Authors' final response

Aude Champouillon, Gerhard Krinner & Juliette Blanchet

September 5, 2026

We thank the two referees who contributed to the discussion. Their thoughtful comments will help us propose a revised version of the paper. In the following, we reply to all their points, which are shown here in italics. We will sometimes use the acronym "ERBC" for "empirical run-time bias correction" in the following.

1 Reply to RC1 (and RC2) (John Scinocca's comments)

We reply here to the first reviewer's initial comments (RC1), which have been completed by a another contribution (RC2) following our initial reply in the open discussion.

1.1 State-dependent ERBC and scope of manuscript

The inclusion of the state-dependent ERBC method (Section 2.2.3 and Appendix A2) raises a fundamental concern regarding the scope of this manuscript. Unlike the other approaches considered, this is a new bias-correction framework in which the corrective term is conditioned on diagnosed weather regimes. However, it receives only a one-paragraph description in the manuscript, and I found the details in Appendix A2 insufficient to fully understand how the methodology operates and how it relates to the other methods discussed in this study. Ultimately, its motivation, theoretical basis, algorithmic formulation, relationship to previous ERBC approaches, expected properties, and sensitivity to key design choices require a much more complete presentation. The authors themselves acknowledge in the Discussion section, "The lack of effect of making correction state dependent with this method may be the sign that our methodology is not adequate and must be adjusted." I do not believe this issue can be addressed through a conventional major revision, as doing so would fundamentally change the scope and balance of the manuscript. The state-dependent ERBC method warrants its own dedicated methodological paper before being incorporated into a comparative study. Consequently, I recommend that the authors remove the state-dependent ERBC from the present manuscript and instead develop it as a standalone study. As part of that study, a section could be reserved to evaluate the new methodology using the same diagnostics and evaluation framework employed here, thereby placing it into the context of the established ERBC approaches while allowing the methodology itself to be properly developed and peer reviewed.

Reply We thank the reviewer for raising this concern, which is also raised by the second reviewer. We also acknowledge that the introduction of this new method might have been too rushed in the first version of the paper.

The goal of this study is to compare multiple run-time bias correction methods, and the state-dependent one only makes the comparison more complete. Therefore, our point of view is that including this method does very well fit the scope of the paper.

In addition, this method has not been published anywhere before, and publishing it in a dedicated paper would only lead to further inflation of scientific publishing without much added value, as a separate paper on this variant would necessarily have to include a comparison at least with the standard bias correction method. We do think that the relatively short treatment of this variant in the present paper is the better option, because it allows a large community to be informed about the existence of this method and its relatively weak added value (possibly preventing others from losing their time).

We hence suggest to keep the state-dependent ERBC in the analysis. Nevertheless, to improve the description of the method and the readability of the paper, we would include a more detailed description of the newly developed state-dependent bias correction method as supplementary material. We agree that we also should make clearer how it differs from the classical cyclostationnary ERBC approaches.

In addition, we would make the novelty of this method clearer in the abstract and in the introduction, hence extending the scope of the manuscript, by stating that its secondary objective is to introduce this new variant. The state-dependent ERBC method is merely an incremental development from the classical nudging-based ERBC method; it is therefore not an entirely new approach. Furthermore, we would briefly comment on the choice of parameters and mention the expected properties as we introduce the method.

1.2 Common conceptual framework

With the removal of the state-dependent ERBC methodology, the manuscript becomes conceptually much simpler. All remaining approaches belong to a common class of run-time bias correction in which a cyclostationary, state-independent forcing is added to the right-hand side of the evolution equations of the corrected variables (u and v in this study). The methods differ only in how this corrective forcing (ERBC) is derived.

The manuscript would benefit from first establishing the common properties of this class of bias correction before introducing the individual ERBC derivation methodologies, perhaps by expanding Section 2.2.1, the Introduction, or both. In particular, the discussion should emphasize the following properties of this class of bias corrections:

1. *The method is applied to a finalized, tuned control model with known climatological biases.*

2. *It does not change the physical formulation of the control model because the correction is a cyclostationary state-independent term added to the RHS of each corrected variable's evolution equation.*

3. *It is designed to explicitly/directly bias correct the climatological annual cycle of the variables it is applied to - "by construction"*

4. *Because it is applied during runtime of the model, the improved climatological annual-cycle basic state of the corrected variables can implicitly/indirectly influence the behaviour of other quantities that are not targeted by the bias correction, which occurs as "a response".*

Establishing these properties at the outset would provide a clearer framework for interpreting many of the later results. For example, from (1) it follows that the control represents a finalized and tuned model, which has climatological biases we would like to correct. It is therefore reasonable to expect that compensating errors exist within the tuned parameterizations to counteract those biases. Combining this with (2) and (3), one can anticipate that directly bias correcting a prognostic variable with a large climatological bias may disrupt those compensating errors as its basic state is altered and so made inconsistent with the tuning of the model physics. Consequently, some aspects of the simulation may degrade as a consequence of detuned physical parameterizations, despite the targeted variable itself improving by construction. This provides a general rationale for why directly bias correcting temperature caused problems in the LMDZ model and explains why other models (and even earlier versions of the LMDZ model) could successfully apply ERBC to T. This framework also clarifies how the different ERBC methodologies should be evaluated. From (3), the primary comparison should focus on the corrected variables (u and v), since these directly reflect the effectiveness of the different ERBC derivation methods. In contrast, from (4), responses in variables such as temperature and precipitation are indirect consequences of the correction and are therefore more likely to depend on the characteristics of the particular host model rather than on the ERBC methodology itself. Introducing this conceptual framework before presenting the individual methodologies would substantially improve the organization of the manuscript and help distinguish conclusions that are broadly applicable to the ERBC methodologies from those that are specific to the LMDZ implementation.

Reply We thank the reviewer for this detailed comment that points out with precision what lacks in our description of the ERBC framework. We therefore plan to propose a revised introduction and method description section by including these relevant elements.

It is true, that the state-dependent approach differs from other methods. The discrepancy must indeed be described in an explicit way. We believe that its introduction remains feasible as it is a "natural" extension of the original state-independent methods. We will make sure to insist on how it differs from the first class of methods.

1.3 Rationale for out-of-sample validation

The rationale for using out-of-sample validation as the primary basis for comparing the ERBC formulations requires further consideration. The objective of this study is to compare alternative methodologies for deriving the ERBCs, that is, to determine which methodology most effectively reduces the climatological biases of the corrected prognostic variables. This is fundamentally a boundary-value problem in which climatological statistics define both the correction and its evaluation. It therefore differs from an initial-value prediction problem, where split-sample validation is the standard approach because information from the period being predicted cannot be used during model initialization or calibration. It is therefore not obvious that the same validation philosophy is appropriate here.

Furthermore, interpretation of the out-of-sample validation is complicated by anthropogenic climate change, which the authors have identified. Separating the training and validation periods changes the target climatology itself. Consequently, if one ERBC formulation performs better than another during the validation period, it is unclear whether this reflects greater methodological robustness or simply differences *between the climatologies of the two periods. Without additional analyses, these competing explanations cannot be distinguished.*

This issue is also difficult to reconcile with the manuscript's conclusion that temporal smoothing of the 6-hourly ERBC provides little benefit. If a 20-year training period is already sufficient to estimate a stable ERBC without smoothing, then the principal motivation for withholding an additional 20 years to guard against sampling variability appears substantially weakened. Conversely, if sampling variability is considered sufficiently important to justify split-sample validation, then the conclusion that an unsmoothed 20-year ERBC is adequately converged requires stronger justification.

If the authors retain the out-of-sample validation, then, at a minimum, Fig. 2 should also include results over the training period together with an assessment of the changing climatology. The in-sample validation would establish the relative performance of the ERBC formulations under the conditions for which the corrections were derived. The influence of climate change could then be quantified by comparing the reference climatological annual cycles of the training and validation periods. If RT(d) and RV(d) denote the corresponding reference annual cycles, then B(d) = RV(d) - RT(d) measures the change in the target climatology itself. The magnitude of B(d) could be summarized using the same global annual RMSE metrics employed in Fig. 2. The figure would then contain three complementary panels: (i) training-period validation, (ii) the existing validation-period results, and (iii) the RMSE associated with the change in the reference climatology. This would establish whether the differences among the ERBC formulations are large relative to the change in the target climatology itself.

Depending on the outcome of these analyses, the authors may wish to reconsider whether out-of-sample validation should remain the primary basis for comparing the ERBC formulations. If the changing climatology contributes substantially to the global RMSE, then the out-of-sample comparison may be unnecessarily confounded. In that case, validation over the training period would provide the clearest assessment of the competing methodologies, with the out-of-sample results serving as a complementary robustness analysis. Moreover, if the changing climatology proves important for the globally averaged RMSE, its influence may be impacting the two-dimensional latitude-longitude bias fields presented throughout the manuscript.

Reply We must admit that we have trouble understanding the reviewer's point, on why the out-of-sample validation might not be relevant here. In our view, out-of-sample validation here is not conceptually different from validating a climate model on a period for which it was not directly designed, such as validating climate models, tuned on present climate, on paleo-climatological periods.

Evaluating ERBC methodologies only on the training period, one would never know if their performances could be transposed to another period. As the main motivation for our use of ERBC is to eventually use it in climate change simulations, it makes sense to evaluate them on a different period than the training period.

It is true, however, that evaluating ERBC only on an evaluation period, prevents from identifying if the remaining biases are caused by a defect in the ERBC method, by over-correction or by the fact that biases are not stationnary. Over-correction refers to the fact that ERBC terms could be too dependent on the period used to derive them.

The potential non-stationnarity of biases appart, there are indeed 2 sources of errors, that cannot be

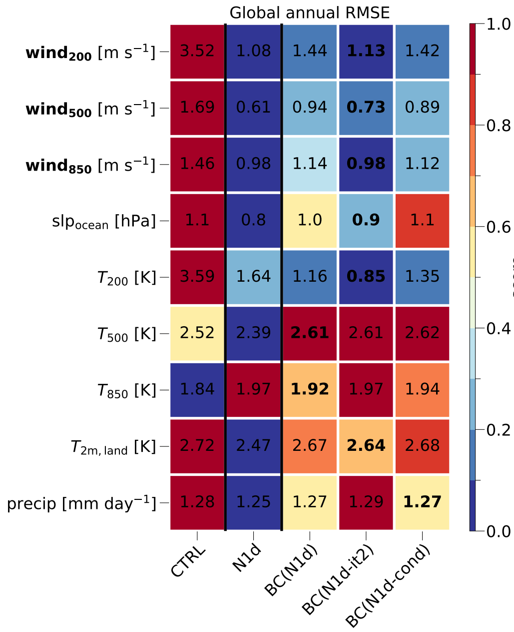

Figure 1: **RMSE of global mean variables and corresponding scores, evaluated on the training period (1981-2000).** Numbers correspond to the RMSE while color shading corresponds to the associated score. Bold font is used for the lowest RMSE for a given variable among BC simulations. For simulation types for which multiple runs were available, the minimum and maximum RMSE are indicated below and above the mean RMSE. The score's and variables' definitions are in the main text.

differentiated. But ultimately, over-correction is also a deficiency of the ERBC method. Being able to distinguish between these sources of errors is interesting in itself and could give hints on how to improve them. It is however not strictly necessary in the context of evaluating and comparing these methods, for a use in simulations of various periods, include future periods. Note that in Krinner et al. (2025), over-correction is carefully examined in the case of the iterated nudging-based method, as it constrains the number of iterations.

To address the reviewer's concerns, about whether in-sample validation should be preferred over out-of-sample validation, we propose to include, as supplementary material, a figure similar to Fig. 2 (in the paper) but for the training period (only for some of the simulations).

Figure 1 is a draft of the proposed additional figure. The conclusions we draw from this additional analysis are that:

1. differences in biases remain modest ;
2. the ranking between the three methodologies retained here is conserved for most variables ;
3. the main conclusions of the present study are not altered.

While including this additional analysis and figure in the paper, we choose to keep the out-of-sample validation as primary basis for comparing the ERBC formulations in this study.

"This issue is also difficult to reconcile with the manuscript's conclusion that temporal smoothing of the 6-hourly ERBC provides little benefit." We believe that the reviewer refers to the statement in 4.5 "Scinocca and Kharin (2024) also study the influence of spatial smoothing of the correction terms. As preliminary tests showed no significant effect, we decided to not investigate this further.". There, we refer to *spatial* smoothing of the correction terms. Correction terms derived from nudged simulations are indeed temporally smoothed, as is described in A1.

1.4 CABCOR properties

The goal of this paper is to compare different methodologies for deriving the cyclostationary correction term (ERBC) appearing on the right-hand side of the evolution equations for u and v, with the hope of drawing conclusions that extend beyond the specific LMDZ model used for this study. It is therefore essential that the leading-order control parameters of each methodology are adequately explored. This has been done for the nudging-based methodologies, where the authors investigate multiple values of the relaxation timescale together with the iterative formulation. However, this has not been done for the CABCOR methodology.

The manuscript implicitly treats the relaxation timescale, denoted by the single symbol tau, as though it represents the same control parameter in both the nudging and CABCOR methodologies. It does not. These two parameters have fundamentally different physical meanings. Following Scinocca and Kharin (2024, hereafter SK24), I will therefore refer to them as τ_R *(nudging) and* τ_C *(CABCOR).*

In the nudging methodology, τ_R *is the timescale with which the model winds are relaxed toward the single realization of ERA5 meteorology. In contrast,* τ_C *is the timescale with which the climatological annual cycle of the model is relaxed toward that of ERA5. Unlike nudging, the individual meteorological realizations in the CABCOR adaptation simulation are entirely different from those in the reference. Consequently, there is no physical basis for assuming that* $\tau_C = \tau_R$*. Nevertheless, by considering only a single value of tau, the manuscript implicitly makes this assumption.*

This assumption is evident in Fig. 1, where the CABCOR correction terms are substantially weaker than those obtained using the strongest nudging configuration, BC(N1d-it2). A meaningful comparison should allow each method to span a comparable range of correction strengths by varying its principal control parameter. Just as the nudging method is examined using different values of τ_R*, the CABCOR method should likewise be examined using different values of* τ_C*. Based on Figure 1, this likely requires substantially smaller values, such as 12 hours and 6 hours for* τ_C*, so that the correction tendencies can reach magnitudes comparable to BC(N1d-it2).*

Also, the BC(CA1d-y30), BC(CA1d-y70), and BC(CA1d-y110) experiments do not span the behaviour of the CABCOR method. They instead represent different stages in the convergence of a single CABCOR configuration with $\tau_C = 1$ *day. The analogy for the nudging method would be to compare ERBCs derived from 2-, 5-, 10-, and 20-year nudged simulations using the same value of* τ_R*. Such experiments do not span the method. They simply demonstrate convergence for a single parameter choice.*

Similarly, CABCOR requires sufficient sampling to converge. In CABCOR, the annually updated correction term is initially derived from only a poor estimate of the bias-corrected model climatology. This can produce unstable behaviour, particularly for smaller values of τ_C*, and is the reason why a gradual spin-up, in which* τ_C *is decreased toward its target value, is required during the early part of the adaptation. Beyond the spin-up, the relaxational nature of the formulation suppresses unstable departures. At the same* *time, as an increasing number of years contributes to the estimate of the bias-corrected climatology, any remaining instability is shifted to progressively longer timescales. This allows the CABCOR adaptation to keep such departures under control relative to the corresponding BC simulation using ERBCs extracted before convergence. Eventually, the metrics of the corresponding BC simulation converge toward those of the CABCOR adaptation, indicating convergence of the derived ERBC.*

An important result of SK24 is that the CABCOR adaptation appears to approach its final metric values well before the corresponding BC simulations converge (see Fig 1 of SK24 – 3d, 2d, 1d). However, the very long convergence times reported in SK24 likely reflect the much more demanding implementation used there, in which corrections were applied throughout the atmospheric column and evaluated from the surface to 10 hPa. In the present study, ERBC is applied only to u and v, only between 850 and 200 hPa, and the evaluation is likewise limited to 200 hPa. The similarity of the BC(CA1d-y30), BC(CA1d-y70), and BC(CA1d-y110) experiments therefore suggests that convergence may be occurring much more rapidly than in SK24. However, the manuscript does not demonstrate this. Instead, it assumes that the convergence characteristics reported by SK24 apply here, which isn't justified without verification.

I therefore suggest the following revisions:

- *Remove material: Remove all but one of the BC(CA1d-y30), BC(CA1d-y70), and BC(CA1d-y110) experiments, since they appear to represent essentially the same converged result for τ_C = 1 day.*
- *Add material: Investigate several values of τ_C (e.g., 1 day, 12 hours, and 6 hours) so that the CABCOR method spans a range of correction strengths comparable to that of the nudging-based approaches.*
- *Convergence of the BC simulation and the CABCOR adaptation: The authors should investigate the convergence of the CABCOR adaptation as it appears to be much faster than in SK24 for the reasons mentioned above. Explicit convergence of the BC simulation to the CABCOR adaptation may not be necessary. As discussed above, the stability of the CABCOR adaptation simulation allows it to approach the final metric values more rapidly than the associated BC simulation. Once the adaptation metrics have stabilized, they should be representative of the eventual converged BC results. However, this should be demonstrated for at least one value of τ_C in the present implementation.*

Reply As the reviewer points out, the temporal scales of the nudging-based approach and the CABCOR methodologist (referred to as τ, without distinction, in the original draft) have different physical meaning. We will therefore adopt different notations for them, τ_R and τ_C for example, following the reviewer's suggestion.

We understand that the reviewer considers that exploring the sensitivity of CABCOR to the value of τ_C makes more sense than exploring the sensitivity of CABCOR to the length of the adaptation simulation. He therefore suggests that we compare simulations with different values of τ_C and fixed length of adaptation simulation, rather than simulations with different lengths of adaptation simulation and fixed values of τ_C as was originally proposed.

The point is well taken and we would therefore like to replace:

- BC(CA1d-y30)

- BC(CA1d-y70)
- BC(CA1d-y110)

by

- BC(CA1d-y150)
- BC(CA12h-y150)
- BC(CA6h-y150)

The proposed fixed length of adaptation simulation here is 150 years, however we will make sure to document convergence of the method with this length.

1.5 Year selected for CABCOR repeated annual cycle simulations

The choice of 1991 as the repeated forcing year for the 100-year CABCOR adaptation simulation raises significant concerns. The CABCOR method relies on a repeated annual cycle simulation to define the climatological correction term. The choice of repeated year therefore determines the forcing environment to which the model is repeatedly exposed throughout the adaptation. The year 1991 is problematic because it includes the eruption of Mt. Pinatubo, one of the largest volcanic aerosol perturbations of the twentieth century. The reference describing the LMDZ6A model states that stratospheric aerosols are prescribed from the CMIP6 forcing data set as monthly and annually varying latitude-height fields of aerosol optical properties that interact radiatively with the model atmosphere. Consequently, the model responds prognostically to the volcanic aerosol forcing associated with the Mt. Pinatubo eruption. Repeating the 1991 forcings for 100 years therefore repeatedly imposes the Pinatubo volcanic forcing throughout the adaptation. The resulting circulation response has been extensively documented. Enhanced stratospheric aerosol loading warms the tropical lower stratosphere, strengthens the equator-to-pole temperature gradient, alters planetary-wave propagation, strengthens the polar vortex, and produces significant anomalies in the Northern Hemisphere zonal circulation and Arctic Oscillation (Stenchikov et al., 2002). Multi-model studies have likewise demonstrated robust responses of the large-scale circulation and zonal wind fields to the Pinatubo eruption across CMIP-class climate models (Barnes et al., 2016). These circulation responses involve precisely the variables being bias corrected in the present study. Consequently, the derived CABCOR ERBC is compensating in part for a repeatedly imposed volcanic circulation anomaly rather than solely for the model's climatological wind bias. This could influence both the magnitude and spatial structure of the derived ERBC and thereby compromise comparisons with ERBCs derived using the other methodologies, none of which rely on repeated annual-cycle simulations. To avoid these issues, the authors should repeat the CABCOR adaptation using a representative non-volcanic forcing year. Alternatively, the repeated annual cycle could be constructed using a climatological volcanic aerosol forcing rather than that associated with the Mt. Pinatubo eruption. Without such a change to the experimental design, the current CABCOR simulations do not provide a reliable basis for evaluating the methodology.

Reply The reviewer's initial comment and response in RC2 have allowed us to realize that we made mistakes and inconsistent choices when implementing CABCOR. In the CABCOR adaptation simulation used for the CABCOR simulations analyzed in the first draft of this paper, the setup included the following elements:

- The reference climatology was computed from the period 1981-2010 (and not 1981-2000 as could be suggested by the reference training period of other ERBC methods).
- The year 1990 (and not 1991 as stated in the draft) was used for repeated solar, GHG and aerosols forcings, but also for sea surface boundary conditions.

We will hence re-run all CABCOR adaptation simulations with the following setup:

- The reference climatology will be computed over the period 1981-2000, making it consistent with the rest of the paper.
- The year 1990 will be used for repeated solar, GHG and aerosols forcing, as it it the central year of the climatological period, and free from major volcanic eruption.
- Climatological (computed over 1981-2000) sea surface boundary conditions will be applied.

2 Reply to RC3 (Anonymous referee's comments)

2.1 Description of the experimental design

This issue was already pointed out in detail by RC1 and I agree with those suggestions. I also think that the state-dependent BC could be removed from the comparison in this manuscript for the clarity of the scope. For the remaining BC methods, I would suggest adding more context in Section 2 to make the purpose clearer to readers. For example, in the comparison of BC(N1d), N2d and N3d in Section 3, the manuscript emphasizes that the spatial distribution of the increments are not identical and their amplitude is not proportional to the inverse of the relaxation factors. I think that this behavior is somewhat expected, considering the nonlinear propagation of the nudging effect through model integration. Therefore it would be helpful to describe this context in the method section to indicate what readers should focus on when comparing the results. Also I think more explanation is needed regarding the choice of 30, 70 and 110 years for CA1d-based CABCOR ERBC, including how the cases were expected to differ from one another.

Reply We thank the reviewer for pointing out the lack of clarity of our context's description. We will take into account the suggestions of added explanations. More specifically, regarding the effect of varying the value of τ in nudging simulations to derive ERBC terms, we will make clearer what results are expected by design, and what is more of a side-effect (positive or negative). Additionally, we will include more details on the common conceptual framework.

Regarding the state-dependent ERBC method, we argue that it is relevant to the present study. The goal of this study is indeed to compare multiple run-time bias correction methods, and the state-dependent one only makes the comparison more complete. Therefore, our point of view is that including this method does very well fit the scope of the paper. In addition, this method has not been published anywhere before, and publishing it in a dedicated paper would only lead to further inflation of scientific publishing without much added value, as a separate paper on this variant would necessarily have to include a comparison at least with the standard bias correction method. We do think that the relatively short treatment of this variant in the present paper is the better option, because it allows a large community to be informed about the existence of this method and its relatively weak added value (possibly preventing others from losing their time).

Nevertheless, to improve the description of the method and the readability of the paper, we will include a more detailed description of the newly developed state-dependent bias correction method as supplementary material. We will also make clearer how it differs from the cyclostationnary classical ERBC approaches. In addition, we will make the novelty of this method clearer in the abstract and in the introduction, hence extending the scope of the manuscript, by stating that its secondary objective is to introduce this new variant.

Regarding CABCOR simulations, following suggestions from the first reviewer, we will revise our analysis framework and therefore propose to replace

- BC(CA1d-y30)
- BC(CA1d-y70)
- BC(CA1d-y110)

by

- BC(CA1d-y150)
- BC(CA12h-y150)
- BC(CA6h-y150)

The proposed fixed length of adaptation simulation here is 150 years, however we will make sure to document convergence of the method with this length.

2.2 The terms bias and RMSE

The term "bias" is often used as the mean error. I found it somewhat confusing when the terms "bias" and "RMSE" appeared to be used interchangeably, for example, in the caption of Figure 3 and Section 3.2.2. Though, if this usage is already standard in the climate model community, it may be acceptable to keep it as it is.

Reply We thank the reviewer for pointing out this amalgam made between bias and RMSE. We will make sure that the revised paper is perfectly clear on this aspect.

2.3 Regional view

3.4 L288 The paragraph beginning with "One might want to select the method that gives the smallest residual biases in their region of interest.", may sound as though it casts doubt on the selection of the BC method based on the performance in a specific region. Indeed, if the simulation is performed with different methods on each region of the earth which shows the smallest RMSE would lead to a rather worse result because of the inconsistency between regions through the global long-term simulation. It might be helpful to add some discussion on this.

Reply This sentence was possibly misleading, and we apologize for the misunderstanding. It was not our intention to suggest applying different methods in different regions within a same simulation. We wanted to adopt the perspective of a regional climate model user, aiming at selecting the best global forcing model. In this perspective, one would focus mainly on regional metrics, maybe disregarding bad performances on other parts of the globe. To avoid any misunderstanding, we suggest to replace the initial sentence by "Taking a regional climate modeler perspective, one might want to select the method that gives the smallest residual biases in their region of interest, disregarding performances in other parts of the globe."

2.4 possible typos

- *Table 1: "defautl"*
- *L406 "correction terms à "*
- *L412 "initial tendency erros "*

Reply Thank you for pointing out these errors, they will be corrected.