, the physical explanation for this mechanism is identical to that for Qu so we expect the Zhai constraint to be negatively correlated with ECS. Of the 27 models included in Zhai et al. (2015), 24 have sufficient feedback and forcing information for our analysis. Correlations using our subsets of models and separating CMIP3 and CMIP5 yield correlations similar to those reported in Zhai et al. (2015) (see Table 2).

k. Tian

Tian (2015) found the amplitude of erroneous convergence and deep convection in the southeast Pacific (the so-called 'double-ITCZ' bias common in GCMs) to be correlated with ECS in a combination of CMIP3 and CMIP5 models. Formally, the Tian constraint is defined as the annual mean precipitation averaged over the box from 0-20°S and 100-150°W. This relationship lacks a solid explanation. The authors do note that Hwang and Frierson (2013) found that models with stronger southern ocean cloud biases tended to have a stronger double ITCZ (though Kay et al. 2016, find this relationship to only hold in models with fixed SST); combining Hwang and Frierson (2013)'s result with the Trenberth constraint, one might predict that a stronger double ITCZ and stronger cloud increases over the southern ocean in the future (and correspondingly weaker ECS) may both be symptoms of underprediction in southern hemisphere clouds. If this was the case, ECS should be negatively correlated with the strength of the double ITCZ across models. Tian also cites Hirota and Takayabu (2012) as finding that slowdown of the Hadley circulation is stronger in models with weaker double ITCZ bias. If this is the case, we might expect the Tian and Su constraints to be related. Tian data is already available for a wide variety of CMIP3 and CMIP5 models so we do not calculate our own values. Unsurprisingly, our correlations between ECS and the Tian constraint are similar to the value from his paper.

l. Brient Cloud Albedo

Brient and Schneider (2016) find that deseasonalized current-climate shortwave cloud albedo response to SST variations in tropical oceanic low clouds regions (defined as the 25% of ocean grid cells between 30°N and 30°S with driest 500 mb relative humidity) is negatively correlated with ECS in CMIP5 models. This is essentially a variant on the Qu et al. (2013) mechanism using a different region and measure of cloudiness, so we expect it to be correlated with Qu and Zhai constraints. That correlation is shown in Sect. 5 to be strong.

m. Lipat

Lipat et al. (2017) find that the present-day latitude of the southern edge of the Hadley cell in austral summer is a good predictor of ECS in CMIP5 models. Their argument is based on shortwave cloud radiative effect changes in the lower mid-latitudes (roughly between 28°and 48°S latitude). Models whose Hadley cell does not extend far into this region experience a large decrease in shortwave cloud radiative effect as the Hadley cell expands, replacing very cloudy midlatitude conditions with a less cloudy subtropical regime. Models whose Hadley cell already extends far into the lower mid-latitudes see less change because most of the radiatively-sensitive area is already filled with subtropics-type clouds. As a result, we expect Hadley-cell edge latitude (signed so further south is more positive) to be negatively correlated with ECS. This is borne out in Table 2. Because both Lipat and Su constraints are both related to Hadley-cell representation in models, one might expect them to be related.

n. Siler

Siler et al. (2017) generalizes upon the Volodin (2008) finding that inter-model differences in ECS are well-predicted by the latitudinal gradient of present-day cloudiness by showing that λ_{cld} is negatively correlated with cloud albedo in regions of SST>27°C and is positively correlated with cloud albedo in regions with SST<14°C, with correlation blending smoothly between positive and negative values in the intervening SST range. Regions of positive and negative correlation are hypothesized to be tied together by the need to tune global-average cloud albedo to match observations, which means that models with little cloud in warm regions are forced to compensate by also having too much cloud in cold regions. Like Volodin (2008), Siler and coauthors argue that if cloud albedo depends on SST in a climate-invariant way and the region of warm SST expands in the future, present-day cloudiness could inform future changes. Further justification for positive correlation in cold-SST regions is provided by McCoy et al. (2015), who notes that models whose clouds glaciate at higher temperatures in the current climate tend to have more negative cloud feedback because warming increases cloud liquid, which is brighter and more long-lived. Justification for negative correlation in warm-SST regions is taken from Zhao (2014), who use convective precipitation efficiency to understand present-day cloudiness and its changes in the future. This constraint is claimed to be independent of other constraints for subtropical low clouds because it operates in both ascending and descending regions and operates at all levels in the troposphere.

Siler et al. (2017) distill their geographic pattern of correlations into a single number for each model by taking the magnitude of the projection of that model's cloud albedo map onto the map of multi-model correlation between cloud albedo and λ_{cld}. Models with smaller present-day cloud albedo in warm-SST regions and larger cloud albedo in cold regions have larger values of this index. Larger index also means greater λ_{cld} and hence larger ECS.

Because almost all models used in Siler et al. (2017) have the output needed for our study, our correlations in Table 2 are almost identical to Siler's. Interestingly, even though the Siler constraint is more sophisticated than Volodin, it does not produce stronger correlation.

o. Cox

Cox et al. (2018) use a simple differential equation for surface temperature response to white-noise radiative forcing in the presence of climate feedbacks to motivate an emergent constraint related to the strength and autocorrelation of global-averaged surface temperature variations. Models with larger temperature variations and stronger year-to-year autocorrelation tend to have larger ECS. Unlike other constraints, application of the historical temperature record to the Cox constraint implies ECS values which are somewhat weaker than the CMIP5 multi-model mean.

The Cox constraint is an interesting fit for our study because its proposed mechanism is related to fluctuation dissipation rather than a particular feedback process. As a result, our decomposition cannot be used to assess the validity of the Cox constraint. We include the Cox study in our analysis because it is currently the subject of great community interest and because our decomposition illuminates the physical mechanisms controlling the temperature response investigated by Cox et al. (2018).

4. Statistical Significance

Most of the potential constraints described above provide some mention that their correlations are significant but provide few details about how this was tested. Significance of correlations can be easily tested either by noting that $r\sqrt{(N-2)/(1-r^2)}$ follows a t-distribution (if ECS and emergent constraint values are normally distributed) or by using bootstrapping (constructing randomized samples by shuffling the model associated with each ECS or emergent-constraint value repeatedly to build up an empirical distribution for the null hypothesis). Both of these approaches typically assume that each model is an independent sample, but Sanderson et al. (2015) note that a quarter to a half of the CMIP5 models they analyzed were functionally redundant. This means that the appropriate number of degrees of freedom for these tests is significantly lower than the number of models evaluated. This is important because large correlations occur by chance more frequently when the number of degrees of freedom is greater.

Low sample size makes it very difficult to say anything definitive at all about relationships in the CMIP archives. One manifestation of this is the likelihood that some previously-proposed constraints are spurious. Identifying such constraints is the main goal of this paper. Unfortunately, small sample size also works against the goal of identifying bad constraints in the sense that a constraint may fail the tests in this paper not because it is incorrect, but instead because of unlucky alignment of available models. An anecdote puts this danger in context. Initially we followed Caldwell et al. (2016) by only using CMIP5 models which had less than 15% error in their clear-sky radiative kernel calculations. Eventually we decided to include all models in our analysis because the increase in sample size was deemed worth the the potential for increased sampling error, particularly because cloud feedbacks are the dominant source of correlation with ECS and their calculation is relatively accurate and only weakly affected by kernel errors. In 17 of the 19 constraints tested here, this change in ensemble composition had little effect. For the Qu constraint, however, correlation dropped from -0.63 to -0.29 when all models were used. Using all models had the opposite effect on Brient Shal - its correlation grew from 0.05 to 0.38. Scatter plots for each of these relationships are presented in Fig. 1. In both cases, correlation changed because models that failed the clear-sky linearity test had systematically different behavior than the rest of the ensemble. Does this mean that Qu is more credible than indicated by the rest of this study? Is Brient Shal less credible? We interpret these findings as an indication of the uncertainty in any correlation obtained from CMIP data. If our results are any indication, results are robust 17/19≈90% of the time and are misleading or ambiguous the other 10% of the time.

Another issue is that the search for emergent constraints naturally lends itself to trying relationships until a strong correlation is found. This is problematic because if one tries n relationships for significance at the $S\%$ level there is a $1-(S/100)^n$ probability of getting at least one false positive relationship, and this probability approaches 1 as $n \rightarrow \infty$. As an example, if you find one relationship out of 5 that is significant at the 95% level, the probability of this relationship occurring by chance is actually 23%. And while this sort of data-mining can be done purposefully, it can also occur unconsciously within a community. In particular, scientists are likely to notice and report strong correlations while keeping silent about their negative results. As a result, even if researchers are conscientious about the significance of their results on an individual level, the publication process will overstate the significance of their work by neglecting to account for unsuccessful attempts to find relationships. Because it is difficult or impossible to honestly say how many attempts were made to find a strong correlation before achieving success, Caldwell et al. (2014) and Klein and Hall (2015) advocate giving up on formal significance testing and instead relying on an undeniable physical explanation as proof of meaningful correlation. That is not to say that strong correlations should be ignored - finding these relationships is the first step to understanding them - but one should retain a healthy skepticism of relationships until they are fully understood. In order to identify obviously insignificant results, relationships that pass a t-test at 90% probability assuming independence between models are printed in bold in Table 2. Note that this test is not meant as a measure of statistical significance, but instead is used to identify constraints that are definitely *not* significant. All constraints except Covey pass this test when using their original data; Covey et al. (2000) themselves note that their metric is not quite significant at 90%. With the exception of constraints being confronted with CMIP5 data for the first time, this weak form of significance is maintained when ensembles are subset to the models with sufficient data to compute feedback and forcing.

5. Are Emergent Constraints Independent?

In this section we focus on the question of whether previously-proposed constraints are truly independent, or whether they are merely different manifestations of the same underlying phenomenon. This is important because as identifying emergent constraints becomes more popular, researchers need to be careful that new constraints are not merely repackaged versions of older constraints. We are in a unique position to answer this question by virtue of the large collection of previously-proposed constraints we have gathered.

Fig. 2 shows correlation coefficients for all combinations of emergent constraints considered. Diagonal values are always 1 because a constraint is perfectly correlated with itself. Cells above the diagonal are redundant because $\mathrm{corr}(x,y) = \mathrm{corr}(y,x)$ for any x and y and therefore has been omitted. Constraints which are negatively correlated with ECS in Table 2 are multiplied by -1 to aid comparison. To maximize sample sizes, correlations are computed using data from all available models rather than only those for which we have feedback decomposition information. The number of models used in each correlation is included in parentheses within each cell of Fig. 2. Pooling CMIP3 and CMIP5 data for these correlations is reasonable because we are only interested in cross-model relationships between constraints. CMIP3 and CMIP5 models are considered separately in Sect. 6 because their decompositions differ.

Correlations which are significant at 90% using a 2-tailed t-test are shown in color, with darker colors indicating stronger correlations. As noted in Sect. 4, the probability that at least one of the 152 correlations below the diagonal of Fig. 2 passes our significance test by chance is $1-0.9^{152} \approx 100\%$. The expected number of significant correlations by chance alone can be computed by noting that significance of correlation between a given pair of constraints is a Bernoulli trial with a 0.1 probability of success and the probability of a given number of successful Bernoulli trials follows a binomial distribution. This yields an expected value of $0.1 \times 152 \approx 15$, which is much smaller than the 64 significant correlations actually found. One purely mathematical reason to expect correlation between constraints is that all constraints were chosen for their high correlation with ECS, and if they all look like ECS then they must look like each other as well. This is probably why almost all correlations in Fig. 2 are positive (reddish). If similarity to ECS was the only reason for correlation between constraints, however, we might expect the constraints best correlated with ECS to be more strongly correlated with each other. With the exception of Covey, which is the most independent constraint and the worst-correlated with ECS, this does not seem to be the case. Fig. 3 shows that strong correlation with ECS does not imply significant correlation with more or better constraints. In other words, mathematics alone cannot explain the large number of significantly correlated constraints in Fig. 2, leading us to turn now to exploration of physical explanations for these relationships.

Overlapping groups of constraints for which we might expect a relationship based on physical grounds are indicated by colored lines and corresponding numbers in Fig. 2. The first grouping involves Lipat, Trenberth, and Volodin, which are all related to present-day southern hemisphere cloudiness. Siler is also included in this group because its definition is so similar to Volodin's; Fig. 2 shows that Volodin and Siler are correlated at 0.8. All constraints in this group are correlated at ≥ 0.5, suggesting that they may all be part of a single southern-hemisphere mechanism.

The constraints in group 2 are related to mean-state clouds and related indicators over geographically-broad areas. Surprisingly, while Siler is well correlated with all constraints in this group, the other constraints are not that well correlated with each other. It seems natural that if Siler is similar to two other constraints, those constraints should be similar to each other. Such behavior is known in math as the triangle inequality, and obeying this constraint is a requirement for all measures of distance. Our correlation matrix does not satisfy the triangle inequality because each correlation is based on a different ensemble of models. Using a single set of models for all constraints would solve this problem but is untenable here because we would be left with 7 models. There are also real reasons to expect mean-state constraints to be uncorrelated. Klein TCA and ctp-tau focus on cloud fraction and optical depth, for example, and these two quantities can change independently.

Group 3 contains constraints based on mean-state RH. If the locations picked by Fasullo are particularly important, one may expect them to show up in the Su constraint. This does not seem to be the case. Group 4 consists of Tian and Fasullo D, which both target convection-related variables in largely overlapped regions. Unsurprisingly, they are correlated at 0.6. Group 5 consists of constraints based on the ability of convection to remove moisture from the tropical boundary layer. Sherwood D and S are uncorrelated, which explains why LTMI=D+S explains a much larger fraction of ECS than D or S in isolation. Brient Shal, which was based on Sherwood's concepts, seems to be an unrelated constraint.

The last group focuses on current-climate response of low clouds to variations in SST. Zhai and Brient Alb do seem to be related to each other, but they are only weakly related to Qu. This could be due to unlucky sampling, but it could also be due to differences in constraint design, including differences in geographical region, sampling time periods, or the fact that Qu removes the component of cloud change coincident with EIS. It is also interesting to note that Volodin and Siler are strongly correlated with Brient Alb at 0.7. Volodin/Siler and Zhai/Brient Alb are similar in that both assume cloud changes track SST in a climate-invariant way, so perhaps this is unsurprising. The fact that Zhai/Brient use temporal variations as their present-day measure while Volodin/Siler use geographic variations raises interesting questions about cloud feedback.

Because Tian cites Trenberth and Fasullo (2010) for support, it is worth noting that the correlation between Tian and Trenberth is 0.5. Tian also makes reference to the strength of the Hadley circulation. Su and Lipat both measure aspects of the Hadley cell, but Su and Lipat are not significantly correlated with Tian or with each other.

Perhaps more interesting than the blocks of expected relationships in Fig. 2 is the region of unexpected correlations. With the exception of the Covey constraint - which is poorly correlated not only with most other constraints, but also with ECS - these unexpected correlations have similar magnitude to those found in the expected-relationship blocks. Several unexpected correlations are over 0.7 in magnitude! Of particular interest is the Cox constraint. The Cox constraint is designed to measure the response properties of global-average surface temperature to forcing, but the feedback process governing that response is unclear. Cox is very strongly correlated with Zhai and Klein TCA, and is significantly correlated with Brient Alb, Sherwood LTMI, and Tian. All of these clouds are related to clouds, suggesting that clouds are the main mechanism controlling surface temperature variations. This hypothesis will be further explored in Sect. 6.

Given the number of unexpected yet apparently significant connections between constraints, an empirical method for identifying groups of related constraints seems warranted. We tried a variety of clustering algorithms, but failure of the triangle equality makes the results very sensitive to the definition of distance between clusters and makes the results difficult to interpret. For example, should Volodin and Klein TCA be considered synonymous with Siler because both are correlated with Siler at 0.8? Or is Klein TCA a separate mechanism because it's correlation with Volodin is only 0.3? Larger model ensembles and/or mechanistic understanding of potential relationships is needed to decide. In the meantime, we simply state that pairs of constraints with large correlation in Fig. 2 are probably related, and that understanding why is important future work. Central to this goal is the need to understand why each constraint has skill in predicting ECS. Doing so is the focus of the remainder of this paper.

6. Decomposing Correlations

So far we have described each of the 19 constraints and tested them against new data where possible. We have also looked for relationships between emergent constraints. In this section we describe a method for decomposing correlation between ECS and an emergent constraint into components associated with individual feedback and forcing processes and into contributions from different geographical regions. In some sense the decomposition described here provides a bridge between predictors of ECS and predictors of individual climate processes by identifying the processes and regions which contribute to correlation with ECS and by clarifying how correlation with a particular process contributes (or is unimportant for) correlation with ECS.

a. Global-Average Decomposition

The first step in this decomposition is to write the net feedback λ as a sum of individual feedback terms λ_i such that $\lambda = \sum_{i \in P} \lambda_i$. For the purposes of this paper, feedback mechanisms in the set P will consist of albedo feedback (Alb), lapse-rate feedback (LR), water vapor feedback (WV), Planck feedback (Pl), shortwave cloud feedback (SW Cld), and longwave cloud feedback (LW Cld). See Bony et al. (2006) for a primer on these feedback mechanisms.

The next step in our decomposition is to approximate ECS by replacing $1/\lambda$ in Eq. 1 with its first-order Taylor expansion around the multi-model mean λ as described in Caldwell et al. (2016):

$$\mathrm{ECS} = -\frac{\overline{F}}{\overline{\lambda}} - \frac{F'}{\overline{\lambda}} + \frac{\overline{F}}{\overline{\lambda}^2} \sum_{i \in P} \lambda_i' + E_{\mathrm{kernel}} + E_{\mathrm{Taylor}} \tag{4}$$

where overbar indicates the multi-model average and primes indicate deviations from this average. E_{kernel} and E_{Taylor} are errors due to nonlinearity in radiative kernel calculations (computed as the difference between ECS calculated using $-F/\lambda$ versus $-F/\sum_{i \in P} \lambda_i$) and error due to the $1/\lambda$ Taylor expansion (computed as the residual in Eq. 4 after accounting for E_{kernel}). Both error terms are shown later to be small. In addition to causing discrepancies between ECS and $-F/\overline{\lambda}^2 \sum_{i \in P} \lambda_i'$ (as captured by E_{Taylor}), the Taylor approximation will result in misleading partitioning between feedback processes unless $\lambda_i' << \overline{\lambda}$. As discussed in Caldwell et al. (2016), this condition is overwhelmingly met for all λ_i except cloud feedback. Nonlinearity related to cloud feedback was shown in Caldwell et al. (2016) to have a predictable and minor role on intermodel spread in ECS.

The term on the right-hand side of Eq. 4 corresponding to a given j in the set $A = P \cup \{F, \text{const}, E_{\text{kernel}}, E_{\text{Taylor}}\}$ will be denoted by T_j. Using this shorthand, we can partition correlation between ECS and an arbitrary emergent constraint X into correlations with individual feedback and forcing terms:

$$\begin{aligned} \text{corr}(X, \text{ECS}) &= \frac{\text{cov}(X, \sum_{j \in A} T_j)}{\sigma(X)\sigma(\text{ECS})} \\ &= \sum_{j \in A} \frac{\text{cov}(X, T_j)}{\sigma(X)\sigma(\text{ECS})} \frac{\sigma(T_j)}{\sigma(T_j)} \\ &= \sum_{j \in A} \frac{\sigma(T_j)}{\sigma(\text{ECS})} \text{corr}(X, T_j) \end{aligned} \tag{5}$$

where $\sigma(\cdot)$ is the standard deviation operator, $\text{corr}(\cdot,\cdot)$ is the Pearson correlation coefficient, and $\text{cov}(\cdot,\cdot)$ is the covariance. The second line uses the identity $\text{cov}(X, \sum_{j=1}^{N} Y_j) = \sum_{j=1}^{N} \text{cov}(X, Y_j)$. In words, this equation says that correlation between an emergent constraint and ECS can be interpreted as the sum of correlations with each T_j term scaled by the relative importance of that term to ECS variations. This means that correlation with ECS is best achieved by being correlated with T_j terms which contribute most strongly to var(ECS).

Decomposition of correlation between ECS and each emergent constraint following Eq. 5 are provided in Fig. 4. As noted in Sect. 2, CMIP3 data (in panel a) and CMIP5 data (in panel b) differ in the following ways:

1. Only net Cld is shown in panel a) because separate SW and LW Cld components are not available for CMIP3.

2. For CMIP3, λ_{Cld} is computed as the residual between the net feedback and the sum of non-cloud feedbacks. As a result, E_{kernel} is absorbed into λ_{Cld} in panel a).

3. Pl, WV, and LR are computed relative to fixed specific humidity in panel a), whereas panel b uses fixed RH.

Note that contributions from E_{kernel} and E_{Taylor} are generally small, indicating that our decomposition is appropriate.

Cloud feedback is the main source of strong correlation with ECS for most emergent constraints, particularly in the CMIP3 ensemble. This can be understood by noting that the correlation of a particular T_j term in Eq. 5 is modulated by that term's contribution to var(ECS). The magnitude of these weighting factors is presented in Table 3. Because λ_{Cld} dominates inter-model spread in ECS (Dufresne and Bony 2008; Caldwell et al. 2016), it receives by far the largest weighting in Table 3. Put simply, inter-model variations in T_{Cld} are so big that they leave a strong imprint in inter-model variations in ECS. This means that fields which are strongly correlated with ECS are probably correlated with T_{Cld} (and vice versa). Emergent constraints which are more weakly correlated with ECS have more latitude to obtain that correlation from other terms. This is reflected in the fact that the constraints with little contribution from clouds (e.g. Covey and Sherwood S from CMIP5) also have relatively low correlation with ECS. The practical implication of this finding is that the search for emergent constraints for ECS should target mechanisms related to cloud feedback. Note in particular that the Cox constraint, which was not proposed as being related to clouds at all, is strongly dominated by SW cloud feedback.

One odd feature of Table 3 is that the Cld term has weight greater than 1. This seems to imply that strong correlations with T_{Cld} could cause $\mathrm{corr}(X, \mathrm{ECS})$ to exceed 1. This is not the case because having a weight greater than 1 is only possible due to anti-correlation with other T_j terms. These anti-correlated terms can be relied on to prevent correlations with ECS from exceeding their allowable bounds. Similarly, while it may at first appear that weights imply certain relationships between ECS and T_j (e.g. one might assume that $\mathrm{corr}(X, \mathrm{ECS})$ is always greater than $\mathrm{corr}(X, T_{\mathrm{Cld}})$ because the weight for Cld is greater than 1), the potential for anti-correlation between T_j terms means that no such relationship exists.

This anti-correlation between T_j terms was documented in Caldwell et al. (2016). Its effect here is to make it impossible for any potential constraint to be correlated with ECS due to only one feedback or forcing mechanism. For example, λ_{Cld} is opposed by λ_{LR} for almost all constraints in Fig. 4. Correlation between processes means that Klein and Hall (2015)'s 'no multiple influences' requirement is probably unworkable. In light of this finding, a better criterion for a promising emergent constraint is that correlation with ECS should come *primarily* from a single T_j. The logic of this criterion is that it is much easier to imagine a physical explanation involving one or at most a pair of feedback mechanisms, while a mechanism comprised of a complex mixture of feedbacks is hard to imagine. We will use the criterion that the dominant constraint should be twice as large as any other term unless there's a good physical reason to expect otherwise as a way to screen for constraints arising from unlikely mixtures of processes. Covey and Fasullo M fail this criterion for both CMIP3 and CMIP5, while Sherwood S, Trenberth, Fasullo D, Lipat, and Qu fail for CMIP5.

Influence from multiple sources seems to be more pronounced in CMIP5 ensembles. Some of this comes from the fact that SW and LW Cld components are included separately for CMIP5 data but not for CMIP3 (because of lack of available data), but CMIP5 data is more complex even when SW and LW cloud feedback is combined. One reason for this is probably due to increasing model complexity with time. The fact that CMIP3 values are computed from runs which include transient aerosol changes while CMIP5 data don't also complicates interpretation. Similarity between decompositions computed using independent ensembles would be a useful indicator of the credibility of an emergent constraint, but cannot be evaluated in this study because of differences in the experimental design of CMIP3 and CMIP5. One aspect of the CMIP5 results which is simpler is the partitioning between LR and WV: in CMIP3 these quantities oppose each other and are of roughly equal size. When computed relative to fixed RH in CMIP5, however, the importance of WV fades and LR is shown to be the dominant source of correlation with ECS.

Fig. 4 can be used to test whether a constraint's correlation with ECS is due to its proposed physical explanation or not. This is only possible for constraints with well-defined physical mechanisms; constraints without an explanation cannot be tested and therefore can never be moved beyond the 'potential constraint' status. Sherwood D and LTMI and Brient Shal and alb pass this test for both CMIP3 and CMIP5 data - they are proposed to operate through changes in low clouds and their correlation with ECS comes primarily through shortwave cloud feedback. Sherwood S and Qu are also meant to operate through shortwave cloud feedback but gain correlation mainly through other terms for CMIP5 data. Non-robustness between ensembles suggests these constraints may be spurious. Further decomposition of the SW cloud feedback term into amount and scattering components (not shown) reveals that Qu - which was originally framed in terms of low cloud amount changes - is operating as intended in the sense that SW cloud amount feedback in stratocumulus regions does actually contribute to negative correlation with ECS but its effect is canceled out by opposing contributions from SW cloud scattering feedback.

b. Regional Decomposition

Eq. 5 can be further dissected to include geographical information. This is useful to test proposed constraints which are meant to target a process specific to a particular region. To do this, we note that for each term in Eq. 5,

$$\begin{aligned}\frac{\sigma(T_j)}{\sigma(\text{ECS})}\text{corr}(X,T_j) &= \frac{\sigma(T_j)}{\sigma(\text{ECS})}\frac{\text{cov}(X,\sum_{k=1}^{N} w_k T_{jk})}{\sigma(X)\sigma(T_j)} \\ &= \sum_{k=1}^{N}\frac{w_k\sigma(T_{jk})}{\sigma(\text{ECS})}\text{corr}(X,T_{jk})\end{aligned} \tag{6}$$

In the first line of this equation, we write the correlation as a covariance and rewrite global-average T_j as an area-weighted average over spatial dimension k, where w_k is the area weighting for grid cell k. The second line follows logic similar to that used to derive Eq. 5. In Eq. 6, the contribution from grid cell k is given by that particular location's correlation with X weighted by the contribution of that cell's area to the global total and the relative importance of process i at location k to var(ECS). Combining Equations 5 and 6 allows us to plot the contribution of each feedback in each grid cell of the model to $\text{corr}(X,ECS)$.

This geographical decomposition is applied to selected constraints in Fig. 5. These constraints were chosen because they are relatively independent of each other, they target different regions, and they are of contemporary interest. Similar figures including SW and LW cloud feedbacks separately are available in the supporting material for all constraints. In these plots, ECS is broken into terms due to net Cld and F, with all other terms combined into a single plot because their spatial variations are unimportant.

One striking feature of Fig. 5 is that the correlation contributions for a particular T_j tend to come from the same geographic regions for all constraints. Analogous to the way that $\lambda_{\text{SW Cld}}$ dominates the global-average contribution to ECS because it contributes most to ECS variations, these geographic locations are most important because their variations are the main source of differences in the global-average of the T_j in question. An example of the geographic scaling factor $w_k \sigma(T_{jk})/\sigma(\mathrm{ECS})$ is presented in Fig. 6 for each term in our decomposition. These maps differ slightly depending on the set of models used for each constraint, but the patterns are quite similar for all constraints. Ability to predict cloud feedback and forcing variations in the tropical Pacific is most important for getting ECS right. Polar regions in the bottom row of Fig. 6 also show up as important due to snow albedo feedback.

Another interesting feature of Fig. 5 is the fact that the spatial distribution of F contribution for each constraint is almost perfectly opposed in the tropics by the net cloud contribution. Anti-correlation between F and λ for simulations without aerosol changes was previously noted for global averages by Ringer et al. (2014). This relationship may be an artifact of the fact that we follow Gregory et al. (2004) in computing feedback and forcing as the slope and y-intercept of the same data. This hypothesis could be tested by getting F from runs with 4xCO_2 and present-day SST, but such analysis is outside the scope of this paper.

While similarities in geographic structure between constraints is interesting, the main goal of Fig. 5 is to test proposed mechanisms. The Lipat constraint is related to cloud changes in the Southern hemisphere at the border between the subtropics and midlatitudes. Geographic decomposition of the cloud contribution from Lipat does show more amplitude in this latitude band than other constraints, but this region is still not the main source of correlation with ECS. Without understanding how Hadley cell extent could affect future changes in *tropical* cloudiness, this constraint remains unconfirmed. Sherwood D predicts tropical low cloud changes due to BL drying by convection. It is computed using data from the tropical Atlantic and East Pacific, though it is unclear whether this is the region where cloud changes are expected. It does have large correlation in the tropical East Pacific and Atlantic, but its correlation with cloud feedback in the West Pacific is even bigger. Brient Alb is related to low clouds in subsiding (eastern subtropical) oceanic regions.

Magnitude of correlation in the subtropical eastern oceans is greater than for other constraints, but equatorial ascent-region clouds again play an unexpectedly large role. Geographic decomposition is also illuminating for the other constraints, but is relegated to Supplementary Material because validity of the corresponding constraint can already be assessed from the other information in this paper.

7. Discussion and Conclusions

This study provides several methods for evaluating the credibility of a proposed emergent constraint. We hope this work triggers an effort to evaluate new and existing emergent constraints, discarding unreliable constraints and developing consensus and trust around confirmed predictors. To that end, we ask which of the 19 emergent constraints tested here are trustworthy. Our assessment is provided in Table 4. Six constraints (Covey, Trenberth, Fasullo D and M, Sherwood S, and Sherwood LTMI) do not appear to be credible because they are either not robust to change of ensemble or their correlation with ECS is not due to their proposed physical mechanism. The credibility of 3 constraints - Lipat, Qu, and Cox - is ambiguous. Lipat gains correlation with ECS from the expected region and mechanism but gains more correlation from unexpected sources. Similarly, Qu also gains correlation from the expected mechanism and region but fails to have a large correlation with ECS for the models used in this study because of unexpected compensation from other terms. Additionally, while Qu fails to be robust to all changes in ensemble, it does have a large correlation with ECS in CMIP3 and in a subset of CMIP5 models and it is conceptually related to the Zhai and Brient Alb constraints, which do seem to be robust. The Cox constraint has a physical explanation which is unrelated to particular feedbacks and regions and hence cannot be tested in our framework. An additional 6 constraints (Volodin, Siler, Klein ctp-tau and TCA, Su, and Tian) cannot be tested because they lack clear physical mechanisms. These constraints should not be considered credible until they are fully understood. Decomposition of these constraints' correlation with ECS may prove useful in uncovering the physical explanations for their skill (if any exist). In this context it is interesting to note that Klein TCA is predominantly related to tropical LW cloud feedback, while Klein ctptau and Tian are related to SW cloud feedback over a broad variety of regions and Su is tied to cloud feedback mainly over the tropics (see Supplementary Figures 3 and 4).

The remaining 4 constraints (Sherwood D, Brient Shal, Zhai, and Brient Alb) pass all tests in this study and thus seem credible. Worryingly, all of the studies introducing these constraints note that their constraint implies higher climate sensitivity than predicted by giving each CMIP5 model equal weight. The Sherwood D constraint in particular is only satisfied by models with ECS greater than 3.4 K, while the Sherwood S and LTMI metrics - which themselves predict relatively high climate sensitivity - are much closer to the centroid of CMIP model values. While the tendency for emergent constraints to predict higher climate sensitivity has been noted in the past (e.g. Tian 2015; Qu et al. 2018), the validity of this finding has been unclear because it was based on *potential* rather than *credible* constraints.

So what does it mean that 4 credible emergent constraint studies all predict warming at the upper end of community expectation? One interpretation is that these studies *reinforce* each other's conclusions - if all agree, they must be right. This is an appropriate interpretation if all constraints are flawed samples of the same underlying underlying distribution/physical process. In this case, the more we sample the underlying distribution, the better we will understand it. If each constraint is instead targeting a different physical process which contributes to ECS, the constraints will contribute *additively* towards determining ECS. In this latter case, having one constraint predict high sensitivity and another predict low sensitivity does not invalidate the constraints - they may simply constrain different drivers of climate sensitivity. The credible constraints identified in this study are all related to tropical low clouds and all except Brient Shal are shown in Fig. 2 to be significantly correlated with each other. Zhai and Brient Alb even share a common physical mechanism. Thus it is tempting to view all constraints as reinforcing each other. It is, however, unsurprising that the best emergent constraints would be related to tropical low clouds because (as noted above) $\lambda_{\mathrm{SW\,Cld}}$ has largest impact on ECS and incident SW radiation is strongest in the tropics. There are also many processes contributing to tropical low cloud changes, so the credible constraints identified here could very well be capturing different mechanisms governing tropical cloud change. Understanding how these constraints relate to each other is important future work. Developing a numerical estimate of ECS by combining constraints would be very useful, but such an estimate will only be possible once we understand clearly how the constituent constraints are related. As noted by Klein and Hall (2015), a complete picture of ECS will only emerge once we are able to constrain every important feedback component in each important climate regime. It is therefore desirable to focus research efforts on developing constraints for individual processes and on identifying the appropriate infrastructure for combining these constraints into a coherent story.

Sect. 5 provides a first step towards identifying related constraints. By comparing constraint definitions and explanations as well as correlations between pairs of constraints, we conclude that Siler and Volodin describe the same physical mechanism, as do Zhai and Brient Alb. Beyond these pairs, we had trouble identifying groups of similar constraints because one constraint would frequently be correlated with two others which weren't themselves correlated with one another. This breakdown of the triangle inequality results from the fact that the models available for each constraint differ coupled with the the extremely-small sample size of the CMIP archives. While the 19 constraints considered here are definitely much more similar to each other than expected by chance, lack of empirical methods for grouping forces us to fall back on physical reasoning to identify related constraints. This is difficult when the mechanisms responsible for potential constraints are not well understood.

It is important to stress that all conclusions in Table 4 should be considered tentative because the number of models used in each correlation calculation is so small. As discussed in Sect. 4, insufficient sample size is underscored by the fact that correlation for 2 of the 19 constraints changed radically when we switched from using only models which passed the clear-sky linearity test to using all models. This is a problem not just with our methodology, but with all studies attempting to identify emergent constraints from the relatively small ensembles available from CMIP. This conclusion is supported by the fact that strong correlation with ECS disappeared in 4 of the 5 constraints in this study confronted with new ensembles. It is interesting to note, however, that several of these failing constraints are strongly correlated with newer constraints which do show strong correlations with CMIP5 data. Perhaps these original studies do have some value, but were overtuned to their training dataset. It is also worth noting that while our criteria of robustness across successive CMIP generations and correlation coming mainly from a single feedback mechanism seem like reasonable rules of thumb, there may be situations where real constraints do not satisfy these criteria. In these cases, the need for an exception should be obvious from the purported physical mechanism. Grounds for such an exception are not clear for any of the constraints evaluated here. Another important caveat to this study is that it focuses entirely on correlations, which only capture linear relationships while climate response may be nonlinearly related to a present-day predictor (see Appendix 2 of Covey et al. 2000, for an example). Compositing a predictor into an average over models with low ECS and a separate average over models with high ECS (as done by Su et al. 2014) and Brient et al. (2015) may be better for identifying nonlinear emergent constraints but is not conducive to our decomposition approach.

Acknowledgments. We would like to acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the World Climate Research Program's Working Group on Coupled Modeling for their roles in making available the CMIP3 and CMIP5 multi-model datasets. Support of these datasets is provided by the U.S. Department of Energy (DOE) Office of Science. We would also like to thank the authors of the emergent constraint studies evaluated here for sharing their data. Finally, a huge thanks to Paul Durack for maintaining the CMIP archive for internal use at LLNL. This work was supported by the Office of Science Biological and Environmental Research (BER) program at Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. All authors were supported by BER's Regional and Global Climate Modeling (RGCM) Program.

References

Andrews, T., J. M. Gregory, and M. J. Webb, 2015: The Dependence of Radiative Forcing and Feedback on Evolving Patterns of Surface Temperature Change in Climate Models. *Journal of Climate*, **28**, 1630–1648, doi:10.1175/JCLI-D-14-00545.1.

Andrews, T., J. M. Gregory, M. J. Webb, and K. E. Taylor, 2012: Forcing, feedbacks and climate sensitivity in cmip5 coupled atmosphere-ocean climate models. *Geophysical Research Letters*, **39 (9)**, n/a–n/a, doi:10.1029/2012GL051607, URL http://dx.doi.org/10.1029/2012GL051607.

Armour, K. C., C. M. Bitz, and G. H. Roe, 2013: Time-varying climate sensitivity from regional feedbacks. *J. Climate*, **26**, 45184534, doi:10.1175/JCLI-D-12-00544.1.

Bony, S., and J.-L. Dufresne, 2005: Marine boundary layer clouds at the heart of cloud feedback uncertainties in climate models. *J. Geophys. Res.*, **32 (20)**, L20 806, doi: 10.1029/2005GL023 851.

Bony, S., and Coauthors, 2006: How well do we understand and evaluate climate change feedback processes? *J. Climate*, **19**, 34453482.

Bretherton, C. S., and P. N. Blossey, 2014: Low cloud reduction in a greenhouse-warmed climate: Results from lagrangian les of a subtropical marine cloudiness transition. *Journal of Advances in Modeling Earth Systems*, **6 (1)**, 91–114, doi:10.1002/2013MS000250, URL http://dx.doi.org/10.1002/2013MS000250.

Brient, F., and T. Schneider, 2016: Constraints on climate sensitivity from space-based measurements of low-cloud reflection. *Journal of Climate*, **29 (16)**, 5821–5835, doi:10.1175/JCLI-D-15-0897.1, URL https://doi.org/10.1175/JCLI-D-15-0897.1.

Brient, F., T. Schneider, Z. Tan, S. Bony, X. Qu, and A. Hall, 2015: Shallowness of tropical low clouds as a predictor of climate models' response to warming. *Climate Dynamics*, 1–17, doi:10.1007/s00382-015-2846-0, URL http://dx.doi.org/10.1007/s00382-015-2846-0.

Caldwell, P. M., C. S. Bretherton, M. D. Zelinka, S. A. Klein, B. D. Santer, and B. M. Sanderson, 2014: Statistical significance of climate sensitivity predictors obtained by data mining. *Geophysical Research Letters*, **41 (5)**, 1803–1808, doi:10.1002/2014GL059205, URL http://dx.doi.org/10.1002/2014GL059205.

Caldwell, P. M., M. D. Zelinka, K. E. Taylor, and K. Marvel, 2016: Quantifying the sources of intermodel spread in equilibrium climate sensitivity. *Journal of Climate*, **29 (2)**, 513–524, doi:10.1175/JCLI-D-15-0352.1, URL https://doi.org/10.1175/JCLI-D-15-0352.1, https://doi.org/10.1175/JCLI-D-15-0352.1.

Charney, and Coauthors, 1979: *Carbon Dioxide and Climate: A Scientific Assessment*. The National Academies Press, Washington, DC, doi:10.17226/12181, URL https://www.nap.edu/

catalog/12181/carbon-dioxide-and-climate-a-scientific-assessment.

Collins, M., and Coauthors, 2013: *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, chap. Long-term Climate Change: Projections, Commitments and Irreversibility. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, (Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex,and P.M. Midgley (eds.)].

Covey, C., and Coauthors, 2000: The seasonal cycle in coupled ocean-atmosphere general circulation models. *Climate Dyn.*, **16**, 775–787.

Cox, P. M., C. Huntingford, and M. S. Williamson, 2018: Emergent constraint on equilibrium climate sensitivity from global temperature variability. *Nature*, **553**, 319 EP –, URL http://dx.doi.org/10.1038/nature25450.

Cox, P. M., D. Pearson, B. B. Booth, P. Friedlingstein, C. Huntingford, C. D. Jones, and C. M. Luke, 2013: Sensitivity of tropical carbon to climate change constrainted by carbon dioxide variability. *Nature*, **494**, 341–344.

Dufresne, J.-L., and S. Bony, 2008: An assessment of the primary sources of spread of global warming estimates from coupled atmosphere-ocean models. *J. Clim.*, **21**, 5135–5144.

Fasullo, J., and K. Trenberth, 2012: A less cloudy future: The role of subtropical subsidence in climate sensitivity. *Science*, **338**, 792–794, DOI:10.1126/science.1227 465.

Fasullo, J. T., B. M. Sanderson, and K. E. Trenberth, 2015: Recent progress in constraining climate sensitivity with model ensembles. *Current Climate Change Reports*, **1 (4)**, 268–275, doi:10.1007/s40641-015-0021-7, URL http://dx.doi.org/10.1007/s40641-015-0021-7.

Flato, G., and Coauthors, 2013: *Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, chap. Evaluation of Climate Models. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, (Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex,and P.M. Midgley (eds.)].

Forster, P. M., and K. E. Taylor, 2006: Climate forcings and climate sensitivities diagnosed from coupled climate model integrations. *J. Climate*, **19**, 6181–6194.

Gleckler, P. J., K. E. Taylor, and C. Doutriaux, 2008: Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres*, **113 (D6)**, n/a–n/a, doi:10.1029/2007JD008972, URL http://dx.doi.org/10.1029/2007JD008972, d06104.

Gordon, N. D., and S. A. Klein, 2014: Low-cloud optical depth feedback in climate models. *Journal of Geophysical Research: Atmospheres*, **119 (10)**, 6052–6065, doi:10.1002/2013JD021052.

Gregory, J., and Coauthors, 2004: A new method for diagnosing radiative forcing and climate sensitivity. *Geophy. Res. Let.*, **31**, 10.1029/2003GL018 747.

Grise, K. M., L. M. Polvani, and J. T. Fasullo, 2015: Reexamining the Relationship between Climate Sensitivity and the Southern Hemisphere Radiation Budget in CMIP Models. *Journal of Climate*, **28**, 9298–9312, doi:10.1175/JCLI-D-15-0031.1.

Hall, A., and X. Qu, 2006: Using the current seasonal cycle to constrain snow albedo feedback in future climate change. *Geophy. Res. Let.*, **33**, L03 502, DOI: 10.1029/2005GL025 127.

Held, I. M., and K. M. Shell, 2012: Using relative humidity as a state variable in climate feedback analysis. *J. Climate*, **25**, 25782582. doi:JCLI–D–11–00 721.1.

Held, I. M., and B. Soden, 2000: Water vapor feedback and global warming. *Annu. Rev. Energy Environ*, **25**, 441–475.

Held, I. M., and B. Soden, 2006: Robust responses of the hydrological cycle to global warming. *J. Climate*, **19**, 5686–5699.

Hirota, N., and Y. N. Takayabu, 2012: Inter-model differences of future precipitation changes in cmip3 and miroc5 climate models. *Journal of the Meteorological Society of Japan. Ser. II*, **90A**, 307–316, doi:10.2151/jmsj.2012-A16.

Huber, M., I. Mahlstein, M. Wild, J. Fasullo, and R. Knutti, 2011: Constraints on climate sensitivity from radiation patterns in climate models. *J. Clim.*, 1034–1052.

Hwang, Y.-T., and D. M. W. Frierson, 2013: Link between the double-intertropical convergence zone problem and cloud biases over the southern ocean. *Proceedings of the National Academy of Sciences*, **110 (13)**, 4935–4940, doi:10.1073/pnas.1213302110, URL http://www.pnas.org/content/110/13/4935.abstract, http://www.pnas.org/content/110/13/4935.full.pdf.

Kamae, Y., H. Shiogama, M. Watanabe, T. Ogura, T. Yokohata, and M. Kimoto, 2016: Lower-tropospheric mixing as a constraint on cloud feedback in a multiparameter multiphysics ensemble. *Journal of Climate*, **29 (17)**, 6259–6275, doi:10.1175/JCLI-D-16-0042.1.

Kay, J. E., C. Wall, V. Yettella, B. Medeiros, C. Hannay, P. Caldwell, and C. Bitz, 2016: Global climate impacts of fixing the southern ocean shortwave radiation bias in the community earth system model (cesm). *Journal of Climate*, **29 (12)**, 4617–4636, doi:10.1175/JCLI-D-15-0358.1, URL https://doi.org/10.1175/JCLI-D-15-0358.1, https://doi.org/10.1175/JCLI-D-15-0358.1.

Klein, S. A., and A. Hall, 2015: Emergent constraints for cloud feedbacks. *Current Climate Change Reports*, **1 (4)**, 276–287, doi:10.1007/s40641-015-0027-1, URL http://dx.doi.org/10.1007/s40641-015-0027-1.

Klein, S. A., Y. Zhang, M. D. Zelinka, R. Pincus, J. Boyle, and P. J. Gleckler, 2013: Are climate model simulations of clouds improving? an evaluation using the isccp simulator. *Journal of Geophysical Research: Atmospheres*, **118 (3)**, 1329–1342, doi:10.1002/jgrd.50141, URL http://dx.doi.org/10.1002/jgrd.50141.

Klocke, D., R. Pincus, and J. Quaas, 2011: On constraining estimates of climate sensitivity with present-day observations through model weighting. *J. Clim.*, **24**, 6092–6099.

Knutti, R., 2010: The end of model democracy? *Clim. Change*, **102**, 395–404.

Knutti, R., and G. C. Hegerl, 2008: The equilibrium sensitivity of the Earth's temperature to radiation changes. *Nature Geo.*, **1**, 735–743.

Knutti, R., D. Masson, and A. Gettelman, 2013: Climate model genealogy: Generation CMIP5 and how we got there. *Geophy. Res. Let.*, **40**, 1194–1199.

Knutti, R., G. A. Meehl, M. R. Allen, and D. A. Stainforth, 2006: Constraining climate sensitivity from the seasonal cycle in surface temperature. *J. Climate*, **19**, 4224–4233.

Knutti, R., M. A. A. Rugenstein, and G. C. Hegerl, 2017: Beyond equilibrium climate sensitivity. *Nature Geoscience*, **10**, 727736, doi:10.1038/ngeo3017.

Lipat, B. R., G. Tselioudis, K. M. Grise, and L. M. Polvani, 2017: Cmip5 models' shortwave cloud radiative response and climate sensitivity linked to the climatological hadley cell extent. *Geophysical Research Letters*, **44 (11)**, 5739–5748, doi:10.1002/2017GL073151, URL http://dx.doi.org/10.1002/2017GL073151, 2017GL073151.

Masson, D., and R. Knutti, 2013: Predictor screening, calibration, and observational constraints in climate model ensembles: an illustration using climate sensitivity. *J. Climate*, 887–898.

McCoy, D. T., D. L. Hartmann, M. D. Zelinka, P. Ceppi, and D. P. Grosvenor, 2015: Mixed-phase cloud physics and southern ocean cloud feedback in climate models. *Journal of Geophysical Research: Atmospheres*, **120 (18)**, 9539–9554, doi:10.1002/2015JD023603, URL http://dx.doi.org/10.1002/2015JD023603, 2015JD023603.

McCoy, D. T., I. Tan, D. L. Hartmann, M. D. Zelinka, and T. Storelvmo, 2016: On the relationships among cloud cover, mixed-phase partitioning, and planetary albedo in gcms. *Journal of Advances in Modeling Earth Systems*, **8 (2)**, 650–668, doi:10.1002/2015MS000589, URL http://dx.doi.org/10.1002/2015MS000589.

Meehl, G. A., C. Covey, T. Delworth, M. Latif, B. McAvaney, J. F. B. Mitchell, R. J. Stouffer, and K. E. Taylor, 2007: The WCRP CMIP3 multi-model dataset: A new era in climate change research. *Bull. Amer. Meteor. Soc.*, **88**, 1383–1394.

Pennell, C., and T. Reichler, 2011: On the effective number of climate models. *J. Clim.*, **24**, 2358–2367.

Qu, X., A. Hall, A. M. DeAngelis, M. D. Zelinka, S. A. Klein, H. Su, B. Tian, and C. Zhai, 2018: On the emergent constraints of climate sensitivity. *Journal of Climate*, **31 (2)**, 863–875, doi:10.1175/JCLI-D-17-0482.1.

Qu, X., A. Hall, S. A. Klein, and P. M. Caldwell, 2013: On the spread of changes in marine low cloud cover in climate model simulations of the 21st century. *Climate Dynamics*, **42 (9)**, 2603–2626, doi:10.1007/s00382-013-1945-z, URL http://dx.doi.org/10.1007/s00382-013-1945-z.

Randall, D., and Coauthors, 2007: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, chap. Climate Models and Their Evaluation. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, Solomon S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M.Tignor, and H.L. Miller (eds.).

Ringer, M. A., T. Andrews, and M. J. Webb, 2014: Global-mean radiative feedbacks and forcing in atmosphere-only and coupled atmosphere-ocean climate change experiments. *Geophy. Res. Let.*, doi: 10.1002/2014GL060347.

Rose, B. E. J., K. C. Armour, D. S. Battisti, N. Feldl, and D. D. B. Koll, 2014: The dependence of transient climate sensitivity and radiative feedbacks on the spatial pattern of ocean heat uptake. *Geophysical Research Letters*, **41 (3)**, 1071–1078, doi:10.1002/2013GL058955, URL http://dx.doi.org/10.1002/2013GL058955.

Sanderson, B. M., 2011: A multimodel study of parametric uncertainty in predictions of climate response to rising greenhouse gas concentrations. *Journal of Climate*, **24 (5)**, 1362–1377, doi: 10.1175/2010JCLI3498.1.

Sanderson, B. M., R. Knutti, and P. Caldwell, 2015: A representative democracy to reduce interdependency in a multimodel ensemble. *Journal of Climate*, **28 (13)**, 5171–5194.

Shell, K. M., J. T. Kiehl, and C. A. Shields, 2008: Using the radiative kernel technique to calculate climate feedbacks in NCAR's community atmospheric model. *J. Hydrometeor.*, **21**, 2269–2282.

Sherwood, S. C., S. Bony, and J.-L. Dufresne, 2014: Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature*, **505 (7481)**, 37–42, doi:10.1038/nature12829, URL http://dx.doi.org/10.1038/nature12829.

Shukla, J., T. DelSole, M. Fennessy, J. Kinter, and D. Paolino, 2006: Climate model fidelity and projections of climate change. *Geophysical Research Letters*, **33 (7)**, n/a–n/a, doi:10.1029/2005GL025579, URL http://dx.doi.org/10.1029/2005GL025579, l07702.

Siler, N., S. Po-Chedley, and C. S. Bretherton, 2017: Variability in modeled cloud feedback tied to differences in the climatological spatial pattern of clouds. *Climate Dynamics*, **48**, 1–12, doi:10.1007/s00382-017-3673-2.

Soden, B. J., A. J. Broccoli, and R. S. Hemler, 2004: On the use of cloud forcing to estimate cloud feedback. *J. Climate*, **17**, 36613665.

Soden, B. J., and I. M. Held, 2006: An assessment of climate feedbacks in coupled ocean-atmosphere models. *J. Climate*, **19**, 3354–3360.

Soden, B. J., I. M. Held, R. Colman, K. M. Shell, J. T. Kiehl, and C. A. Shields, 2008: Quantifying climate feedbacks using radiative kernels. *J. Climate*, **21**, 3504–3520.

Su, H., J. H. Jiang, C. Zhai, T. J. Shen, J. D. Neelin, G. L. Stephens, and Y. L. Yung, 2014: Weakening and strengthening structures in the hadley circulation change under global warming and implications for cloud response and climate sensitivity. *Journal of Geophysical Research: Atmospheres*, **119 (10)**, 5787–5805, doi:10.1002/2014JD021642, URL http://dx.doi.org/10.1002/2014JD021642.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl, 2012: An Overview of CMIP5 and the Experiment Design. *Bull. Amer. Meteor. Soc.*, **93**, 485498, doi:10.1175/BAMS-D-11-00094.1.

Tian, B., 2015: Spread of model climate sensitivity linked to double-intertropical convergence zone bias. *Geophysical Research Letters*, **42 (10)**, 4133–4141, doi:10.1002/2015GL064119, URL http://dx.doi.org/10.1002/2015GL064119, 2015GL064119.

Trenberth, K. E., and J. T. Fasullo, 2010: Simulation of present day and 21st century energy budgets of the southern oceans. *J. Clim.*, **23**, 440–454.

Volodin, E. M., 2008: Relation between temperature sensitivity to doubled carbon dioxide and the distribution of clouds in current climate models. *Izvestiya, Atmospheric and Oceanic Physics*, **44**, 288–299.

Waugh, D. W., and V. Eyring, 2008: Quantitative performance metrics for stratospheric-resolving chemistry-climate models. *Atmospheric Chemistry and Physics*, **8 (18)**, 5699–5713, doi:10.5194/acp-8-5699-2008, URL http://www.atmos-chem-phys.net/8/5699/2008/.

Webb, M. J., and Coauthors, 2015: The impact of parametrized convection on cloud feedback. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, **373 (2054)**, doi:10.1098/rsta.2014.0414, URL http://rsta.royalsocietypublishing.org/content/373/2054/20140414, http://rsta.royalsocietypublishing.org/content/373/2054/20140414.full.pdf.

Williams, K. D., W. J. Ingram, and J. M. Gregory, 2008: Time variation of effective climate sensitivity in gcms. *J. Climate*, **21**, 50765090, doi:10.1175/2008JCLI2371.1.

Winton, M., K. Takahashi, and I. M. Held, 2010: Importance of ocean heat uptake efficacy to transient climate change. *J. Climate*, **23**, 23332344, doi:10.1175/2009JCLI3139.1.

Zhai, C., J. H. Jiang, and H. Su, 2015: Long-term cloud change imprinted in seasonal cloud variation: More evidence of high climate sensitivity. *Geophysical Research Letters*, **42 (20)**, 8729–8737, doi:10.1002/2015GL065911, URL http://dx.doi.org/10.1002/2015GL065911, 2015GL065911.

Zhao, M., 2014: An investigation of the connections among convection, clouds, and climate sensitivity in a global climate model. *Journal of Climate*, **27 (5)**, 1845–1862, doi:10.1175/JCLI-D-13-00145.1, URL https://doi.org/10.1175/JCLI-D-13-00145.1.

LIST OF TABLES

TABLE 1. Short description of each emergent constraint tested in this paper.

Name	Description
Covey	Amplitude of seasonal cycle of surface temperature
Volodin	Difference between tropical and southern-hemisphere midlatitude total cloud fraction
Trenberth	Net TOA radiation averaged over the southern hemisphere
Fasullo D	Southern hemisphere zonal-average mid-tropospheric RH in dry-zone between 8.5°-20°S
Fasullo M	Tropical zonal-average lower-tropospheric RH in moist-convective region
Qu	BL cloud amount response to SST variations in subtropical stratocumulus regions (after removing EIS contribution)
Klein ctp-tau	Error in the distribution of cloud-top pressure and optical thickness for regions between 60°N and S
Klein TCA	Error in total cloud amount for regions between 60°N and S
Su	Error in vertically-resolved tropospheric zonal-average RH between 40°N and 45°S
Sherwood D	Strength of resolved-scale mixing between BL and lower troposphere in tropical E Pacific and Atlantic
Sherwood S	Strength of mixing between BL and lower troposphere in tropical convective regions
Sherwood LTMI	Sum of Sherwood S and D constraints
Brient Shal	Fraction of tropical clouds with tops below 850 mb whose tops are also below 950 mb
Zhai	Seasonal response of BL cloud amount to SST variations in oceanic subsidence regions between 20-40°latitude
Tian	Strength of double-ITCZ bias
Brient Alb	Sensitivity of cloud albedo in tropical oceanic low-cloud regions to present-day SST variations
Lipat	Latitude of the southern edge of the Hadley cell in austral summer
Siler	Extent to which cloud albedo is small in warm-SST regions and large in cold-SST regions
Cox	Strength of global-average surface temperature variations and temporal autocorrelation

TABLE 2. Correlations between emergent constraints as reported in their original papers and as computed using the subsets of models for which we have constraint information as well as forcing and feedback components. Except for reported values from Covey (which used CMIP1 data), columns 2 and 3 report a single number if the study combined CMIP3 and CMIP5 models and otherwise reports individual CMIP3 and CMIP5 values separated by a '/'. Values in bold are significant at 90% confidence using a t-test assuming independent models (which is an overly-permissive test, see text for details). An asterisk is used where no data is available.

	Reported # Models	Reported Values	# CMIP3 Models	CMIP3 Values	# CMIP5 Models	CMIP5 Values
Covey	17	0.40	12	-0.36	27	**0.35**
Volodin	18 / *	**-0.82** / *	12	-0.42	27	**-0.60**
Trenberth	13 / *	**-0.73** / *	12	**-0.56**	27	-0.22
Fasullo D	16 / *	**-0.81** / *	9	**-0.78**	23	-0.26
Fasullo M	16 / *	**0.65** / *	9	**0.74**	23	0.15
Qu	18 / 18	* / *	11	**-0.61**	16	-0.29
Klein ctp-tau	6 / 9	* / *	*	*	9	**-0.74**
Klein TCA	6 / 9	* / *	*	*	9	**-0.71**
Su	* / 14	* / *	*	*	13	**0.58**
Sherwood D	43	**0.46**	11	0.47	26	**0.40**
Sherwood S	43	**0.50**	11	**0.64**	26	**0.37**
Sherwood LTMI	43	**0.68**	11	**0.62**	26	**0.65**
Brient Shal	* / 21	* / *	*	*	21	**0.38**
Zhai	27	**-0.64**	9	**-0.80**	15	**-0.73**
Tian	44	**-0.64**	11	-0.52	25	**-0.60**
Brient Alb	* / 29	* / **-0.67**	*	*	28	**-0.71**
Lipat	* / 21	* / **-0.48**	*	*	21	**-0.46**
Siler	* / 20	* / **0.54**	*	*	19	**0.54**
Cox	* / 16	* / *	*	*	22	**0.63**

TABLE 3. Values of $\sigma(T_j)/\sigma(\mathrm{ECS})$ for each $j \in A$. Weights differ for each constraint depending on the models available. Values given are means over all constraints $\pm 1\sigma$. No value is given for CMIP3 E_{kernel} because closure error for CMIP3 was absorbed into ECS values.

	Pl	WV+LR	Alb	Cld	F	E_{kernel}	E_{Taylor}
CMIP3	0.10 ± 0.00	0.32 ± 0.04	0.18 ± 0.01	1.16 ± 0.01	0.21 ± 0.02		0.18 ± 0.01
CMIP5	0.12 ± 0.02	0.43 ± 0.05	0.32 ± 0.06	1.22 ± 0.06	0.62 ± 0.04	0.45 ± 0.05	0.25 ± 0.02

TABLE 4. Assessment of proposed constraints.

Name	Credible?	Why?
Covey	no	not robust to change in ensembles
Volodin/Siler	unclear	no testable mechanism
Trenberth	no	not robust to change in ensembles, proposed mechanism is not the main source of correlation
Fasullo D	no	not robust to change in ensembles, no testable mechanism
Fasullo M	no	not robust to change in ensembles, no testable mechanism
Qu	uncertain	not robust to change in ensembles, CMIP5 correlation not due to proposed mechanism
Klein ctp-tau	unclear	no proposed mechanism
Klein TCA	unclear	no proposed mechanism
Su	unclear	no testable mechanism
Sherwood D	yes	correlation is due to proposed mechanism and region
Sherwood S	no	CMIP5 correlation is not due to the proposed mechanism
Sherwood LTMI	no	combination of credible and non-credible mechanisms
Brient Shal	yes	correlation is mainly due to proposed mechanism and region
Tian	unclear	mechanism isn't clear enough to test
Zhai/Brient Alb	yes	correlation is due to proposed mechanism and region
Lipat	uncertain	proposed region is important, but isn't the dominant source of correlation
Cox	uncertain	proposed mechanism is unrelated to individual feedbacks and regions

LIST OF FIGURES

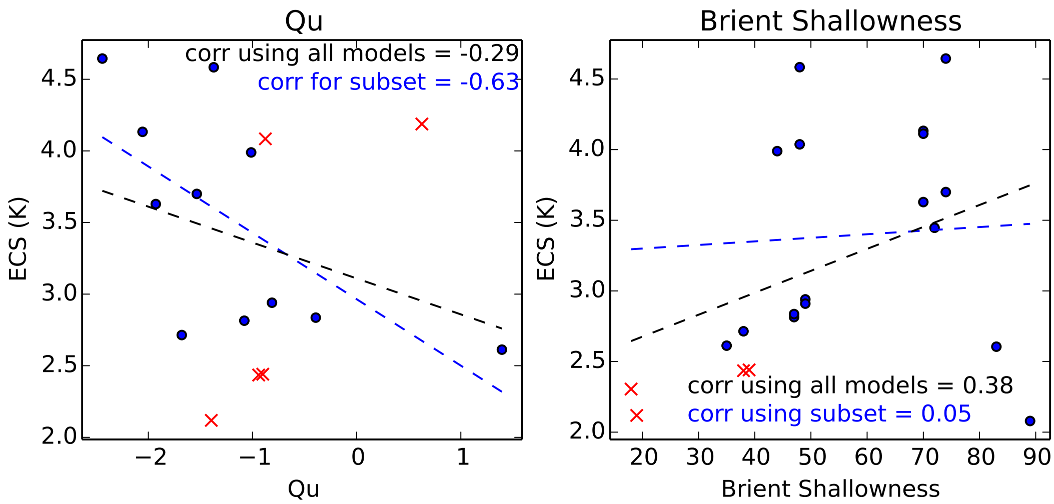


FIG. 1. Scatter plot of constraint value versus ECS for CMIP5 models passing the clear-sky linearity test for radiative kernel decomposition at the 15% level (blue) and for CMIP5 models failing this test (red). Qu is tested in the left panel, Brient Shal in the right.

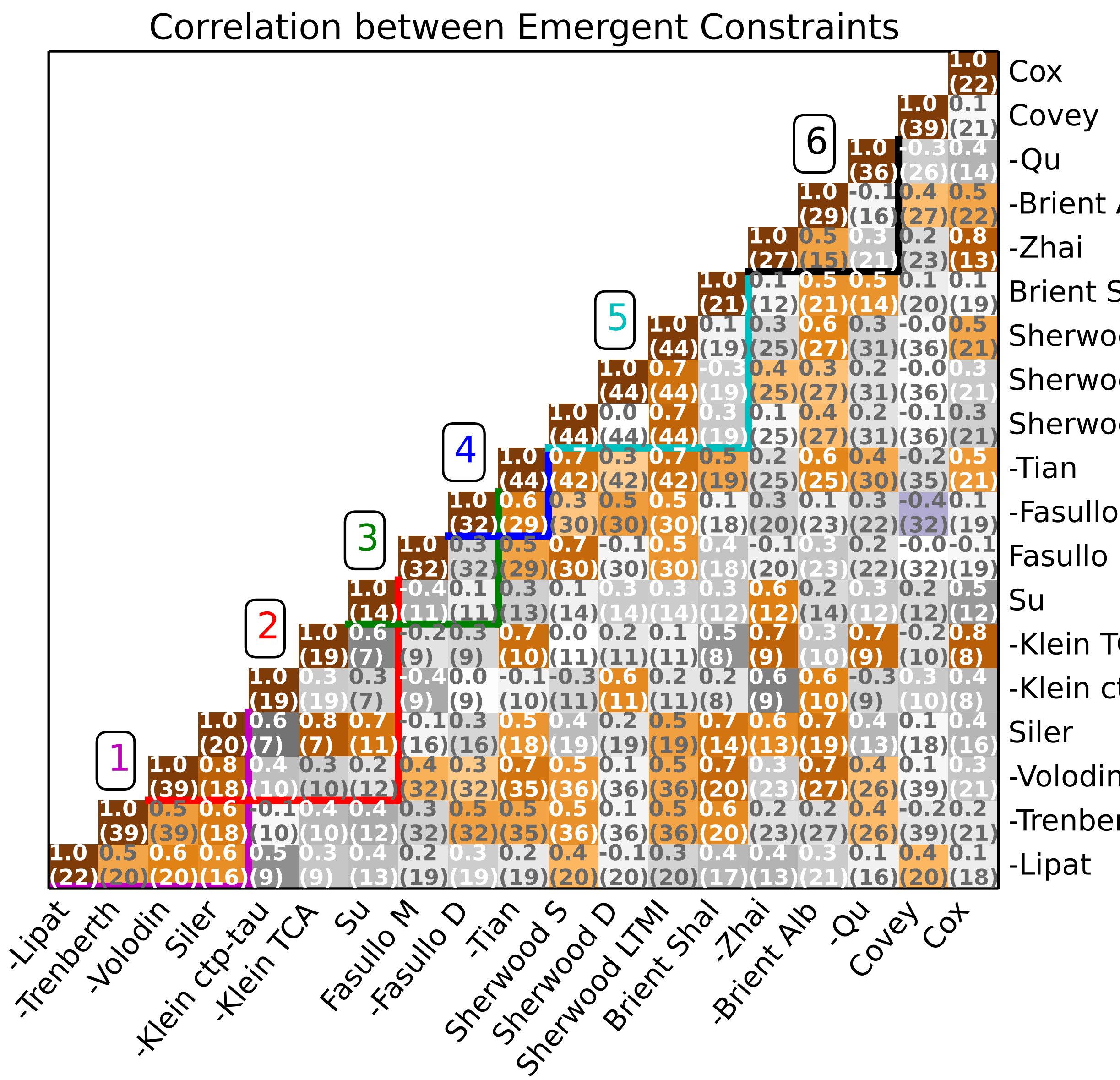


FIG. 2. Correlation between pairs of emergent constraints. Boxes with correlations significant at 90% using a 2-tailed t-test are colored, with insignificant correlations in gray. Darker shades indicate larger correlation. Positive correlations are reddish and negative correlations are blueish. In each cell, the first number is the correlation between quantities listed on the x and y axes. The number in parentheses is the number of models used in this calculation. Dark boxes (high correlation) have white text and light boxes (low correlation) have black text. The sign of emergent constraints expected to be negatively correlated with ECS has been reversed so positive values in this plot indicate both constraints have the same effect on ECS. Each correlation is calculated using data from all available CMIP3 and CMIP5 models. Colored lines and accompanying numbers reflect groups of constraints which are discussed in the text.

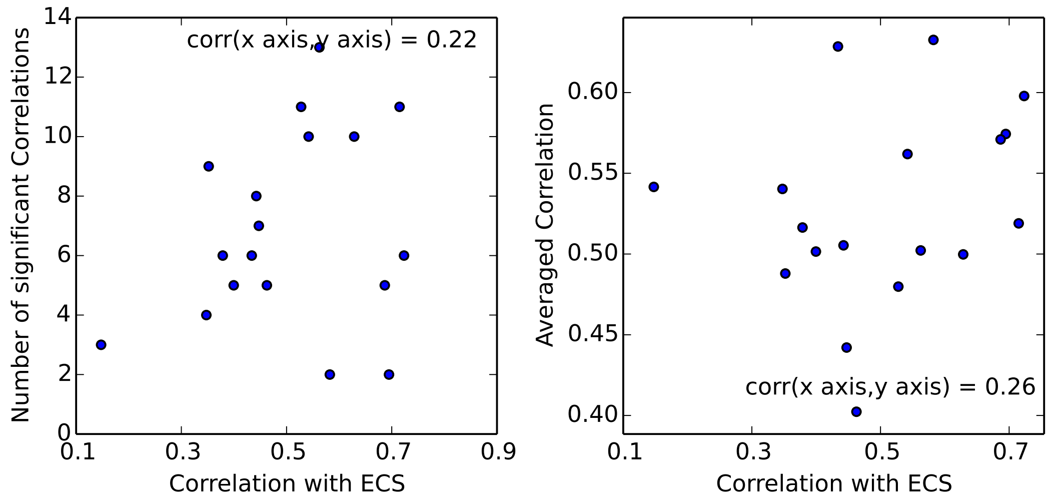


FIG. 3. Left: number of constraints a given constraint is significantly correlated with (y axis) as a function of that constraint's correlation with ECS (x axis). Right: Average correlation with ECS of all constraints significantly correlated with a given constraint (y axis) as a function of that constraint's correlation with ECS (x axis). The Covey constraint was omitted from both plots because it was an outlier.

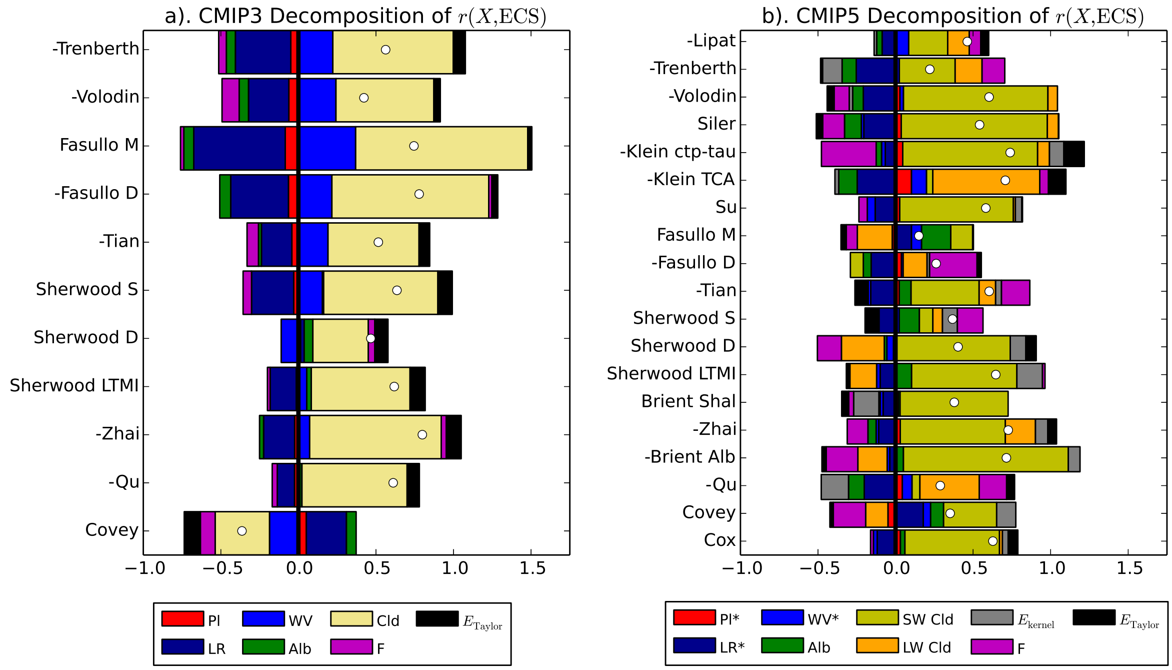


FIG. 4. Decomposition of correlation between the emergent constraints listed on the y axis and ECS into components due to forcing and feedback terms (identified in the legend). Constraints negatively correlated with ECS in their original paper are multiplied by -1 for easy comparison with other constraints. The correlation with ECS is the sum of positive and negative terms and is indicated for each emergent constraint as a white dot.

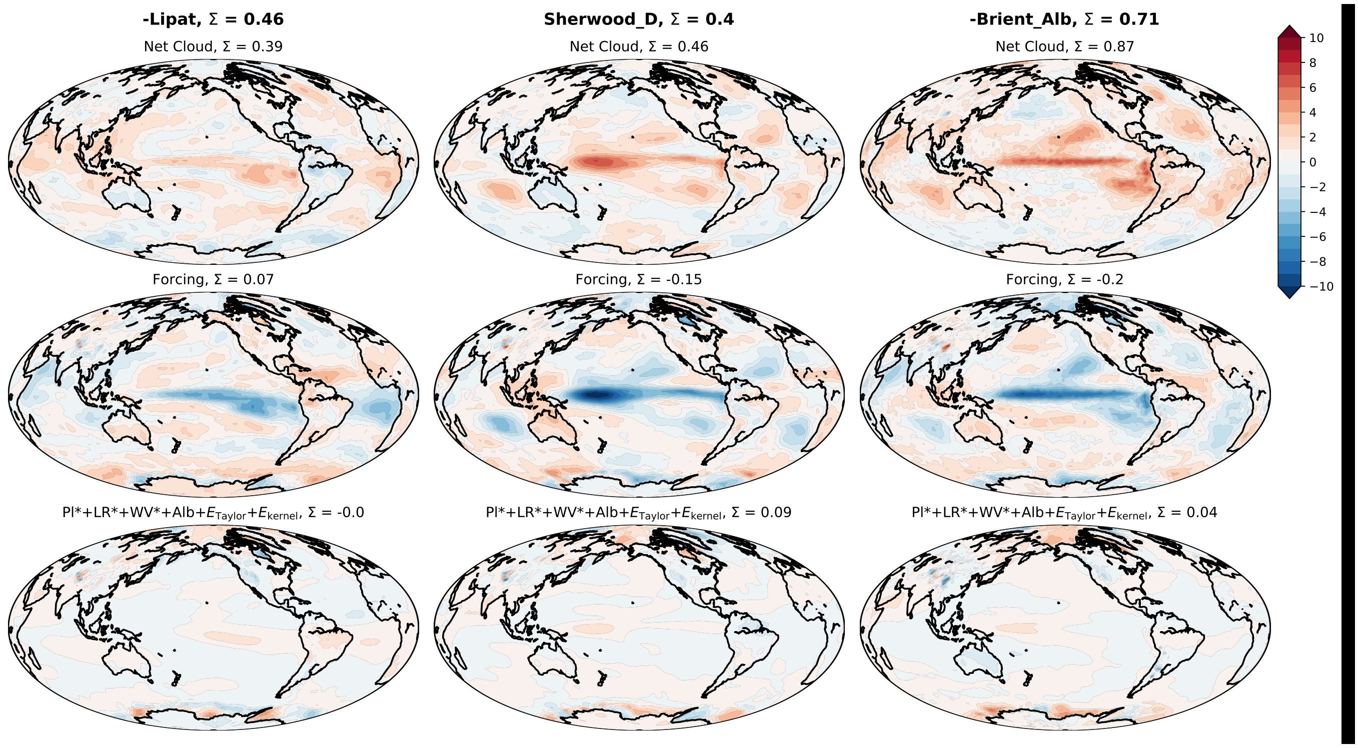
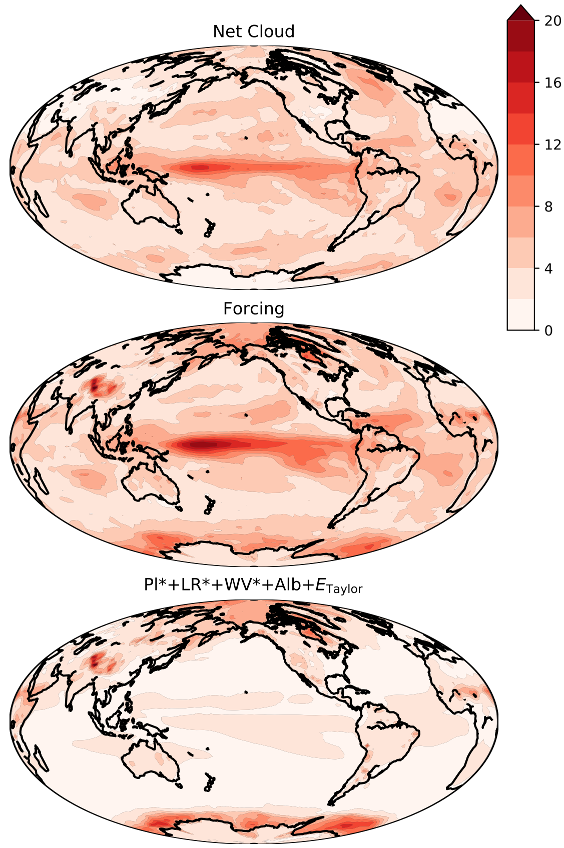


FIG. 5. Decomposition of selected emergent constraints (columns) into dominant terms (rows). Titles in bold at the top of each column list the constraint tested and the correlation of that constraint with ECS computed as the sum of all panels in that column. Sums in the title for each panel give the global sum of the geographic decomposition of that term following Eq. 6, which is comparable to the global-average contribution to that term as plotted in Fig. 4.



1163 FIG. 6. Weighting function $\sigma(T_{jk})/\sigma(\mathrm{ECS})$ from Eq. 6. Weighting maps differ slightly for different con-
1164 straints because of changes in available models; these maps are for Sherwood D. Cell-area w_k is omitted from
1165 this plot since the plot geometry already gives less space to smaller cells.