Calibrating Simple Climate Models to Individual Earth System Models: Lessons Learned From Calibrating Hector

Kalyn Dorheim1, Robert Link1, Corinne Hartin1, Ben Kravitz1,2, and Abigail Snyder1

1Joint Global Change Research Institute, Pacific Northwest National Laboratory, Richland, WA, USA, 2Department of Earth and Atmospheric Sciences, Indiana University, Bloomington, IN, USA

Abstract Simple climate models (SCMs) are computationally efficient and capable of emulating global mean output of more complex Earth system models (ESMs). In doing so, SCMs can play a critical role in climate research as stand-ins for the computationally more expensive models, especially in studies involving low, spatial, and/or temporal resolution, providing more computationally efficient sources of climate data. Here we use Hector v2.5.0 to emulate the multiforcing historical and RCP scenario output for 31 concentration and seven emission-driven ESMs. When calibrating Hector, sufficient calibration data must be used to constrain the model; otherwise, climate and/or carbon parameters affecting physical processes may be able to trade off with one another, allowing for solutions to use physically unreasonable fitted parameter values as well as limiting the application of the SCM as an emulator. We also present a novel methodology that uses the ESM range as a calibration data, which can be adopted when faced with missing variable output from a specific model.

Plain Language Summary We calibrated a Simplified Climate Model (SCM), Hector, to efficiently reproduce (emulate) the output of more complex climate models. Using SCMs as emulators is a common practice in the climate science and climate-economic research communities where the aim is to simulate long-term changes in climate for other (nonclimate) sectors, meaning that fine-scale climate simulation requires too much computational effort. In order to use Hector as an emulator that satisfactorily reproduces an ESM's modeled climate, we found that Hector must be constrained with multiple output variables. Otherwise, parameters can trade off with arbitrarily large values; for example, an unrealistically high climate sensitivity balanced by unrealistically large aerosol forcing can produce temperature change comparable to those produced by an ESM. We also document a new method that can use the multimodel range as calibration data if need be.

1. Introduction

Simple climate models (SCMs) and Earth system models (ESMs) represent the carbon cycle and climate system at different scales. Because ESMs represent more physical and chemical processes in more detail and operate on fine temporal and spatial resolutions, they are computationally expensive and therefore can only be run for a limited length of time for a limited number of scenarios. Conversely, SCMs trade off spatial and temporal resolution to reduce computational expense by containing simplified representations of the major processes governing the carbon-climate system (Hartin et al., 2015; Meinshausen et al., 2011). Given that SCMs are computationally inexpensive to run, they can be used to generate large ensembles of results where high-resolution climate data are not needed for the problem at hand, such as many classes of uncertainty quantification studies (Clarke et al., 2014; Fawcett et al., 2015; Meinshausen et al., 2009). SCMs can also be used in model coupling exercises with economic models (Schultz & Kasting, 1997; Smith & Edmonds, 2006; van Vuuren et al., 2011). Additionally, SCMs can be tuned or calibrated to emulate the more complex ESMs. When acting as emulators, SCMs can be used to generate novel scenarios that were not conducted by the ESMs, improve our understanding of the components of more complex models, and examine the structural uncertainty of ESMs (Meinshausen et al., 2011).

There are a range of calibration procedures within the SCM community. The most common approach used to calibrate SCMs to individual ESMs is a least squares regression (Caldeira & Myhrvold, 2013;
Geoffroy et al., 2013; Huntingford et al., 2009; Khabbazan & Held, 2019; Meinshausen et al., 2011). In a least squares regression, an optimization routine solves for the optimal parameter fits that minimize a goodness-of-fit metric between the SCM output data and the ESM calibration data for Earth system variables of interest. Mean or total square error are often the metrics used, and different goodness-of-fit metrics can yield different degrees of calibration success (Caldeira & Myhrvold, 2013; Huntingford et al., 2009; Khabbazan & Held, 2019). Khabbazan and Held (2019), Caldeira and Myhrvold (2013), and Huntingford et al. (2009) used only global mean temperature data in their calibration procedures. While this can produce a good value for the goodness-of-fit metric, Huntingford et al. (2009) also suggested that the calibration procedure must be able identify a unique solution, for which using temperature data alone may be insufficient. Other papers, such as Meinshausen et al. (2011), used both temperature and ocean heat flux as calibration data and were able to find unique calibration solutions for their SCM for concentration-driven ESM simulations.

In this paper we tested four SCM calibration procedures on Hector v 2.5.0. Our objective was to design a calibration protocol that successfully calibrates Hector as an ESM emulator. To this end, we defined several criteria that indicate a satisfactory calibration. First, SCM results should capture the aggregate behavior of the ESM output. Second, the optimization routine should be able to identify a unique solution. Finally, the fitted parameter values should be physically reasonable. The uniqueness criterion is essential for the purpose here, which is to calibrate Hector to a specific ESM for future use in chains of coupled modeling experiments. If we were trying to calibrate Hector to historical observations (i.e., to the physical Earth system), we would want our calibration to capture the uncertainty in many Earth system variables. However, in calibrating Hector to act as an emulator for a specific ESM, a single set of parameters that allow Hector to reproduce all calibration data for that specific ESM is needed. When a unique solution is not identified, there are multiple equally good parameter fits. Supporting multiple model fits that are equally good is not only impractical but may limit the ability to use Hector as an emulator for novel scenarios.

The results from this paper provide best practices for calibrating Hector, which will help inform future Hector calibration exercises. The procedures and concepts described herein may also be relevant to other exercises in calibrating SCMs to ESMs, particularly when calibrating to an individual ESM with missing variables. We outline a procedure that incorporates the ESM multimodel range as additional calibration data when an individual ESM is missing output variable data for that are otherwise important for the calibration protocol.

2. Hector v2.5.0

Hector is an open-source, object-oriented, reduced-form global carbon-cycle climate model (Hartin et al., 2015). Hector operates on an annual time step taking in emissions of greenhouse gases, non-CO₂ gases, and aerosols. Hector converts emissions to concentrations where needed and calculates the global change in radiative forcing from which global mean temperature and other climate and carbon cycle variables are calculated. Hector v2.5.0 (Vega-Westhoff et al., 2019) incorporates ocean-atmosphere heat flux dynamics using a one-dimension diffusive ocean heat flux model, DOECLIM (Kriegler, 2005). The carbon-cycle component of Hector is divided into a one box atmosphere, a three box terrestrial system with vegetation, soil, and detritus, where net primary production is a function of atmospheric CO₂ and heterotrophic respiration is a function of global mean temperature and a four box ocean.

The model is driven with exogenous global emissions or atmospheric concentrations. In the emissions-driven mode, Hector is driven with CO₂, non-CO₂ greenhouse gases, and aerosol emissions as time series inputs. When driven in emissions mode, carbon-cycle feedbacks are active, such as CO₂ fertilization and heterotrophic respiration. Alternatively, Hector can be run in the concentration-driven mode, where instead of using CO₂ emissions, CO₂ concentrations are prescribed, whereby design the carbon cycle is not active. These two modes, or model configurations, are comparable to the ESM Coupled Model Intercomparison Project Phase 5 (CMIP5) emissions and concentration-driven runs (Taylor et al., 2012). From an emulation perspective, we calibrate to the two different kinds of runs separately, as if they were different models of different physical processes (Hourdin et al., 2017).

In this paper we calibrated both the concentration and emissions-driven versions of Hector. During the concentration-driven experiments, four Hector climate parameters were tuned: climate sensitivity $S$, ocean heat diffusivity $\kappa$, the aerosol forcing scaling factor $\alpha_a$, and the volcanic forcing scaling factor $\alpha_v$. $S$ ($^\circ$C)
characterizes Hector's long-term temperature changes in response to changes in CO₂ concentrations, and \( \kappa \) (cm² s⁻¹) is the coefficient in the diffusion equation that controls the rate at which heat diffuses into the deep ocean. The remaining climate parameters, \( \alpha_a \) and \( \alpha_v \), are unitless scale factors used to adjust Hector's temperature sensitivity to aerosol and volcanic radiative forcing, respectively. During the emissions-driven calibration experiments, three Hector carbon-cycle parameters were calibrated in addition to the four Hector climate parameters previously mentioned. The three carbon-cycle parameters are preindustrial CO₂ concentration \( C_0 \) (ppmv), CO₂ fertilization factor \( \beta \) (unitless), and heterotrophic respiration temperature sensitivity factor \( Q_{10} \) (unitless). Although the preindustrial CO₂ concentration is well constrained in observations, preindustrial CO₂ concentrations vary by about 10 ppmv across individual emission-driven ESM runs (Friedlingstein et al., 2014). In order for Hector to accurately emulate the emissions-driven ESM CO₂ concentrations over time, Hector must capture the preindustrial CO₂ concentration of the ESM. Therefore, preindustrial CO₂ concentration was included as a calibration parameter.

Parameters within Hector’s ocean carbon cycle are not directly calibrated in this exercise. Given the active marine carbonate chemistry within the ocean, there are numerous parameters that are more difficult to constrain based on more data availability. Instead of direct calibration, the ocean carbon cycle is indirectly captured via the other parameters as they change the climate and carbon cycle. The ocean carbon flux is a function of the atmospheric CO₂ concentrations and surface temperature, which vary based on the other parameters. As in Meinshausen et al. (2011), we strive to avoid overfitting to minimize the number of parameters being calibrated.

### 3. Calibration Data Processing

In this paper, Hector parameters were calibrated to output from individual ESMs in participating in the Coupled Model Intercomparison Project Phase 5 (CMIP5) (Taylor et al., 2012). CMIP5 results from the historical, rcp26, rcp45, rcp60, rcp85, esmHistorical, and esmrcp85 scenarios were used to produce the global mean temperature, atmospheric CO₂ concentration, and ocean heat flux ESM calibration data. First, CMIP5 outputs were annually and globally averaged to match the resolution of Hector results. Since a limited number of ESMs provided ocean heat flux data, we calculated ocean heat flux using six atmospheric variables: incoming shortwave and longwave radiation at the surface (\( S_d \) and \( L_d \)), upward shortwave and longwave radiation at the surface (\( S_u \) and \( L_u \)), and sensible and latent heat flux (\( F_s \) and \( F_l \)). For each grid cell at each time step, total heat flux \( F_T \) is equal to

\[
F_T = S_d + L_d - L_u - S_u - F_s - F_l.
\]

Mean annual ocean heat flux calculations were isolated to the ocean by using land masks for each model. Incomplete data were discarded to ensure consistent and comparable calibration across all models. Additionally, ESMs that did not have historical temperature results and temperature results for at least one future scenario were also excluded from the calibration experiments.

### 4. Calibration Experiments

We used the downhill simplex algorithm (Nelder & Mead, 1965) to minimize the distance (measured as mean squared error, MSE; also referred to as the goodness of fit) between standardized ESM and Hector output data. Standardizing the ESM and Hector data enables cross-comparison between different variables by preventing the optimal goodness of fit from being dominated by output from a single variable. We also adopted an equal weighting methodology to prevent scenarios with multiple ensemble members from having a greater influence on optimal goodness-of-fit. By minimizing the multiscenario, multivariable MSE average, the optimization routine ensured that all variable and scenario output was being fitted equally well.

### 5. Concentration-Driven Fits

Of the models that participated in CMIP5 (Taylor et al., 2012), we were able to gather concentration-driven temperature and ocean heat flux results from 31 ESMs (listed in Table 1). During the concentration-driven runs, the models were forced with prescribed atmospheric CO₂ concentrations, non-CO₂ emissions, and aerosols for the historical period (historical experiment) and four alternative future scenarios that span...
Table 1

| Model          | Temperature only | Temperature and heat flux | Temperature only | Temperature and heat flux |
|----------------|------------------|----------------------------|------------------|----------------------------|
| ACCESS1-0      | 2.21             | 2.17                       | 0.70             | 1.28                       |
| ACCESS1-3      | 1.95             | 2.01                       | 1.54             | 1.27                       |
| CanESM2        | 1.93             | 1.88                       | 0.60             | 1.04                       |
| CCSM4          | 1.37             | 1.13                       | 0.33             | 1.03                       |
| CESM1-BGC      | 1.23             | 1.10                       | 0.41             | 0.99                       |
| CESM1-CAM5     | 2.00             | 1.93                       | 0.79             | 0.95                       |
| CESM1-WACCM    | 1.08             | 0.97                       | 0.17             | 0.62                       |
| CMCC-CESM      | 2.44             | 2.41                       | 0.06             | 0.21                       |
| CMCC-CM        | 1.80             | 1.75                       | −0.36            | −0.38                      |
| CMCC-CMS       | 1.83             | 1.79                       | −0.14            | −0.19                      |
| CNRM-CM5       | 1.27             | 1.35                       | 1.28             | 3.15                       |
| CSIRO-Mk3-6-0  | 1.57             | 1.69                       | 2.18             | 1.44                       |
| CSIRO-Mk3L-1-2 | 0.75             | 1.13                       | 2.91             | 1.09                       |
| EC-EARTH       | 1.96             | 1.92                       | 0.56             | 1.44                       |
| FGOALS-g2      | 1.94             | 1.93                       | −0.06            | 0.07                       |
| GFDL-CM3       | 2.22             | 2.10                       | 0.77             | 1.18                       |
| GFDL-ESM2G     | 1.41             | 1.26                       | 1.01             | 2.06                       |
| GISS-E2-H      | 0.28             | 0.13                       | 1.80             | 1.45                       |
| GISS-E2-H-CC   | −0.81            | 0.46                       | 4.66             | 1.82                       |
| GISS-E2-R      | 1.35             | 0.85                       | 0.68             | 1.46                       |
| GISS-E2-R-CC   | 1.16             | 1.08                       | 0.63             | 1.31                       |
| inmcm4         | 0.83             | 1.16                       | 1.80             | 0.59                       |
| IPSL-CM5A-MR   | 1.10             | 1.18                       | 1.10             | 0.92                       |
| IPSL-CM5B-LR   | −0.44            | 0.51                       | 4.51             | 1.28                       |
| MIROC-ESM      | 1.45             | 1.15                       | 1.96             | 1.84                       |
| MIROC-ESM-CHEM | 1.72             | 1.16                       | 0.79             | 1.38                       |
| MIROC5         | 1.38             | 1.24                       | 1.67             | 1.55                       |
| MPI-ESM-LR     | 1.57             | 1.47                       | 0.44             | 1.08                       |
| MPI-ESM-MR     | 1.82             | 1.77                       | 0.55             | 1.21                       |
| MRI-CGCM3      | 1.72             | 1.66                       | 1.36             | 1.00                       |
| NorESM1-ME     | 1.97             | 1.77                       | 0.11             | 0.48                       |

Note. The temperature-only calibration experiment used ESM temperature to tune Hector climate parameters, whereas the temperature-heat flux calibration experiment used a combination of ESM temperature and ocean heat flux data to tune Hector climate parameters.

DORHEIM ET AL. 4 of 14
Figure 1. Emulated Hector temperature for CCSM4, GFDL-CM3, and MRI-CGCM3 using temperature-only fitted parameterizations (blue) compared with ESM temperature output (gray) each individual gray line represents ESM output from a different ensemble realization. These three fits were selected out of 31 Hector fits (representing the 31 ESMs) to illustrate a range in Hector’s performance as an emulator. All Hector fits are visualized in the supporting information.

Three of the 31 fits were selected and visualized in Figure 1 to illustrate a range of Hector’s performance as an emulator. Visualizations for the remaining fits are provided in the supporting information. The fitted values of $S$, $\kappa$, $\alpha_v$, and $\alpha_a$ for all 31 models are reported in Table 1.

Even though Hector successfully emulates ESM temperature (Figure 1), the calibration results are unsatisfactory because they have physically unreasonable parameter values. Knutti et al. (2017) and Stocker et al. (2013) estimate that $S$ is unlikely ($\leq 5\%$ probability) to lie outside of the interval $(1, 6) ^\circ$C. Diffusion of heat into the ocean must be strictly positive, so we consider an additional constraint that $\kappa \geq 10^{-6}$. As indicated in Figure 2, not all of the fitted values of $S$ and $\kappa$ fall within or near these reasonable ranges. In fact, only 12 of the 31 temperature-only calibration results are reasonable parameter values for all parameters considered.

A primary issue is that the temperature-only calibrations fail to identify unique solutions. The inability to identify unique solutions limits the practical use of Hector as an emulator of an individual ESM for novel emission pathways. When the calibration protocol fails to produce unique solutions, it means the system is insufficiently constrained. This can result in equally good calibration fits that use both physically reasonable and physically unreasonable parameters for Hector. We explored a series of temperature-only concentration-driven calibration experiments in which $S$, $\alpha_v$, and $\alpha_a$ were calibrated for the same three models (illustrated in Figure 1) at fixed values of $\kappa$. Figure 3 indicates equal goodness of fit for all prescribed values of $\kappa$. There is no defined minimum because the higher values of $\kappa$ are compensated with higher calibrated values of $S$, which ultimately produce well-calibrated versions of Hector. Similar behavior of the tradeoff between $S$ and $\kappa$ was observed by Forest et al. (2002) and Sansó and Forest (2009).

Figure 2. The concentration-driven temperature-only fitted values of $S$ (a) and $\kappa$ (b). In both (a) and (b) the shaded gray area marks the reasonable range of $S$ and $\kappa$ we adopted in this paper based on the Knutti et al. (2017) and Stocker et al. (2013) unlikely range for climate sensitivity and our assumption that physically reasonable values of $\kappa$ are $\geq 10^{-6}$. Only 12 of the 31 temperature-only calibrations result in reasonable parameter values.
Figure 3. Results from temperature-only calibration of the three sample models, (CCSM4, GFDL-CM3, and MRI-CGCM3), at fixed values of $\kappa$. Panel (a) compares $\chi$, which represents the calibration results goodness of fit to the ESM temperature output as $\kappa$ is varied. If there were a unique solution, then there would a minimum in the $\chi$ as $\kappa$ is varied. Panel (b) compares the fitted values of $S$ as $\kappa$ is varied. The reason why there is no defined minimum in panel (a) is because $S$ is able to compensate for the changes in ocean heat flux.

5.2. Temperature-Heat Flux Fits

To address the problems of uniqueness illustrated in Figure 3, we explored an additional calibration experiment in which ocean heat flux was added as a constraint. This was inspired by Huntingford et al. (2009), who postulated that ocean heat uptake could be used in instances where global temperature data were an insufficient calibration constraint.

This calibration protocol requires ocean heat flux data for at least one future scenario. However, we were only able to calculate ocean heat flux for 18 of the 31 ESMs to which we were trying to calibrate. For these 18 ESMs, we were able to follow the calibration procedure described in section 4. In lieu of excluding the 13 ESMs without ocean heat flux values from the Hector calibration exercises, we approximated the ocean heat flux constraint based on the range of available CMIP5 ocean heat flux values. Doing this allows us to calibrate Hector to each of the remaining 13 ESMs that lack ocean heat flux data.

To approximate these missing ocean heat flux data, we assume that the missing ocean heat flux values would lie somewhere within the range of the other CMIP5 ocean heat flux output. We then do the following: For any set of Hector parameters (including $\kappa$), Hector will output global mean temperature and global mean heat flux values. The temperature RMSE is computed as previously described. For ocean heat flux time series data, $x$, computed by Hector, we calculate the penalty function

$$L(x) = -\log \left( \frac{E_2 \left( \frac{x_{\text{max}} - x}{\sigma \sqrt{2}} \right) - E_2 \left( \frac{x_{\text{min}} - x}{\sigma \sqrt{2}} \right)}{2(x_{\text{max}} - x_{\text{min}})} \right),$$

(2)

where $x_{\text{min}}$ and $x_{\text{max}}$ are set to the lowest and highest heat flux values observed in the available ESM ocean heat flux data, $\sigma$ is set to a value of 0.15 to force the sides of the penalty function to be steep, and $E_2(x)$ is the Gaussian error function. Figure 4 provides a graph of $L(x)$, illustrating an equally low response for all values within the range $[x_{\text{min}}, x_{\text{max}}]$ and a high response outside of that range. The goodness of fit for these 13 models is then the average of temperature RMSE and $L(x)$. For values of
Figure 5. Change in fitted values of $S$ (a), $\kappa$ (b), $\alpha_a$ (c), and $\alpha_v$ (d) from the temperature-only (gray) to the temperature-heat flux (blue) calibration results for the 31 concentration-driven ESMs. The arrows indicate the direction of how fitted parameter value changed when the heat flux constraint was added. Adding the heat flux comparison data has a notable effect on the fitted values of $S$ and $\kappa$. With the additional constraint the unreasonable fitted values of for $S$ and $\kappa$ shift to a more reasonable range.

Ocean heat flux outside of the range $[x_{\text{min}}, x_{\text{max}}]$, the value of $L(x)$ will be sufficiently high that the goodness of fit is necessarily poor, and the nonlinear optimization algorithm will not choose those parameter values. For values of ocean heat flux within the range, the value of $L(x)$ is sufficiently low that the goodness of fit is dominated by temperature RMSE, with no preference given to any particular heat flux value within the range.

Adding the ocean heat flux constraint caused the fitted parameter values to change while having a negligible impact on Hector’s emulated temperature. The difference in temperature output from Hector between the temperature-only and temperature-heat flux fits, measured as the RMSE, is 0.01 °C. Figure 5 illustrates how adding the heat flux constraint affected the fitted Hector climate parameters. The newly calibrated fitted values of $S$ and $\kappa$ span a range of 2.4° to 7.4°C and 1.0 to 18.3 cm² s⁻¹, respectively. These values of $S$ and $\kappa$ are more aligned with expected ranges based on the physical understanding of the processes involved and values reported in the literature, especially compared to the temperature-only fitted values (Figure 5).
Figure 6. Results from temperature-heat flux calibration of the three sample models (CCSM4, GFDL-CM3, and MRI-CGCM3), at fixed values of $\kappa$. Panel (a) compares $\chi$, which represents the calibration results goodness of fit to the ESM temperature and ocean heat flux results as $\kappa$ is varied. There is a unique solution because there is a global minimum in the $\chi$ as $\kappa$ is varied, indicated by the black asterisks at $\kappa$ is equal to 3.9, 2.8, and 4.45, for CCSM4, GFDL-CM3, and MRI-CGCM3, respectively. It is now possible to identify unique solutions, because now $\kappa$ affects the goodness of fit independently of $S$ because of the effects $\kappa$ has on the ocean heat flux goodness of fit.

By including heat flux constraints, the optimization routine is able to identify unique combinations of Hector parameters (Figure 6) with more realistic ocean heat flux results as visualized in the supporting information.

6. Emissions-Driven Fits

Unlike the concentration-driven scenarios, the CMIP5 emissions-driven (esmHistorical and esmrcp85) runs were driven with CO$_2$, other non-CO$_2$ greenhouse gases, and aerosol emissions. Only a handful of the CMIP5 generation ESMs had the ability to represent climate-carbon feedbacks (Taylor et al., 2012). As such, finding sufficient comparison data to use for calibrating Hector presented a challenge.

The emission-driven calibration required complete esmHistorical and esmrcp85 output for atmospheric CO$_2$ and temperature as well as esmrcp85 heat flux until year 2050. We were able to obtain a total of six complete sets of emissions-driven results that met these requirements. One additional model, MIROC-ESM, had insufficient esmrcp85 ocean heat flux output, so we used the CMIP5 range (Equation 2) for the ocean heat flux constraint. In total, Hector was calibrated to emulate seven different emissions-driven ESMs.

The emissions-driven calibration experiments calibrated a total of seven Hector parameters: four climate and three carbon-cycle parameters. The four climate parameters were the same parameters calibrated during the concentration experiments $S$, ocean heat diffusivity $\kappa$, the aerosol forcing scaling factor $\alpha_a$, and the volcanic forcing scaling factor $\alpha_v$. The three new carbon-cycle parameters are preindustrial CO$_2$ concentration $C_0$ (ppm), the CO$_2$ fertilization factor $\beta$ (unitless), and the heterotrophic respiration temperature sensitivity factor $Q_{10}$ (unitless).

6.1. Temperature-Heat Flux-CO$_2$ Fits

Although Hector is able to emulate emissions-driven ESM results fairly well, Hector overestimates historical temperature and underestimates future temperature for GFDL-ESM2G (Figure 7). Hector calibration fits for CanESM2, CESM1-BGC, and GFDL-ESM2G were selected for main text illustrations of Hector’s ability to emulate temperature, atmospheric CO$_2$, and future ocean heat flux (all Hector fits are visualized in the supporting information). The fitted values of $S$, $\kappa$, $\alpha_a$, and $\alpha_v$ (Table 2) are comparable to the fitted values from the concentration-driven temperature-heat flux calibration (Table 1).
**Figure 7.** Emulated Hector atmospheric CO2 (a–c), ocean heat flux (d–f), and temperature (g–i) for CanESM2, CESM1-BGC, and GFDL-ESM2G (blue) using the temperature-heat flux-CO2 calibration protocol compared with ESM temperature data (gray). No historical ocean heat flux data were used in the calibration process.

**Table 2**

*The Fitted Parameter Values From Two Different Emissions-Driven Experiments*

| Model         | $a_{q}$ | $a_{v}$ | $S$  | $\kappa$ | $C_{0}$ | $\beta$ | $Q_{10}$ | $a_{q}$ | $a_{v}$ | $S$  | $\kappa$ | $C_{0}$ | $\beta$ | $Q_{10}$ |
|---------------|---------|---------|------|----------|---------|---------|----------|---------|---------|------|----------|---------|---------|----------|
| CanESM2       | 1.60    | 0.53    | 3.69 | 1.01     | 281.68  | 0.05    | 1.25     | 1.87    | 1.81    | 3.88 | 0.98     | 282.35  | 0.08    | 1.75     |
| CESM1-BGC     | 0.77    | 0.32    | 3.23 | 3.16     | 280.58  | 0.00    | 2.04     | 0.46    | 2.12    | 2.03 | 12.01    | 289.24  | 0.07    | 1.76     |
| GFDL-ESM2G    | 0.29    | 1.80    | 2.08 | 10.03    | 289.60  | 0.10    | 5.23     | 0.64    | 2.02    | 5.83 | -0.43    | 283.31  | 0.02    | 1.77     |
| MIROC-ESM     | 1.20    | 1.24    | 7.17 | 5.16     | 283.38  | 0.00    | 1.38     | 1.05    | 2.02    | 5.83 | 6.39     | 283.31  | 0.02    | 1.77     |
| MPI-ESM-LR    | 0.90    | 1.15    | 3.32 | 3.48     | 288.28  | 0.20    | 1.07     | 1.22    | 0.70    | 3.66 | 2.93     | 289.13  | 0.28    | 1.75     |
| MRI-ESM1      | 1.50    | 1.33    | 2.37 | 3.16     | 285.70  | 0.51    | 1.00     | 1.42    | 0.27    | 2.04 | 4.21     | 289.49  | 0.66    | 1.76     |

*Note.* The temp-heat flux-CO2 calibration experiment used ESM temperature, future ocean heat flux, and atmospheric CO2 to calibrate Hector, whereas the temp-heat flux-CO2 penalty calibration experiment used the same comparison data but added a penalty to the goodness of fit based on the value of $Q_{10}$. 

Figure 8. Results from temperature-heat flux-CO$_2$ emission-driven calibration experiment of CanESM2, CESM1-BGC, and GFDL-ESM2G at fixed values of $\beta$. Panel (a) compares $\chi$, which represents the calibration results goodness of fit to the ESM temperature, ocean heat flux, and atmospheric CO$_2$ concentration comparison data at fixed values of $\beta$. If there was a unique solution, then there would be a minimum in $\chi$ as $\kappa$ is varied. Panel (b) compares the fitted values of $Q_{10}$ as $\beta$ is varied. The reason why there is no defined minimum in panel (a) is because $Q_{10}$ is able to compensate for changes as $\beta$ is varied.

In the concentration-driven temperature-only calibration protocol, the optimization algorithm was unable to identify unique parameterizations of $S$ and $\kappa$; similarly, here the emissions-driven calibration was unable to identify unique parameterizations of $\beta$ and $Q_{10}$, where $\beta$ controls the vegetation productivity response to CO$_2$ concentration while $Q_{10}$ controls the response of heterotrophic respiration to temperature. Since these parameters affect physical processes that are related to one another, it is hard for the optimization routine to identify a unique combination of these parameters to use in ESM emulation. Over a wide range of fixed $\beta$ values, the optimized goodness-of-fit value is fairly constant (Figure 8), because $\beta$ and $Q_{10}$ can trade off. As $\beta$ increases, the CO$_2$ fertilization effect increases, and Hector’s terrestrial module responds to the increasing CO$_2$ concentrations with an increase in productivity. To offset the atmospheric CO$_2$ that is being used by the terrestrial ecosystems while maintaining the RCP atmospheric CO$_2$ concentration and global mean temperature, $Q_{10}$ must also increase.

6.2. Temperature-Heat Flux-CO$_2$ Penalty Fits

Further constraining these tradeoffs between $Q_{10}$ and $\beta$ would ideally be handled by introducing additional carbon-cycle-related ESM output data such as Net Primary Production (NPP), Net Ecosystem Exchange, terrestrial carbon flux, and carbon flux from heterotrophic respiration, in a similar fashion to our previous description of the concentration-driven calibration procedure (section 5.2). However, because of the limited availability of the emissions-driven scenario results (including unavailability for many ESMs of data for many of these variables), we instead elected to penalize the calibration fit with $Q_{10}$ parameter values that fell outside of empirically derived ranges as defined by the results of Todd-Brown et al. (2018) and Shao et al. (2013). The $Q_{10}$ parameter penalty function was equal to the negative logarithm of a truncated normal distribution with a lower bound of 0, a mean of 1.75, and a standard deviation of 0.4. The $Q_{10}$ parameter penalty function was designed to impact the goodness of fit, which is now the average of the temperature MSE, ocean heat flux MSE, and the $Q_{10}$ penalty. When the fitted value of the $Q_{10}$ parameter is physically unreasonable, the value returned by the $Q_{10}$ parameter penalty function is large, penalizing the goodness of fit.
Figure 9. The change in fitted values of $S$ (a), $\kappa$ (b), $\alpha_a$ (c), $\alpha_v$ (d), $C_0$ (e), $\beta$ (f), and $Q_{10}$ (g) for the two emissions-driven calibration experiments, where one uses a $Q_{10}$ parameter penalty (blue) and one does not (gray). The arrows indicate the direction of change in parameter values when the $Q_{10}$ penalty was incorporated into the calibration protocol.

After applying the $Q_{10}$ parameter penalty, the difference between the emulated temperature, ocean heat flux, and atmospheric CO$_2$ output between the two different emissions-driven calibration experiments was minor. Each RMSE values were 0.05°C, 0.2 W m$^{-2}$, and 47 ppm CO$_2$. These differences are less than 5% of the end of the century output values. Applying the $Q_{10}$ parameter penalty had the most dramatic effect on the fitted values of $Q_{10}$ but also impacted the six other calibrated Hector parameter values (Figure 9). The $Q_{10}$ parameter penalty function narrowed the range where $\beta$ and $Q_{10}$ parameter values could compensate for one another without impacting the goodness-of-fit value. Although the goodness of fit is still somewhat constant between the fixed values of 0.0 to 0.2 for $\beta$, when $\beta$ was larger than 0.1, the goodness of fit steadily gets worse as $\beta$ increases (Figure 10a). Furthermore, it appears that models require small $\beta$ values in order to have reasonable $Q_{10}$ values: Three of the fitted values of $\beta$ are less than 0.1 (Table 2). While $\beta$ values less than 0.1 are small, these values are not impossibly small (Jiang et al., 2020; Jones et al., 2018). This calibration protocol restricted the fitted values of $Q_{10}$ (Figure 10b) from 1.3 to 1.8, which is more consistent with $Q_{10}$ reported by Todd-Brown et al. (2018) and Shao et al. (2013) than the fitted values of $Q_{10}$ from the first emissions-driven calibration experiment.
Figure 10. Results from temperature-heat flux-CO$_2$ emission-driven calibration experiment of CanESM2, CESM1-BGC, and GFDL-ESM2G at fixed values of $\beta$ with the $Q_{10}$ parameter penalty. Panel (a) compares the optimized goodness of fit, and panel (b) compares the fitted values of $Q_{10}$ at the fixed values of $\beta$. There is a range where, as $\beta$ is varied, the quality of the fit changes because $Q_{10}$ is limited in its ability to compensate for the change in carbon fertilization.

7. Discussion and Conclusions

Calibrating a SCM to act as an emulator of an ESM is a nontrivial task. In order to act as a successful emulator, the emulator should be able to reproduce the characteristics of the comparison data utilizing physically reasonable parameter values. To successfully use Hector as a concentration-driven CMIP5 ESM emulator, Hector must be calibrated using a combination of temperature and ocean heat flux comparison data. Similar to the tradeoff between $S$ and $\kappa$ during the temperature-only calibration experiment, we identified two carbon cycle parameters, $\beta$ and $Q_{10}$ that were able to compensate for each other in the emissions-driven calibration experiments when insufficient comparison data were used. For both calibration exercises, we introduced data-based penalty functions (such as Equation 2 and the $Q_{10}$ parameter penalty) to constrain tradeoff parameters $\kappa$ (in concentration driven exercises) and $Q_{10}$ (in emissions driven). The use of penalty functions is broadly applicable to other SCMs wishing to calibrate to individual ESMs with missing output variables. We presented penalty methods that relied on data from both ESM archives and the experimental literature, illustrating the broad usefulness of the technique. In the future, other penalty-based calibration exercises may be performed with Hector, particularly as CMIP6 model results become available. The inclusion of such penalties to calibrate to ESMs with missing variables must be balanced, however, with caution regarding overfitting.

If left unconstrained, SCM parameters can compensate for one another, making it difficult for nonlinear optimization routines to identify unique solutions. This finding has implications for other SCM calibration exercises. As we saw in the concentration-driven calibrations, insufficiently constrained calibration protocols can lead to unreasonable fitted parameter values, limiting the ability to use the SCM as an emulator to test and design new scenarios. Both the concentration-driven and emission-driven protocols depend on heat flux ESM output in order to identify unique fitted values of $S$ and $\kappa$ that use reasonable fitted parameter values. The key findings of this paper have implications for the broader climate modeling and scenario communities by illustrating the information necessary for successful SCM calibration. Although future SCM calibration exercises may be able to take advantage of new scenarios and ESM output from the Phase 6 of the Coupled Model Intercomparison Project (CMIP6), ensuring that sufficient comparison data that are used will continue to be crucial to SCM calibration. Furthermore, using the ESM output range as an additional
calibration constraint will allow calibration exercise to leverage existing ESM output data for variables and scenarios with limited available output.

Conflict of Interest
The authors declare that they have no conflicts of interest.

Data Availability Statement
CMIP5 data were obtained from the Earth System Grid Federation data portal, which is supported by the interagency and international efforts and the German Climate Computing Center Deutsches Klimaschutzzentrum GmbH Long Term Archiving Service. We thank the climate modeling groups for producing and releasing their model output data. The data and code related to the conclusions of this paper are available at https://doi.org/10.5281/zenodo.3515153 (Dorheim & Link, 2019).

Acknowledgments
We thank Claudia Tebaldi for invaluable discussions regarding mathematical assumptions in the calibration procedure. This research was supported by the U.S. Department of Energy, Office of Science, as part of research in Multi-Sector Dynamics, Earth, and Environmental System Modeling Program. The Pacific Northwest National Laboratory is operated for the U.S. Department of Energy by Battelle Memorial Institute under Contract DE-AC05-76RL01830. Support for B. K. was provided in part by the National Science Foundation through Agreement CBET-1931641, the Indiana University Environmental Resilience Institute, and the Prepared for Environmental Change Grand Challenge initiative.

References
Caldeira, K., & Myhrvold, N. P. (2013). Projections of the pace of warming following an abrupt increase in atmospheric carbon dioxide concentration. Environmental Research Letters, 8, 034039. https://doi.org/10.1088/1748-9326/8/3/034039
Clarke, L., Jiang, K., Akimoto, K., Babiker, M., Blanford, G., Fisher-Vanden, K., et al. (2014). Assessing transformation pathways. In O. Edenhofer et al. (Eds.), Climate change 2014: Mitigation of climate change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change (chap. 6). Cambridge University Press.
Dorheim, K. R., & Link, R. (2019). Code and data for two hector calibration papers. https://doi.org/10.5281/zenodo.3515153
Fawcett, A. A., Iyer, G. C., Clarke, L. E., Edmonds, J. A., Hultman, N. E., McJeon, H. C., et al. (2015). Can Paris pledges avert severe climate change? Science, 350(6265), 1168–1169.
Forest, C. E., Stone, P. H., Sokolov, A. P., Allen, M. R., & Webster, M. D. (2002). Quantifying uncertainties in climate system properties with the use of recent climate observations. Science, 295, 113–117.
Friedlingstein, P., Meinshausen, M., Arora, V. K., Jones, C. D., Anav, A., Liddicoat, S. K., & Knutti, R. (2014). Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks. Journal of Climate, 27(2), 511–526. https://doi.org/10.1175/JCLI-D-12-00579.1
Geoffroy, O., Saint-Martin, D., Olivié, D. J. L., Voldoire, A., Bellon, G., & Tytéca, S. (2013). Transient climate response in a two-layer energy-balance model. Part I: Analytical solution and parameter calibration using CMIP5 AOGCM experiments. Journal of Climate, 26(6), 1841–1857.
Hartin, C. A., Patel, P., Schwarzer, A., Link, R. P., & Bond-Lamberty, B. P. (2015). A simple object-oriented and open-source model for scientific and policy analyses of the global climate system—Hector v1.0. Geoscientific Model Development, 8(4), 939–955. https://doi.org/10.5194/gmd-8-939-2015
Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., et al. (2017). The art and science of climate model tuning. Bulletin of the American Meteorological Society, 98(3), 589–602.
Huntingford, C., Lowe, J. A., Booth, B. B. B., Jones, C. D., Harris, G. R., Gohar, L. K., & Meir, P. (2009). Contributions of carbon cycle uncertainty to future climate projection spread. Tellus B: Chemical and Physical Meteorology, 61(2), 355–360.
Jiang, M., Medlyn, B. E., Drake, J. E., Duursma, R. A., Anderson, I. C., Barton, V. M., et al. (2020). The fate of carbon in a mature forest under carbon dioxide enrichment. Nature, 580(7802), 227–231. https://doi.org/10.1038/s41586-020-2128-9
Jones, A. D., Calvin, K. V., Shi, X., Di Vittorio, A. V., Bond-Lamberty, B., Thornton, P. E., & Collins, W. D. (2018). Quantifying human-mediated carbon cycle feedbacks. Geophysical Research Letters, 45, 11,370–11,379. https://doi.org/10.1029/2018GL079350
Khabbazan, M. M., & Held, H. (2019). On the future role of the most parsimonious climate module in integrated assessment. Earth System Dynamics, 10, 135–155.
Knutti, R., Rugenstein, M. A. A., & Hegerl, G. C. (2017). Beyond equilibrium climate sensitivity. Nature Geoscience, 10, 727–736. https://doi.org/10.1038/ngeo3017
Kriegler, E. (2005). Imprecise probability analysis for integrated assessment of climate change (Ph.D. Thesis). Potsdam: Universität Potsdam. https://www.pik-potsdam.de/members/edenh/theses/PhDKriegler
Meinshausen, M., Meinshausen, N., Hare, W., Raper, S. C. B., Frieler, K., Knutti, R., et al. (2009). Greenhouse-gas emission targets for limiting global warming to 2°C. Nature, 458(7242), 1158.
Sansó, B., & Forest, C. (2002). Quantifying uncertainties in climate system properties. Journal of the Royal Statistical Society: Series C (Applied Statistics), 51, 485–503.
Schultz, P. A., & Kasting, J. F. (1997). Optimal reductions in CO₂ emissions. Energy Policy, 25(5), 491–500.
Shao, P., Zeng, X., Moore, D. J. P., & Zeng, X. (2013). Soil microbial respiration from observations and Earth system models. Environmental Research Letters, 8(3), 034034. https://doi.org/10.1088/1748-9326/8/3/034034
Smith, S. J., & Edmonds, J. A. (2006). The economic implications of carbon cycle uncertainty. Tellus B: Chemical and Physical Meteorology, 58(5), 586–590. https://doi.org/10.1111/j.1600-0889.2006.00217.x
Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., et al. (Eds.) (2013). Climate change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge, UK and New York, NY: Cambridge University Press.
Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2012). An overview of CMIP5 and the experiment design. Bulletin of the American Meteorological Society, 93(4), 485–498. https://doi.org/10.1175/bams-d-11-00094.1
Todd-Brown, K., Zheng, B., & Crowther, T. W. (2018). Field-warmed soil carbon changes imply high 21st-century modeling uncertainty. Biogeosciences, 15(12), 3659–3671. https://doi.org/10.5194/bg-15-3659-2018

DORHEIM ET AL. 13 of 14
van Vuuren, D. P., Lowe, J., Stehfest, E., Gohar, L., Hof, A. F., Hope, C., et al. (2011). How well do integrated assessment models simulate climate change? *Climatic Change, 104*(2), 255–285. https://doi.org/10.1007/s10584-009-9764-2

Vega-Westhoff, B., Sriver, R. L., Hartin, C. A., Wong, T. E., & Keller, K. (2019). Impacts of observational constraints related to sea level on estimates of climate sensitivity. *Earth’s Future, 7*, 677–690. https://doi.org/10.1029/2018EF001082