Geophysical Research Letters

RESEARCH LETTER
10.1029/2021GL095084

Investigating Parametric Dependence of Climate Feedbacks in the Atmospheric Component of CNRM-CM6-1

S. Peatier[1], B. M. Sanderson[1], L. Terray[1], and R. Roehrig[2]

[1]CERFACS, Toulouse, France, [2]CNRM, Université de Toulouse, CNRS, Toulouse, France

Key Points:

- Parameters of CNRM-CM6-1 are varied to produce an ensemble of perturbed atmospheric simulations
- We propose a method to optimize model performance conditional on a target value of climate feedback
- Model variants with comparable score to the reference version span a large climate feedback range

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

S. Peatier,
peatier@cerfacs.fr

Citation:

Peatier, S., Sanderson, B. M., Terray, L., & Roehrig, R. (2022). Investigating parametric dependence of climate feedbacks in the atmospheric component of CNRM-CM6-1. *Geophysical Research Letters*, *49*, e2021GL095084. https://doi.org/10.1029/2021GL095084

Received 8 JUL 2021
Accepted 22 APR 2022

Abstract

We sample 30 calibration parameters of the CNRM-CM6-1 atmospheric component to produce a perturbed parameter ensemble. An error score aggregating four annual-averaged metrics is used here as an example to represent the model performance. We propose a method using statistical emulators of climatic fields and feedback parameters to produce model candidates spanning the maximal range of net feedback while minimizing the error score. Optimal candidates are found for a range of climate sensitivity chosen to span 3.8–10.1 K. The candidates show large errors in the extremes of the climate sensitivity range and none of them have better scores than the released version of CNRM-CM6. However, 65% of the optimal candidates have errors within the inter-model standard deviation of the seven models participating in the Cloud Feedback Model Intercomparison Project, spanning and ECS range of (4.14–7.22 K). These optimal calibrations will be used to assess feedback diversity in coupled experiments.

Plain Language Summary

In climate simulations, an important source of uncertainty comes from the representation of subgrid-scale processes and the unconstrained parameters that it introduces. This parametric uncertainty has a potential impact on the global mean surface temperature response to a doubling in atmospheric CO_2 concentration, measured as the climate sensitivity. Model calibration is often based on expert judgment and produce a single reference version of the model, used for global projections. In this study, we test different calibrations of the CNRM-CM6-1 atmospheric component in order to explore the diversity of estimated climate sensitivities. These model versions form a set of simulations called perturbed parameter ensemble (PPE). We evaluate the performance of the PPE members by comparing them to observational data sets of different fields (temperature, precipitation, radiative fluxes) and propose an optimal subset of calibrations maximizing performance across a range of climate sensitivities. Some of these optimal candidates show comparable performance than the reference model version, though with different climate sensitivities.

1. Introduction

The equilibrium climate sensitivity (ECS) refers to the steady state change in the annual global mean surface temperature following a doubling of the atmospheric equivalent CO_2 concentration (Pachauri et al., 2014). The ECS depends on the climate feedback strength, defined as the change in radiative forcing at the top of atmosphere (TOA) following a unit change in the annual global mean surface temperature. The recent study of Sherwood et al. (2020), combine three lines of evidence including feedback process understanding, historical climate record and paleoclimate record and find a stronger constraint on ECS than previously assessed with a 66% probability range of $2.3 < \mathrm{ECS} < 4.5$ K. In contrast, in the sixth phase of the Coupled Model Intercomparison Project (CMIP), 10 general circulation models (GCMs) out of 27 have values of ECS exceeding 4.5 K (Zelinka et al., 2020), including CNRM-CM6-1, which shows an ECS of 4.9 K (Voldoire et al., 2019).

The unconstrained parameters introduced in the climate model representation of subgrid-scale processes (i.e., in so-called parameterizations) strongly impact climate sensitivity (Murphy et al., 2004; Stainforth et al., 2005). One way to explore this parametric uncertainty is to create perturbed parameter ensembles (PPE), in which parameters are varied within expert-defined ranges. PPEs can be used to better calibrate climate models (Bellprat et al., 2012; Hauser et al., 2012; Williamson et al., 2015) or to explore uncertainty associated with climate change (Piani et al., 2005; Sanderson et al., 2008; Shiogama et al., 2014). In 2005, PPE of the HadAM3 GCM found versions with a 2−11 K ECS range (Murphy et al., 2004; Stainforth et al., 2005), albeit with varying present-day climatological skill (Sanderson et al., 2008). This wide ECS range could not be reproduced with a similar PPE constructed with the community atmosphere model (Sanderson, 2011), which produced a range of 2.2–3.2 K, highlighting that two models can respond very differently to similar parameter perturbations. Recently, a PPE of HadGEM3 atmosphere-only simulations showed a larger range of climate feedbacks, from −1.80 to −0.93 W m^{-2} K^{-1} (D. D. Sexton et al., 2019; Karmalkar et al., 2019). However, the performance filtering applied to this atmosphere-only PPE limited the diversity of sensitivity in the HadGEM3-coupled perturbed simulations, especially in the low-end of the distribution (Rostron et al., 2020).

The primary purpose of this study is to propose a tractable, semi-autonomous approach for producing model candidates which span a range of responses. We explore the diversity of climate feedbacks in a PPE based on the CNRM-CM6-1 atmospheric component and assess feedback likelihood through the use of statistical emulators and observational data sets. We discuss how model performance evaluated on different fields constrains the model climate feedbacks. Finally, we propose an approach for determining an optimal subset of calibrations maximizing performance across a range of feedback values.

2. Model and Methods

This work presents a PPE using ARPEGE-Climat 6.3 (Roehrig et al., 2020), the CNRM-CM6-1 atmospheric component (Voldoire et al., 2019). ARPEGE-Climat 6.3 as used here has a spatial resolution of about 150 km, with 91 vertical levels from near 6 m to 82 km (or 0.01 hPa). The released version of this model will be referred to as the reference model version and thereafter named CNRM-CM6-1. Alternative versions of this model will be referred to as CNRM-CM. Two experiments are considered: (a) an AMIP (Eyring et al., 2015) simulation, where the model is forced by observed sea surface temperatures (SSTs) and sea ice concentrations (SICs); (b) the AMIP-future4K experiment following the cloud-feedback model inter-comparison project (CFMIP; Webb et al., 2017) protocol, where a composite SST warming pattern derived from the CMIP3 coupled models is added to the AMIP SSTs and scaled such that the global mean SST increase averaged over the ice-free oceans is +4 K. Both simulations use observed CO_2 concentrations from 1979 to 1981. They are 3-years long (1979–1981), which is a reasonable length to achieve the stabilization on the net radiative feedback in different CFMIP models (Figure S1 in Supporting Information S1). A total of 30 parameters are selected and their uncertainty limits are determined by expert solicitation with model developers (Table S1 in Supporting Information S1). The parameter values simultaneously vary and are sampled with a Latin Hypercube (LH) design where the minimum distance between points is maximized in order to have a uniform sampling of the parameter space (Urban & Fricker, 2010). After the crash of 68 simulations, the PPE considered here has 102 members.

2.1. Model Performance

In order to assess the present-day climatological performance of the PPE members, we take a subset of atmospheric variables from the AMIP simulation and compare them with observational datasets (Table S2 in Supporting Information S1; Adler et al., 2016; Loeb et al., 2018; Rohde & Hausfather, 2020). Rather than using a full root mean square error (RMSE), we chose to use a metric that focuses on the part of the error related to parameter changes.

For each variable, we consider the model annual mean climatology and calculate EOFs. In contrast to conventional EOFs, the temporal dimension is replaced by the ensemble member dimension (as in Biegler et al., 2011; Higdon et al., 2008; Sanderson et al., 2008; D. M. Sexton et al., 2012). This analysis provides compact description of the spatial variability of the ensemble variance control climate. The resulting EOFs are spatial patterns, while their principal components (PCs) are the expansion coefficients showing the projection of each ensemble member onto the respective EOF. We then project the annual mean climatology of the observations of a given variable s on the EOF basis and reconstruct the three-dimensional fields of both the models and the observations, using only the first five modes of the EOFs. The error score is then represented by the area-weighted RMSE of these reconstructed fields.

The truncation of the EOF after five modes is a subjective choice that allows a skillful emulation while applying the same truncation to all the climatic fields. This truncation explains 63% (for the precipitation) to 85%–90% (for the other variables) of the PPE variance. This metric is, by construction, strongly correlated to a standard RMSE (Figure S2 in Supporting Information S1) but has the benefit of assessing the model performance within the space defined by the chosen parameter range. For each PPE member, the error associated with each variable E_s can be standardized by the error of the CNRM-CM6-1 AMIP simulation $E_{CM6,s}$, allowing the combination of errors associated with the different variables and the estimation of a total error E_{tot}:

$$E_{\text{tot}} = \frac{1}{P}\sum_{s=1}^{P}\frac{E_s}{E_{\text{CM6},s}} \tag{1}$$

where P is the number of variables considered. As described in Table S2 in Supporting Information S1, we use four climatological fields to calculate the total error (thus $P = 4$). The aggregated metric E_{tot} represents a standardized distance between observations and model, considering the four variables described in Table S2 in Supporting Information S1. In this framework, when $E_{\text{tot}} < 1$, the given member is considered better than the reference model version.

2.2. Statistical Predictions of Model Results

The 102-member PPE is a sparse sample of a 30 dimensional parameter space, with ensemble size limited by finite computational resources. We therefore consider statistical emulators to predict, based on the perturbed parameter values, the model's outputs. The emulators are used to predict both the climatological mean state and the global net feedback in response to an idealized warming signal. Different emulators were tested, including nonlinear models as neural networks and linear models as multilinear regressions (MLR) and LASSO (least absolute shrinkage and selection operator) (Ranstam & Cook, 2018) models. Because of the reduced size of the PPE, the neural network (multilayer perceptron) tended to produce an over-fitted result (Figure S3 in Supporting Information S1).

The LASSO model is a regression method that performs both variable selection and regularization in order to enhance the prediction accuracy. The LASSO and the MLR models show comparable overall performance, with lower out-of-sample errors for the LASSO model. However, both models, and especially the LASSO, have difficulty predicting extreme feedback values (Figure S4 in Supporting Information S1). The MLR model is used in the rest of the study, in order to better predict the extreme values of feedback (Figure S5 in Supporting Information S1). The control mean state emulator is built and trained to predict the five first PCs of the EOF analysis described in the previous section, expressed as follows:

$$Y = \sum_{j=1}^{K} a_j x_j + R \tag{2}$$

with $Y = \text{PC}_i$, the ith PC of the EOF analysis (from a total $N = 5$), x_j the parameter value, a_j the regression coefficient estimated based on the training of the model, R the intercept and $K = 30$ the number of perturbed parameters. The feedback emulator is trained to predict directly the global net feedback values (in W m^{-2} K^{-1}, calculated as for Equation 2 with $Y = \lambda$). During the evaluation process, both emulators are trained on 80 members of the PPE, the 22 other members being used to estimate the out-of-sample emulator error: $OSE = \sqrt{\sum_{i=1}^{22}\frac{(pred_i - true_i)^2}{22}}$.

2.3. Optimization With Constraint and Relative Likelihood Function

The objective here is to use the statistical emulator to find optimal model versions spanning the whole range of net feedback. For this purpose, we optimize the emulated response described above in order to obtain a subset of the best calibrations given a specified feedback value. In other words, we aim to find the calibrations which minimize the error E_{tot} given by the mean state emulator, while conditioning the predicted global net feedback value to lie within a chosen bin. We use a constrained linear minimization optimizer (Virtanen et al., 2020) based on sequential least squares programming.

The optimization process is iterative and carried out as follows:

1. Step 1—The emulators are used to produce a 100,000-member ensemble. A vector of discrete global feedback values is created, of length n_λ, such that λ_i spans for all values of i the range of feedbacks seen in the ensemble in increments of 0.01 W m^{-2} K^{-1}.
2. Step 2—For each bin of feedback parameter, we select, within the 100,000 Latin Hypercube sample, the calibration which shows the lowest E_{tot} when used in the mean state emulator while lying within a given bin (i.e., $\lambda_i - 0.005$ W m^{-2} K^{-1} $< \lambda < \lambda_i + 0.005$ W m^{-2} K^{-1}). These parameter values are used to initialize the optimizer.

3. Step 3—The optimizer is used to produce $\mathbf{P_1}$—a 30 by n_λ matrix with initial optimal parameter estimates for each discrete bin of λ_i.
4. Step 4—$\mathbf{P_1}$ seems like a good guess of optimal parameter values but is very noisy. We make the hypothesis that a smooth parameter pathway exists along the feedback range (as suggested by Neelin et al. (2010)) and repeatedly apply a three-point moving average along this axis to each parameter value individually.
5. Step 5—These smoothed values of parameter are then used to reinitialize the optimizer, which produces a smoothed final estimate of optimal calibrations $\mathbf{P_2}$, validating our previous hypothesis.
6. Step 6—Finally, we sample 24 calibrations from $\mathbf{P_2}$, along the feedback range, and use it to perform a new set of simulations with the CNRM-CM atmospheric component.

After the optimization process, the total error associated with the optimal calibration depends on λ_i, which allows us to estimate the relative likelihood of the feedback parameters, within the range of λ we can simulate. The likelihood function is usually expressed as $L(\theta, x_i) \sim \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]$ (Hastie et al., 2001), considering the parametric model $\theta = (\mu, \sigma^2)$ with μ the mean and σ the standard deviation in the case of a normal distribution. Here, the error associated with the optimal calibration $E_i(\lambda_i)$ is considered drawn from a Gaussian distribution of width σ. The parametric model of λ is then written as $\theta = (E_i(\lambda_i), \sigma^2)$ and the likelihood becomes:

$$L(\theta, \lambda_i) \sim \exp\left[-\frac{E_i(\lambda_i)^2}{2 \times \sigma^2}\right] \tag{3}$$

The error $E_i(\lambda_i)$ and the Gaussian width σ remains to be defined. In this study, two estimates of feedback likelihood are proposed. The first one is based on the emulated total errors associated with the optimal calibrations. In this case, the emulated error $E_{\mathrm{tot},i}(\lambda_i)$ is scaled by the sum of the out-of-sample error of the emulator (representing the emulator error) and the standard deviation of E_{tot} within the CFMIP multimodel ensemble (representing the diversity of acceptable errors in climate models): $E(\lambda_i) = E_{\mathrm{tot},i}(\lambda_i)$ and $\sigma = (\mathrm{OSE} + \sigma_{E\mathrm{tot,CFMIP}})$.

A revised estimate of feedback likelihood is then proposed, based on the total errors associated with actual CNRM-CM runs of the optimal calibrations (as described in Step 6 of the optimization). A cubid spline under tension is used to interpolate between the errors of sampled optimal candidates along the feedback range, and this interpolation $\hat{E}_{\mathrm{tot},i}(\lambda_i)$ is used in the likelihood estimate. In this other case, the out-of-sample error of the emulator can be removed from the likelihood equation: $E(\lambda_i) = \hat{E}_{\mathrm{tot},i}(\lambda_i)$ and $\sigma = \sigma_{E\mathrm{tot,CFMIP}}$.

3. Results

3.1. Feedback Distribution

The most accurate method to estimate of the feedback parameter λ from the CMIP6 simulations is based on the increase of the atmospheric CO_2 concentration in fully coupled atmosphere-ocean models. However, the high computational cost of coupled simulations and the need of running them for hundreds of years to get an equilibrated control simulations renders the production of fully coupled PPE practically impossible. A cheaper way of estimating the total feedback parameter is to raise the sea surface temperatures (SSTs) in atmosphere-only models (Cess et al., 1990), even though it has been shown that this estimate of the total net feedback parameter can differ from the coupled-derived estimate due to differences in the clear-sky feedbacks (Ringer et al., 2014). In the PPE context of this study and given the limits of our computational resources, the total net feedback parameter λ (W m^{-2} K^{-1}) is based on the difference (Δ operator) between the forced AMIP-future4K and the control AMIP experiments. In both experiments, the net radiative budget N, in W m^{-2}, is the difference between the downward and the upward radiative fluxes at the top of atmosphere and T, in K, is the annual global mean surface temperature.

$$\lambda = \frac{\Delta N}{\Delta T} \tag{4}$$

Assuming the CNRM-CM6-1 forcing of $F_{2x} \approx 6.5$ W m^{-2} (Voldoire et al., 2019), we can estimate the climate sensitivity S (in K); (Andrews et al., 2018):

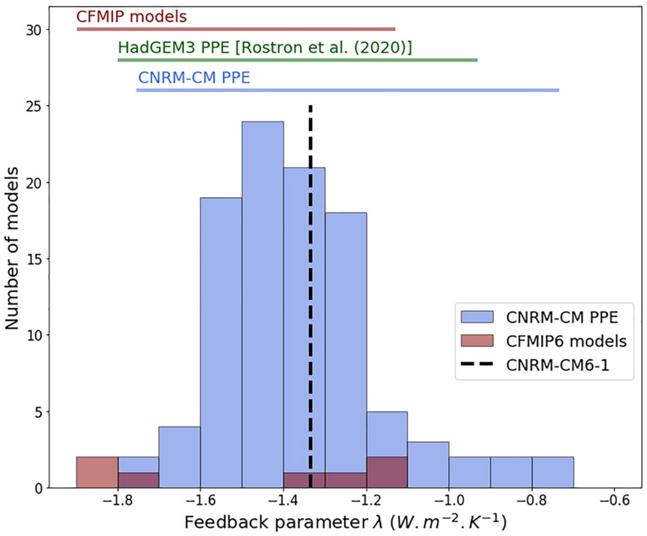

Figure 1. Feedback parameter λ (W m^{-2} K^{-1}) distribution in the ARPEGE-Climat PPE (blue histogram), the CFMIP models (red histogram), and the HadGEM3 ensemble (green line). The black dashed line shows the feedback associated with CNRM-CM6-1.

$$S = \frac{-F_{2x}}{\lambda} \tag{5}$$

The PPE provides a large diversity of climate feedback parameters (Figure 1), which range from −0.7 to −1.8 W m^{-2} K^{-1}. The CNRM-CM6-1 feedback value (−1.33 W m^{-2} K^{-1}) falls near the center of the distribution. The PPE feedback parameter range is notably wider than the one simulated by the multimodel CFMIP ensemble (from −1.7 to −1.1 W m^{-2} K^{-1}) (Webb et al., 2017). For comparison, the feedback distribution obtained in the HadGEM3 atmosphere-only PPE (Karmalkar et al., 2019) ranges from −1.7 to −0.9 W m^{-2} K^{-1}, which is comparable to the results obtained with our current PPE. We note that both PPEs could not reach the lower end of the CFMIP feedback distribution, corresponding to the lowest climate sensitivities.

3.2. Model Performance and Constraints on the Feedback Range

The PPE members indicate that some CNRM-CM versions simulate mean surface temperature, precipitation, and radiative fluxes with comparable or better skill than the reference model version (Figure 2). However, none of the CNRM-CM models show better results than CNRM-CM6-1 in the aggregated metric (Figure 2e and Figure S7e in Supporting Information S1).

We now consider potential optimal performance and its relationship to ECS in a larger emulated ensemble. The emulation of 100,000 new members predicts model versions performing as well as or better than the reference

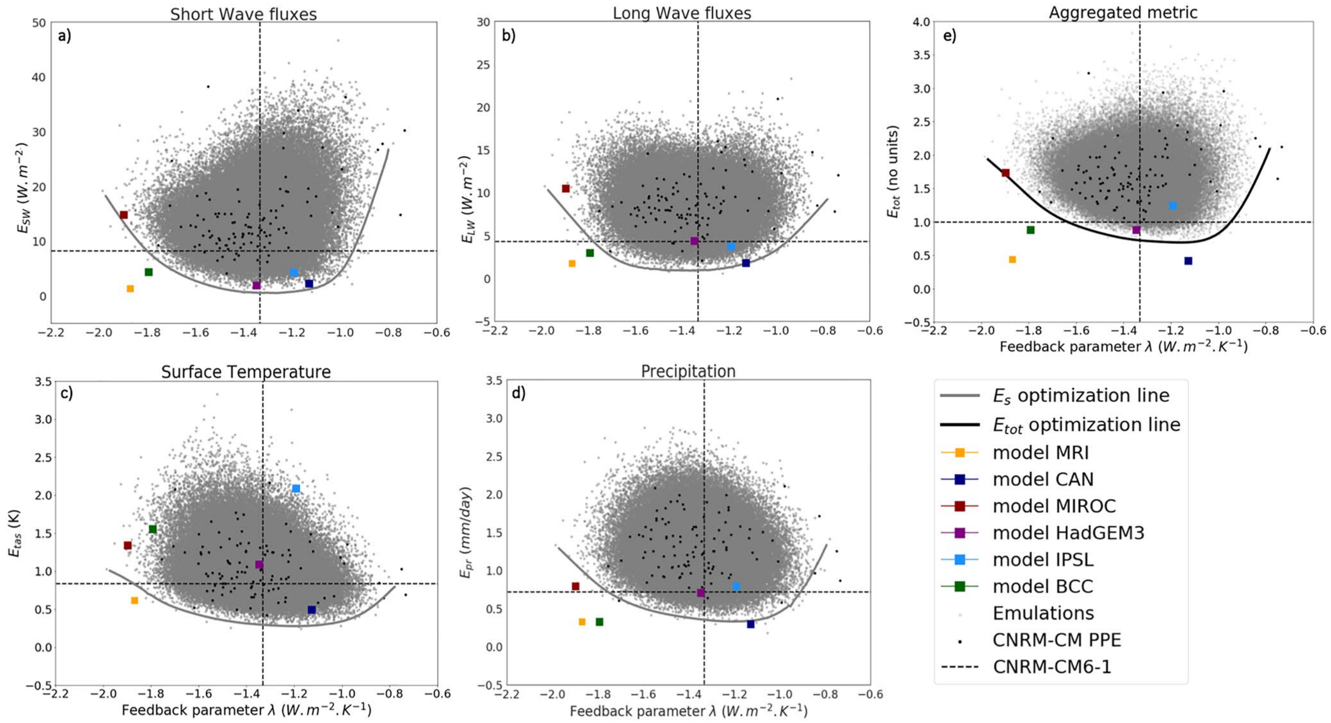

Figure 2. Errors E_s as measured between the simulations and the observations (y axis) as a function of the feedback parameter λ (x axis) in the CNRM-CM PPE (black dots), the emulated ensemble (gray dots), and the CFMIP models (colored squares). The solid lines are the results of the optimization with constraints on feedback values for the individual metrics (gray lines) and the aggregated metric (black line). The errors are estimated for (a) the short wave fluxes (SW) and (b) the long-wave fluxes (LW) at the top of the atmosphere, (c) the air surface temperature (tas) and (d) the total precipitation (pr). These four errors E_s are standardized and averaged to estimate the aggregated metric E_{tot} presented in plot (e). The black dashed lines show the errors and the feedback associated with CNRM-CM6-1.

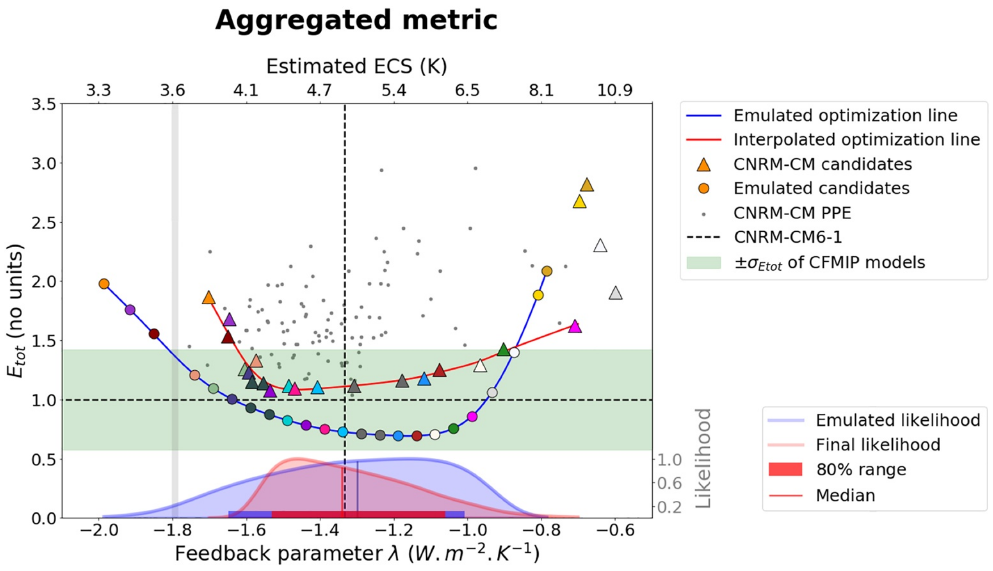

Figure 3. CNRM-CM simulations with optimized calibrations conditional on target values of feedback parameters λ. The plot shows the multivariate error E_{tot} (y axis) as a function of the feedback parameter λ (bottom x axis) for the emulated optimal candidates (blue line). Twenty-three calibrations have been selected along the feedback range (colored disks) and used in ARPEGE-Climat to produce actual climate runs (colored triangles): 1 run crashed (gray thick line) and 15 others have associated errors within the standard deviation of the CFMIP models errors (green area). Disks and triangles of the same color correspond to the same calibration. The climate sensitivity is estimated from λ, assuming the default model forcing (Equation 5). The two relative likelihoods of the feedback parameter, estimated with Equation 3, are represented by the shaded distributions, with the 10th to 90th percentiles range (80% range) and the median in bright colors. The blue distribution is the initial likelihood function, estimated from the emulated optimal candidates (blue line). The red distribution is an adjusted likelihood function based on a smooth fit (using a cubic spline under tension interpolation) of CNRM-CM optimal candidates error metric along the feedback range (red line).

version, spanning a wide range of feedback values. The total error E_{tot} described in Equation 1 is presented on Figure 2e. It is notable that CNRM-CM6-1 has the lowest E_{tot} of all the PPE members, and only 1% of the emulated members show a better performance—indicating a rather successful expert calibration of the model when a multivariate metric is considered.

The PPE results are also compared with those of the six CFMIP climate models. The global annual means of the four climatic fields of the CFMIP models are projected onto the PPE-derived EOF basis. Then, the RMSE between these projections and the observations are computed. In the aggregated metric as well as in the individual metrics, the other CFMIP models mostly show better performance than CNRM-CM6-1 and than most of the emulated CNRM-CM versions.

In Figure 2, and following Equation 5, lower feedback parameters $|\lambda|$ correspond to higher estimated ECS values. The emulated CNRM-CM versions with lower-than-reference errors span a reduced range of ECS compared to the full emulated ensemble (Figure S6 in Supporting Information S1). This range is rather centered around the CNRM-CM6-1 value for the individual metrics, but is slightly shifted to the high ECS for the aggregated metric (Figure S6e in Supporting Information S1). The slope in the optimization line of Figure 2e seems to indicate that the best performing calibration correspond to model version with ECS higher than the reference (Figure 2e).

3.3. Selection of Optimal Calibrations Along Feedback Range

This section now confirms the previous results with a subset of 23 calibrations selected among ones with the lowest E_{tot} along the feedback values λ_i. This subset is used to produce CNRM-CM candidate versions, discretely sampling the feedback range. This selection is presented in Figure 3, in which colored disks show the emulations of the selected calibrations and the colored triangles are the corresponding CNRM-CM simulations. The disks and triangles of the same color correspond to the same calibrations. There is a shift toward the lowest $|\lambda|$ between the emulations and the simulations in the lowest values of feedback. This result is consistent with the emulator's overestimate of small feedback magnitudes noted in the emulator evaluation phase (Figure S1 in Supporting Information S1). By the end of the optimization process, none of the selected calibrations perform better than the reference model, highlighting the success of the CNRM-CM6-1 calibration, which achieve the best possible performance, given this metric. However, we observe 15 model versions showing errors within the standard deviation of the multimodel ensemble CFMIP associated with a diversity of feedback.

The selection of calibrations along the E_{tot} optimization line results in candidates with relatively low total error, but unequal performance in each of the climate fields considered (Figure S8 in Supporting Information S1). This implies trade-offs between performance in this different fields. Different choices could be made depending on the variable of interest. We could imagine choosing a weighted metric that favors a good performance in some of the fields over others. However, in terms of SW fluxes and surface temperature, most members in our selection perform better than the default model. The LW fluxes and precipitation errors are rather high for our selection compared to the reference model. Nevertheless, these errors are comparable with other PPE members (Figure S8d in Supporting Information S1). Additionally, our metric does not require a near-zero top-of-atmosphere radiative budget. Therefore, most of the candidates presenting high ECS are not balanced at the top-of-atmosphere (Figure S9 in Supporting Information S1). We note in Figure S9 in Supporting Information S1, that the balanced runs are all showing estimated ECS lower than the reference and that a relationship is present between the top-of-atmosphere radiative budget and the net feedback for most candidates, except the four with the highest ECS. The presence of this clear linear constraint suggests that a single process could drive and constrain the net feedback variation in a big region of the CNRM-CM PPE. Further study is needed to identify this process and to understand what differs between model versions with low and high estimated ECS. The last four candidates showing large errors and very unbalanced top-of-atmosphere radiative budget suggest that such high ECS values are associated with an unrealistic mean climate and are therefore ruled out from the interpolation and likelihood function estimate.

The estimated likelihood profiles for λ differ depending on the use of emulated value (where the profile peaks for ECS higher than reference) or actual CNRM-CM runs (where the profile peaks for ECS lower than reference) in Equation 3. Even though the 80% ranges and quite similar, the use of actual climate simulations in the estimate of the final likelihood allows to obtain a more realistic constraint on feedback range. The median of the final likelihood profile is very close to the reference feedback, with an estimated ECS of 4.86 K. The estimated 10th to 90th percentile range of ECS is (4.26–6.15 K).

Some of the perturbed atmospheric parameters correlate with the net feedback in the optimized emulations (Figure S10 in Supporting Information S1). Among these parameters, the ones with the highest coefficient of linear regressions are the minimum drag for the convective updraft vertical velocity (RKDN), the liquid cloud heterogeneity coefficient in the shortwave spectrum (RSWINHF_LIQ) and the critical ice content for ice auto-conversion at high negative temperatures (RQICRMAX). These results confirm that the calibration of convection and cloud optical properties strongly affect the ECS (Saint-Martin et al., 2020) identified the cloud radiative response as the main driver of the ECS increase between the current and the previous version of the CNRM-CM climate model, with a predominant contribution of the tropical longwave response. According to Saint-Martin et al. (2020), the change of convection scheme between the two model versions is the main reason of the cloud changes, causing the increase in ECS. The present CNRM-CM PPE could be sharing the same characteristics, with the calibration of convection scheme and cloud properties affecting the tropical longwave response, leading to different ECS, but more work is required to confirm this hypothesis.

To conclude, 15 of the optimal CNRM-CM versions exhibit good performance in the individual metrics, with errors comparable to or better than CNRM-CM6-1 for the short-wave fluxes and surface temperature. In terms of the aggregated metric, none of the model versions show better performance than CNRM-CM6-1, highlighting the difficulty of finding a better model version than the reference. However, 15 of the optimized CNRM-CM versions have an aggregate errors within the standard deviation of the CFMIP models errors, with feedback values ranging from -1.6 to -0.9 W m^{-2} K^{-1}. This range corresponds to estimated ECS from 4.14 to 7.22 K.

4. Discussions and Conclusions

In this paper, we propose a novel approach for optimizing model performance conditional on a target value of net climate feedback, based on a 102-member perturbed physics ensemble (PPE) of the atmospheric component of CNRM-CM6-1. Our method allows the determination of potentially plausible model configurations which span a wide range of net climate response. This optimization with feedback constraint allows us to predict 23 conditionally optimized calibrations spanning a wide range of net feedbacks, which we subsequently tested through realized additional simulations. None of these 23 CNRM-CM versions exhibit better performance than CNRM-CM6-1, but 15 of them have error scores within the standard deviation of CFMIP models errors. These 15 CNRM-CM versions span an ECS range of (4.14–7.22 K) and the likelihood profile of their ECS suggests an 80% range of (4.26–6.15 K).

In all of the individual metrics as well as in the aggregated metric, the other CFMIP models seem to perform better than the CNRM-CM emulated versions. This specificity could be the result of the basis choice. By projecting the different global means of the CFMIP models onto the CNRM-CM PPE-derived EOF basis, we do not consider any parts of the model errors orthogonal to this chosen basis. This projection might give CFMIP models an unfair advantage by hiding an important part of their error. This point will be investigate in future study.

The present work does not seek to recommend calibrations suitable for operational use, given optimal candidates are conditional on a subjective and nonextensive multivariate metric used here to demonstrate the technique. Further study will consider how constraints result in trade-offs between performance in different variables. For optimization purposes, we have made defensible but subjective choices in terms of the variables chosen, the EOF truncation length, the nature of the feedback constraints for the optimization, and the error metric. Moreover, the top-of-atmosphere radiative balance is not yet constrained by the method, which results in imbalanced model versions in the final candidate selection. These are all elements which an operational tuning procedure may wish to expand on, but are out of the scope of the present study.

The metric considered here suggests the existence of model variants with comparable climatological performance to CNRM-CM6-1, with both lower and higher net feedback strengths. However, the model with higher climate sensitivities tend to exhibit strong radiative unbalanced at the top-of-atmosphere. To further understand this result, future studies should consider a broader range of metrics, ideally with a range of CMIP6 atmospheric models.

In conclusion, the method proposed here succeeded in producing model candidates spanning a range of climate responses. Our work sets out to understand the inconsistency between observational evidence which finds high values of ECS to be increasingly less likely (Sherwood et al., 2020), and a new generation of models in which many members exhibited ECS values outside of the previously assessed range. Our model optimization exercise suggests that, for the simulation of mean climatological performance, optimal model configurations may exhibit lower or higher values of ECS. Understanding how these findings integrate into a general assessment of ECS is thus a priority for future research.

Data Availability Statement

CMIP6 data are available through a distributed data archive developed and operated by the Earth System Grid Federation (ESGF): https://esgf-node.ipsl.upmc.fr/search/cmip6-ipsl/. Data files and code for reproducing the plots are available here: https://doi.org/10.5281/zenodo.6077885.

Acknowledgments
This work is partly funded by the French National Research Agency, project no. ANR-17-MPGA-0016. Benjamin Sanderson is an affiliate scientist with the National Center for Atmospheric Research, sponsored by the National Science Foundation. The authors thank the CNRM-CERFACS modeling group for developing and supporting the CNRM-CM6-1 model. In particular Marie-Pierre Moine, for the technical support provided for the simulations.

References

Adler, R., Wang, J., Sapiano, M., Huffman, G., Chiu, L., Xie, P., et al. (2016). *Global Precipitation Climatology Project (GPCP) Climate Data Record (CDR), Version 2.3 (Monthly)* (Vol. 10, p. V56971M6). National Centers for Environmental Information.

Andrews, T., Gregory, J. M., Paynter, D., Silvers, L. G., Zhou, C., Mauritsen, T., et al. (2018). Accounting for changing temperature patterns increases historical estimates of climate sensitivity. *Geophysical Research Letters*, *45*(16), 8490–8499. https://doi.org/10.1029/2018gl078887

Bellprat, O., Kotlarski, S., Lüthi, D., & Schär, C. (2012). Objective calibration of regional climate models. *Journal of Geophysical Research*, *117*(D23). https://doi.org/10.1029/2012jd018262

Biegler, L., Biros, G., Ghattas, O., Heinkenschloss, M., Keyes, D., Mallick, B., et al. (2011). *Large-scale Inverse Problems and Quantification of Uncertainty* (Vol. 712). John Wiley & Sons.

Cess, R. D., Potter, G., Blanchet, J., Boer, G., Del Genio, A., Deque, M., et al. (1990). Intercomparison and interpretation of climate feedback processes in 19 atmospheric general circulation models. *Journal of Geophysical Research*, *95*(D10), 16601–16615. https://doi.org/10.1029/jd095id10p16601

Eyring, V., Bony, S., Meehl, G. A., Senior, C., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2015). Overview of the coupled model intercomparison project phase 6 (CMIP6) experimental design and organisation. *Geoscientific Model Development Discussions*, *8*(12).

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning* (Vol. 1, No. 10). Springer series in statistics.

Hauser, T., Keats, A., & Tarasov, L. (2012). Artificial neural network assisted Bayesian calibration of climate models. *Climate Dynamics*, *39*(1–2), 137–154. https://doi.org/10.1007/s00382-011-1168-0

Higdon, D., Gattiker, J., Williams, B., & Rightley, M. (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association*, *103*(482), 570–583. https://doi.org/10.1198/016214507000000888

Karmalkar, A. V., Sexton, D. M., Murphy, J. M., Booth, B. B., Rostron, J. W., & McNeall, D. J. (2019). Finding plausible and diverse variants of a climate model. Part ii: Development and validation of methodology. *Climate Dynamics*, *53*(1), 847–877. https://doi.org/10.1007/s00382-019-04617-3

Loeb, N. G., Doelling, D. R., Wang, H., Su, W., Nguyen, C., Corbett, J. G., et al. (2018). Clouds and the Earth's radiant energy system (CERES) energy balanced and filled (EBAF) top-of-atmosphere (TOA) edition-4.0 data product. *Journal of Climate*, *31*(2), 895–918.

Murphy, J. M., Sexton, D. M., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., & Stainforth, D. A. (2004). Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, *430*(7001), 768–772. https://doi.org/10.1038/nature02771

Neelin, J. D., Bracco, A., Luo, H., McWilliams, J. C., & Meyerson, J. E. (2010). Considerations for parameter optimization and sensitivity in climate models. *Proceedings of the National Academy of Sciences*, *107*(50), 21349–21354. https://doi.org/10.1073/pnas.1015473107

Pachauri, R., Meyer, L., & Team, C. W. (2014). Annex ii: Glossary [Mach, KJ, s. Planton & C. In von stechow (Eds.)]. *Climate change 2014: Synthesis report. Contribution of working groups I, II and III to the fifth assessment report of the intergovernmental panel on climate change* (pp. 117–130).

Piani, C., Frame, D., Stainforth, D., & Allen, M. (2005). Constraints on climate change from a multi-thousand member ensemble of simulations. *Geophysical Research Letters*, *32*(23). https://doi.org/10.1029/2005gl024452

Ranstam, J., & Cook, J. (2018). Lasso regression. *Journal of British Surgery*, *105*(10), 1348. https://doi.org/10.1002/bjs.10895

Ringer, M. A., Andrews, T., & Webb, M. J. (2014). Global-mean radiative feedbacks and forcing in atmosphere-only and coupled atmosphere-ocean climate change experiments. *Geophysical Research Letters*, *41*(11), 4035–4042. https://doi.org/10.1002/2014gl060347

Roehrig, R., Beau, I., Saint-Martin, D., Alias, A., Decharme, B., Guérémy, J.-F., et al. (2020). The CNRM global atmosphere model Arpege-Climat 6.3: Description and evaluation. *Journal of Advances in Modeling Earth Systems*, *12*(7), e2020MS002075. https://doi.org/10.1029/2020ms002075

Rohde, R. A., & Hausfather, Z. (2020). The Berkeley Earth land/ocean temperature record. *Earth System Science Data*, *12*(4), 3469–3479. https://doi.org/10.5194/essd-12-3469-2020

Rostron, J. W., Sexton, D. M., McSweeney, C. F., Yamazaki, K., Andrews, T., Furtado, K., et al. (2020). The impact of performance filtering on climate feedbacks in a perturbed parameter ensemble. *Climate Dynamics*, *55*, 521–551. https://doi.org/10.1007/s00382-020-05281-8

Saint-Martin, D., Geoffroy, O., Voldoire, A., Cattiaux, J., Brient, F., Chauvin, F., et al. (2020). Tracking changes in climate sensitivity in CNRM climate models. *Journal of Advances in Modeling Earth Systems*, *13*(6), e2020MS002190.

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. https://doi.org/10.1175/2010jcli3498.1

Sanderson, B. M., Knutti, R., Aina, T., Christensen, C., Faull, N., Frame, D., et al. (2008). Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes. *Journal of Climate*, *21*(11), 2384–2400. https://doi.org/10.1175/2008jcli1869.1

Sexton, D., Karmalkar, A., Murphy, J., Williams, K., Boutle, I., Morcrette, C., et al. (2019). Finding plausible and diverse variants of a climate model. Part 1: Establishing the relationship between errors at weather and climate time scales. *Climate Dynamics*, *53*(1), 989–1022. https://doi.org/10.1007/s00382-019-04625-3

Sexton, D. M., Murphy, J. M., Collins, M., & Webb, M. J. (2012). Multivariate probabilistic projections using imperfect climate models part I: Outline of methodology. *Climate Dynamics*, *38*(11–12), 2513–2542. https://doi.org/10.1007/s00382-011-1208-9

Sherwood, S., Webb, M., Annan, J., Armour, K., Forster, P., Hargreaves, J., et al. (2020). An assessment of Earth's climate sensitivity using multiple lines of evidence. *Reviews of Geophysics*, *58*, e2019RG000678.

Shiogama, H., Watanabe, M., Ogura, T., Yokohata, T., & Kimoto, M. (2014). Multi-parameter multi-physics ensemble (MPMPE): A new approach exploring the uncertainties of climate sensitivity. *Atmospheric Science Letters*, *15*(2), 97–102. https://doi.org/10.1002/asl2.472

Stainforth, D. A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D. J., et al. (2005). Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, *433*(7024), 403–406. https://doi.org/10.1038/nature03301

Urban, N. M., & Fricker, T. E. (2010). A comparison of Latin hypercube and grid ensemble designs for the multivariate emulation of an Earth system model. *Computers & Geosciences*, *36*(6), 746–755. https://doi.org/10.1016/j.cageo.2009.11.004

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SCIPY 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-020-0772-5

Voldoire, A., Saint-Martin, D., Sénési, S., Decharme, B., Alias, A., Chevallier, M., et al. (2019). Evaluation of CMIP6 deck experiments with CNRM-cm6-1. *Journal of Advances in Modeling Earth Systems*, *11*(7), 2177–2213. https://doi.org/10.1029/2019ms001683

Webb, M. J., Andrews, T., Bodas-Salcedo, A., Bony, S., Bretherton, C. S., Chadwick, R., et al. (2017). The cloud feedback model intercomparison project (CFMIP) contribution to CMIP6. *Geoscientific Model Development*, *10*(1), 359–384.

Williamson, D., Blaker, A. T., Hampton, C., & Salter, J. (2015). Identifying and removing structural biases in climate models with history matching. *Climate Dynamics*, *45*(5), 1299–1324. https://doi.org/10.1007/s00382-014-2378-z

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., & Taylor, K. E. (2020). Causes of higher climate sensitivity in CMIP6 models. *Geophysical Research Letters*, *47*(1), e2019GL085782. https://doi.org/10.1029/2019gl085782