GFDL FMS Coupler Cost

GFDL Modeling Systems Division
Technical Memorandum GFDL202002
2 July 2020

Document Prepared by Rusty Benson and Niki Zadeh

Abstract

A review of the results for GFDL models from the Computational Performance Model Intercomparison Project (CPMIP) raises the question of why the coupling costs appear so extreme, especially in light of tests demonstrating minimal cost for the FMS coupling framework. This technical memorandum seeks to explain the seemingly high cost associated with the FMS coupler in pre-industrial control (piControl) simulations using the GFDL ESM4 Earth system model and two configurations of the CM4 climate model[1]. While this technical note presents results for three specific piControl simulations, the findings are assumed to be general and applicable to the full range of configurations and experiments. Among the findings are:

- Load imbalances in coupled models are not limited to inequalities in the time spent within each component, but also when not every timestep in the components takes the same wall-clock time, due to internal timesteps longer than the coupling timestep.
- Making each timestep take the same amount of computation time, even if the individual timesteps become more expensive, can significantly alleviate high-frequency load imbalance, and decrease the expense of the coupled model.
- Diagnostics are not “free” and every attempt should be made to ensure the logic is performant and justify their output from an experiment.

Background

In GFDL coupled models, the Earth system is separated at the highest level into atmospheric (ATM) and ocean (OCN) components. To minimize time to solution and best exploit the resources available, the ATM and OCN components execute at the same time on different groups of MPI-ranks, also referred to as PE-Lists. During execution, the two components perform distinctly different operations and also need to frequently exchange information. When configuring a long-duration simulation, every attempt is made to ensure the individual cumulative runtimes for the ATM and OCN are as close to identical as possible. Having the times match is imperative to reduce load-imbalance, where the resources associated with a component are idled while waiting for the other to arrive at a data exchange point.

Coupling cost is expressed as a fractional value defined using resource-weighted time values (T_xP_x) where T is the time and P is the hardware resources used.

$$CC = \frac{T_{ML}P_{ML} - \sum_C T_C P_C}{T_{ML}P_{ML}}$$

The coupling cost equation is designed to take into account the load imbalance amongst other coupling overheads, such as data exchange and determination of fractional quantities associated with overlapping grids, while ignoring the initialization and termination phases[2].

[1] A description of ESM4 and CM4 can be found on the GFDL website and associated reference list.

[2] Computational Earth system models have three distinct phases of execution: initialization; time integration or “Main Loop”; and termination.

Methodology

From the volume of data created during the most recent CMIP cycle, timing information was sampled to establish a coupling cost of 0.28 for a one-year run of the piControl ESM4 piControl (see table A). Prior to initiating further investigations as to probable causes for the steep coupling cost, a decision was made to update to newer versions of certain components along with reducing the simulation to a more manageable one month duration. The baseline resulting from these changes mimicked the production sample with a similar coupling cost and parity in timings for ATM and OCN.

TABLE A

Time Integration Step	Component Timing
Main Loop	9802s
DO NUM_CPL	
ATM	7073s
ATM: atmos loop	5992s
LND: update_land_model_slow	77s
ICE: update_ice_model_slow slow	1017s
OCN	7159s

Analysis of all clocks embedded within the Main Loop did not provide any insight into the discrepancy between component timing and that associated with the Main Loop and showed minimal time (O(10s)) spent in these coupling-related functions. A review of the source code within the Main Loop identified a series of untimed, PE-List[3], context switches (*mpp_set_current_pelist*). To determine if these untimed synchronization points could be hiding load imbalance, an existing optional argument to bypass the barrier was utilized. Removal of these implicit barriers resulted in increased maximum time and significant load imbalance for two coupling related functions that had previously not accounted for much time. These coupling functions were rejected as the source of the imbalance as they previously did not account for significant time when entered synchronously. Explicit, timed barriers were added as the first executable statement inside of the Main Loop and again at the conclusion of the concurrent stage of the ATM and OCN components. The coupling functions all reverted to having minimal impact and the time spent waiting at the post-concurrent stage barrier accounted for the missing time.

TABLE B

Time Integration Step	Component Timing
Main Loop	937s
DO NUM_CPL	
Top Barrier	3s
ATM	697s
OCN	682s
Post-concurrent Barrier	252s

[3] PE-List, or Processing Element list, is a grouped set of MPI-ranks for a specific task.

Although the missing time was found, it didn't point to a systemic load-imbalance as every MPI-rank spent almost the exact same time waiting at the barrier as evidenced by the almost identical maximum and minimum times 233s vs. 252s for post-concurrent barrier in table B.

Finer granularity was needed and logic was added to time each core from the beginning of the coupling step to the end of the concurrent phase for every MPI-rank, using MPI_Wtime, and output the time series for analysis. The following sections give a detailed analysis of the time series for the GFDL ESM4 and CM4 models.

ESM4 piControl

FIgure 1 is a time-series plot of the per-coupler step times for specific MPI-ranks in the ATM and OCN component, respectively. The ocean data clearly exhibits an alternating pattern, with a large delta between the maxima and minima. This corresponds to the thermodynamic and tracer advection step being computed on a timescale of one-half the coupling frequency (DT_THERM=7200s). Further, while the atmospheric processes have much less variation in the maxima and minima, one can still clearly see the pattern of the radiation timescale being computed at one-third the coupling frequency. Thus the data clearly demonstrates a load-balance oscillation of approximately 0.675s with OCN taking longer than ATM on even coupling steps while reversing the roles of ATM and OCN at about 0.625s for odd coupling steps.

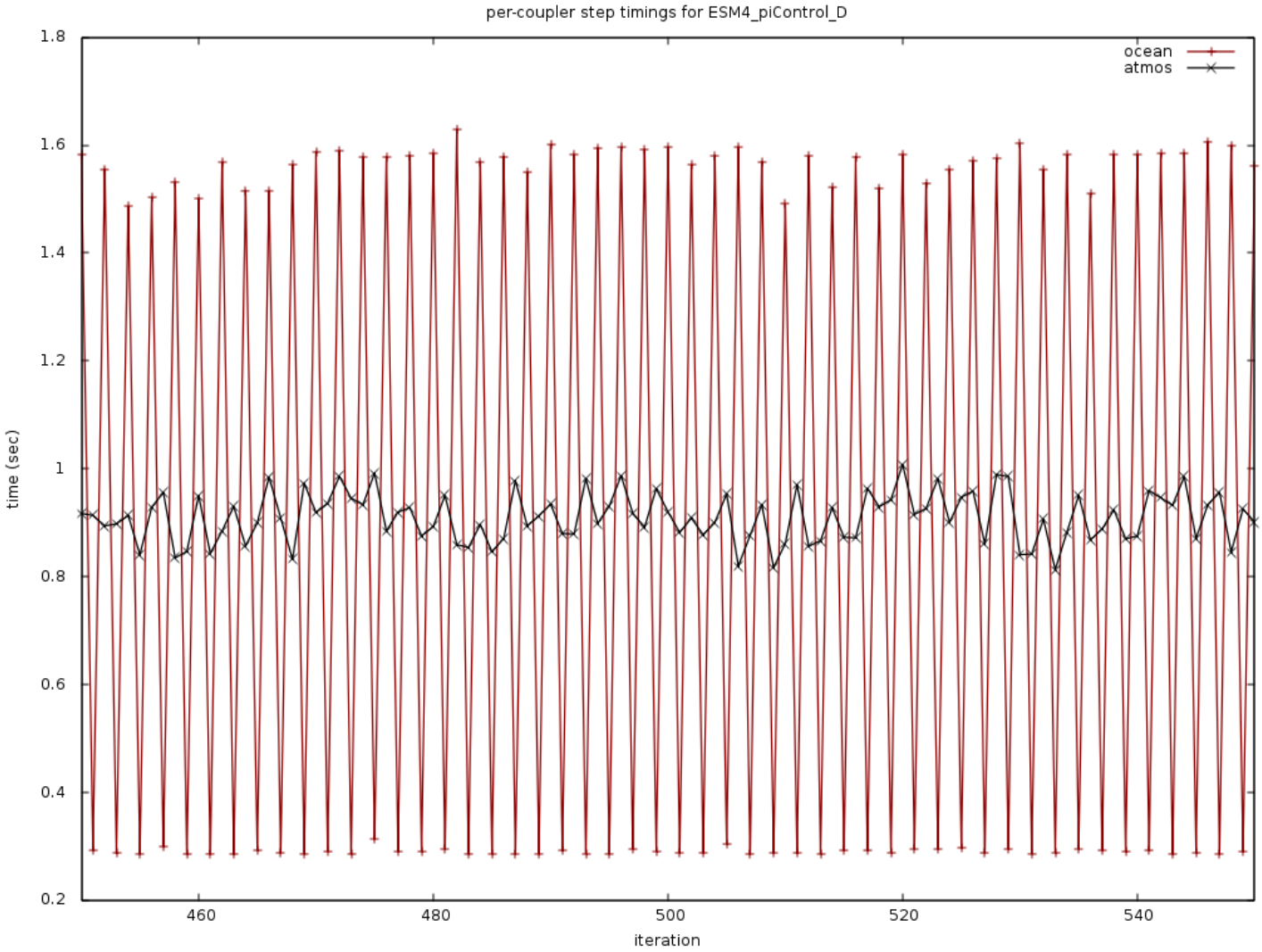

Figure 1: Time series plot of the per-coupler time step data for the ocean (OCN-rank 1436) and atmospheric (ATM-rank 1727) components for the ESM4 piControl baseline.

There are two approaches that can be taken to mitigate this cumulative, cancelling load imbalance. One is to improve the performance of tracer advection, thermodynamics, and diagnostics in MOM6. The other is to utilize appropriate timescales relative to the coupling step to promote uniformity in the component runtime

FIgure 2 has an example of this same ESM4 run optimized to give better elapsed time with fewer resources. On the coupling side, the atmospheric radiation time step and ocean tracer advection and thermodynamics timescale were both reduced to run every coupling step (3600s). To balance the components, the ATM core count was reduced by 33% from 3456 to 2304 total cores and the OCN core count was increased by 42% from 1437 to 2044 total cores. Even with increased complexity and a reduction in resource utilization of 12%, the Main Loop time was reduced by 6%. Coupling cost was reduced by a factor greater than ten to 0.02.

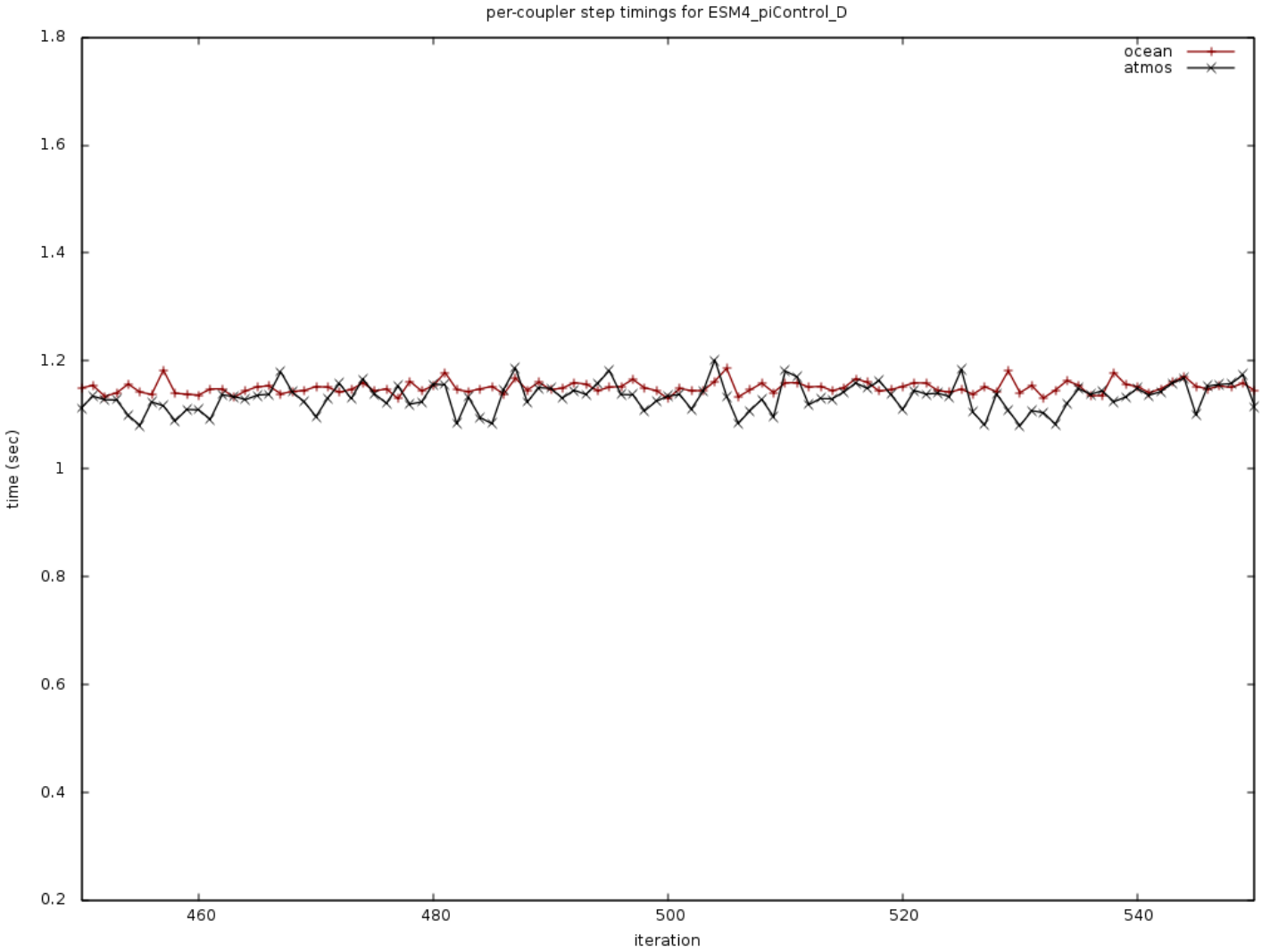

Figure 2: Time series plot of the per-coupler time step data for the ocean (OCN-rank 2043) and atmospheric (ATM-rank 1151) components for the optimized ESM4 piControl.

Diagnostics were mentioned previously as a potential inefficiency for the OCN component. To explore the impacts of diagnostics on performance, the optimized ESM4 piControl was re-run with diagnostics removed ("empty" diag table) with the results plotted in Figure 3. Analysis of the clocks attribute a mere 6% cost specific to diagnostics in the ATM component. The diagnostic cost for the OCN is more severe, with an unexpected penalty of 22.5%.

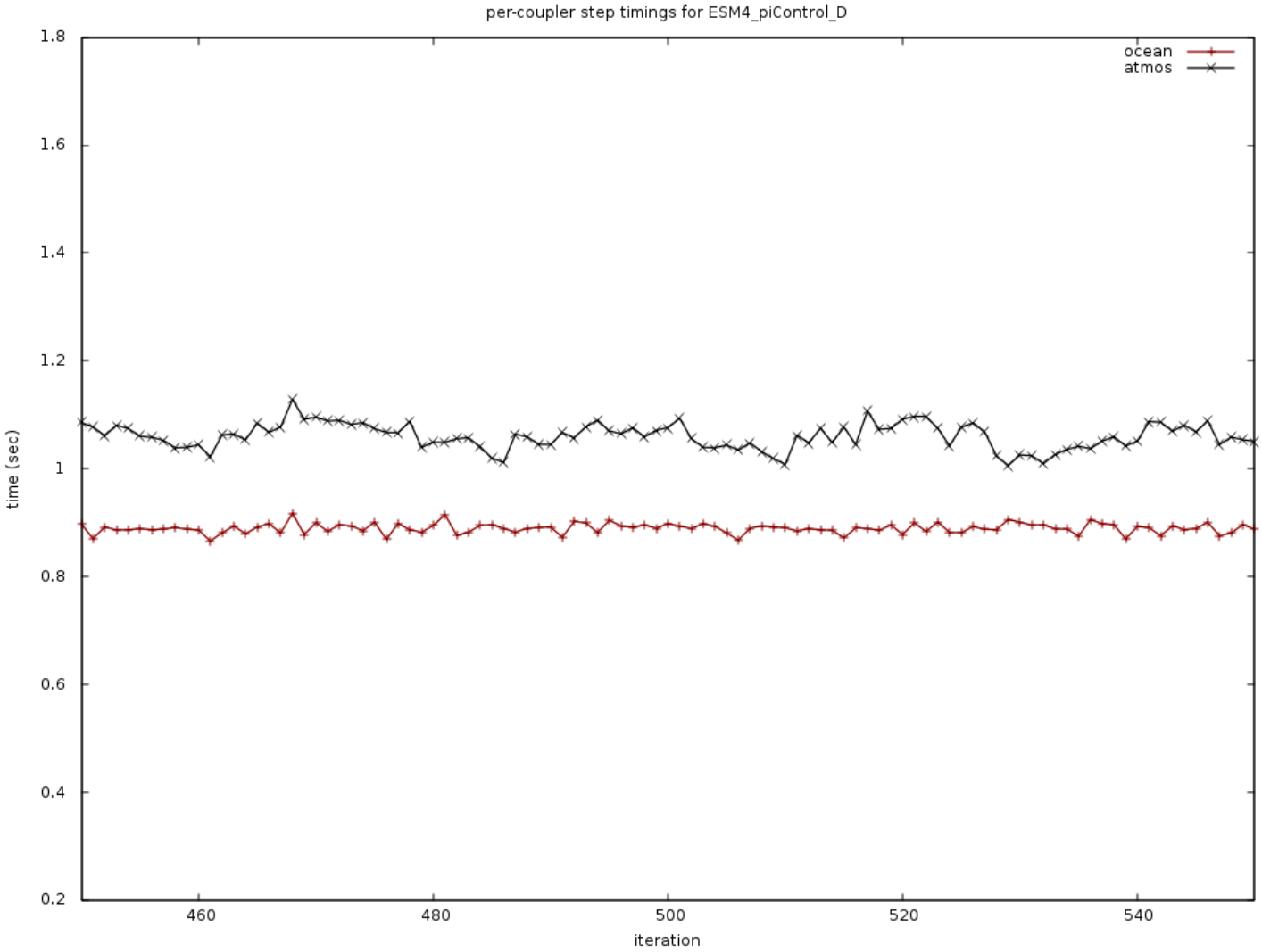

Figure 3: Time series plot of the per-coupler time step data for the ocean (OCN-rank 2043) and atmospheric (ATM-rank 1151) components for the optimized ESM4 piControl with diagnostics removed.

CM4-BLING piControl

The time series plot of a piControl CM4-BLING[4] baseline displays similar behavior to the ESM4 baseline with various peaks and valleys occurring in an easily discernible pattern. With the reduction in the coupling step from 3600s to 1800s, the OCN tracer advection and thermodynamics are now seen every fourth coupling step (Figure 4). Two distinct patterns are present in the ATM component. The first is associated with the radiation computing every sixth coupling step and a second higher frequency pattern linked to the radiation short-wave timescale being computed every other coupling step[5]. Because of these various patterns, the ATM component is on par or faster than the ocean component for every coupling step - except those including a full radiation calculation.

[4] Biology Light Iron Nutrient and Gas model - a simplified model of biogeochemical cycling (https://www.gfdl.noaa.gov/simplified-ocean-biogeochemistry-bling/)

[5] Tests have demonstrated the assertion, but results are not presented herein.

It is interesting and important to note that the different timescales for the expensive steps in ATM and OCN never overlap. The costly step for ATM alternates being the step before and then after the costly step for the OCN component. This is directly related to the radiation scheme needing to compute on the first iteration and then utilizing the time step from that reference.

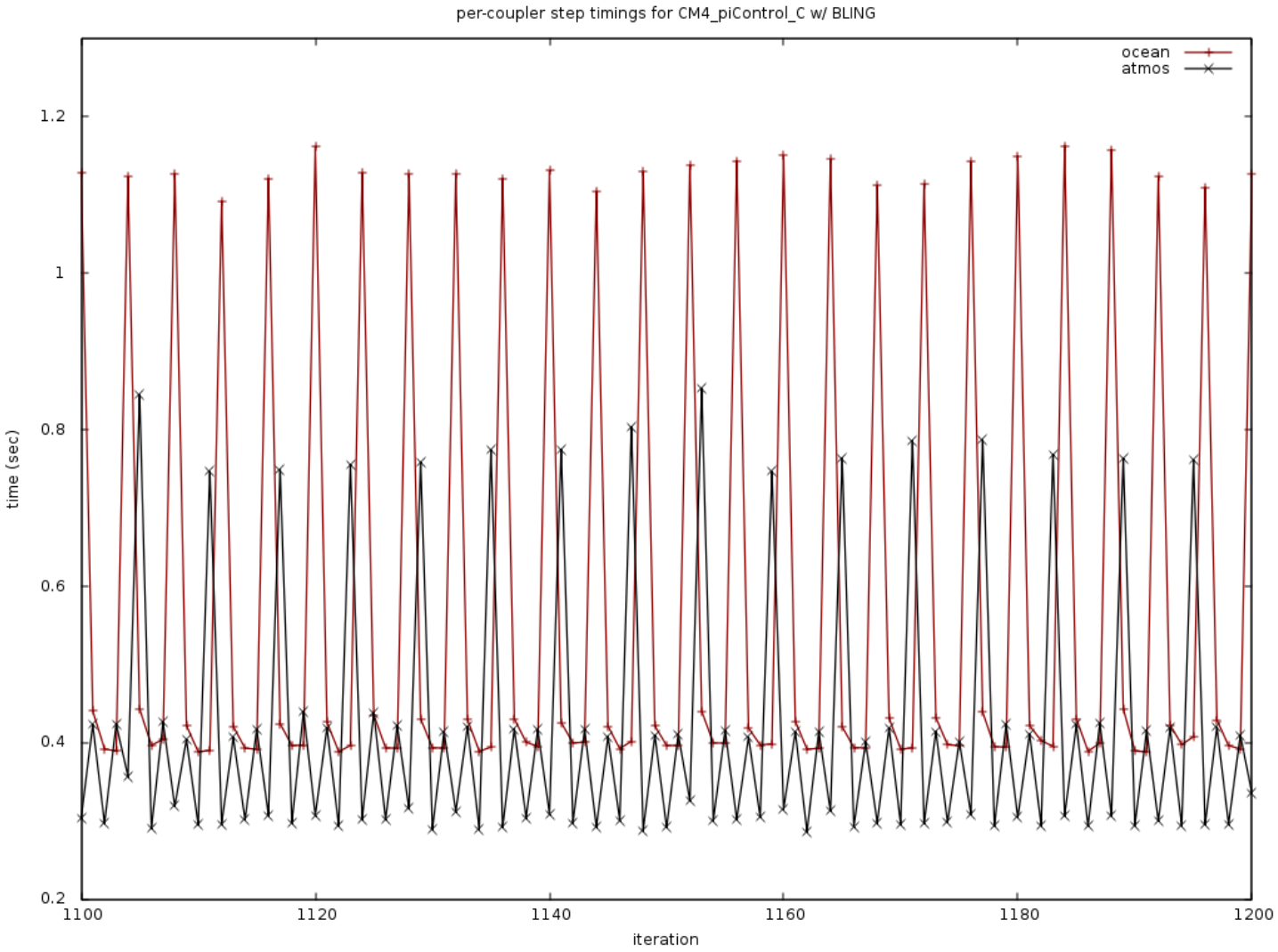

Figure 4: Time series plot of the per-coupler time step data for the ocean (OCN-rank 2996) and atmospheric (ATM-rank 431) components for the CM4-BLING baseline.

CM4 piControl

The only difference between CM4 and CM4-BLING is the absence of the simplified ocean biogeochemistry. Because of this, the ATM time series profile is basically unchanged. The OCN time series varies only in the upper limit (Figure 5). This change is directly related to the removal of tracers associated with the BLING scheme.

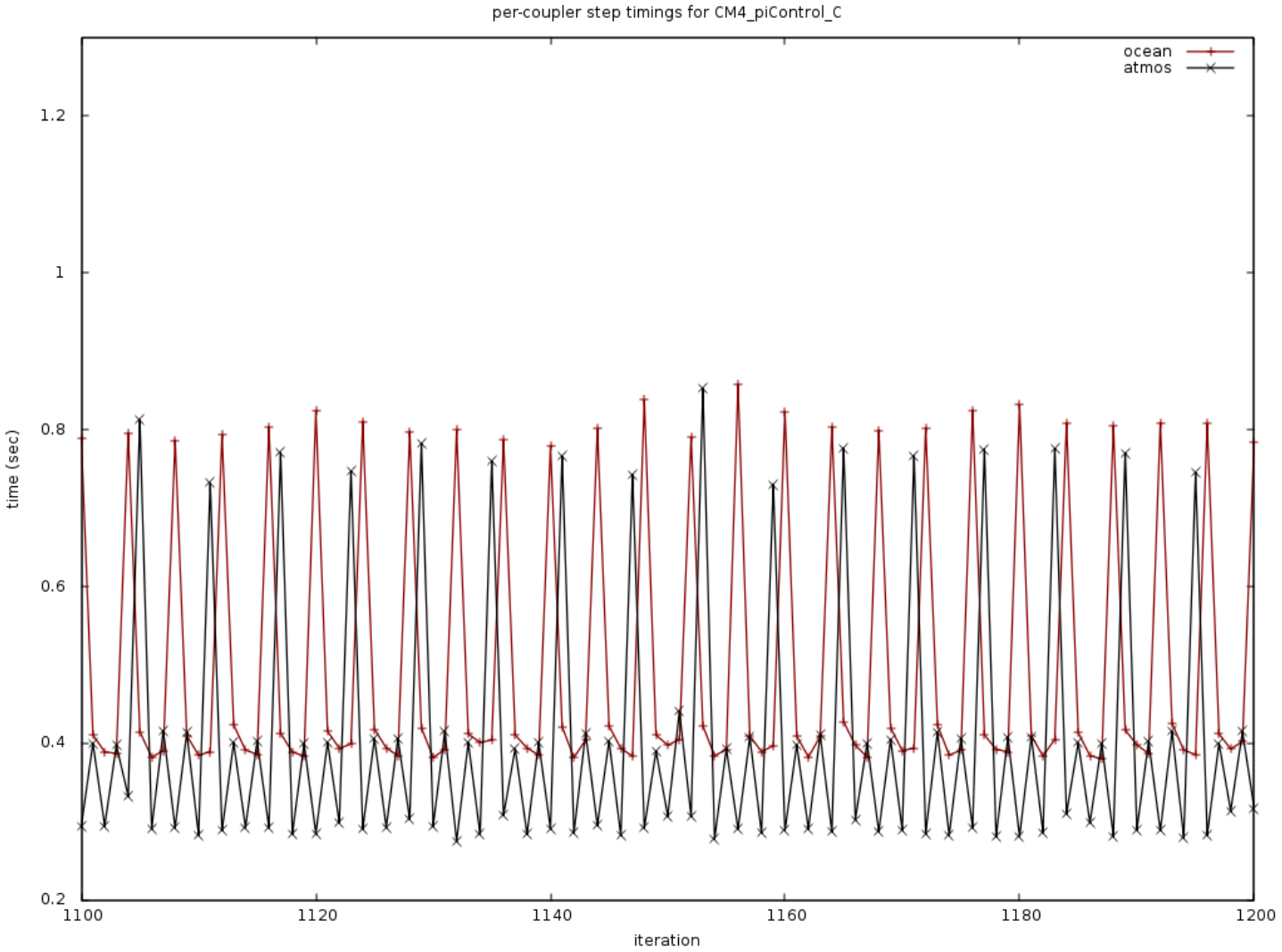

Figure 5: Time series plot of the per-coupler time step data for the ocean (OCN-rank 2976) and atmospheric (ATM-rank 431) components for the CM4 baseline.

Conclusions

It is often the case that choices for model component time scales are made attempting to trade off the cost of the component vs. fidelity of the science. As is evident in the data and figures presented, all compromises related to time scales in the model for performance reasons need to be re-examined as reduced component cost and potentially reduced fidelity do not necessarily lead to a faster model wall clock. Additionally, an examination of performance of OCN diagnostic logic should be undertaken and operational configurations must always begin with a review of requested diagnostics to ensure the list is comprehensive, but not excessive.

Acknowledgements

We thank the GFDL Front Office for their support of this Technical Note. We also would like to thank Lucas Harris, Jeffrey Durachta, and Venkatramani Balaji who, through their reviews and suggestions, made this technical note stronger. Special thanks to Kristen Schepel for managing the publication process, along with ensuring compliance with accepted publishing standards and preparation of the final copy, including creation of the title page.

References

Balaji, V., Maisonnave, E., Zadeh, N., Lawrence, B. N., Biercamp, J., Fladrich, U., Aloisio, G., Benson, R., Caubel, A., Durachta, J., Foujols, M.-A., Lister, G., Mocavero, S., Underwood, S., and Wright, G.: CPMIP: measurements of real computational performance of Earth system models in CMIP6, Geosci. Model Dev., 10, 19–34, https://doi.org/10.5194/gmd-10-19-2017, 2017.

Held, I. M., Guo, H., Adcroft, A., Dunne, J. P., Horowitz, L. W., Krasting, J., Shevliakova, E., Winton, M., Zhao, M., Bushuk, M., Wittenberg, A. T., Wyman, B., Xiang, B., Zhang, R., Anderson, W., Balaji, V., Donner, L., Dunne, K., Durachta, J., Gauthier, P. P. G., Ginoux, P., Golaz, J. C., Griffies, S. M., Hallberg, R., Harris, L., Harrison, M., Hurlin, W., John, J., Lin, P., Lin, S. J., Malyshev, S., Menzel, R., Milly, P. C. D., Ming, Y., Naik, V., Paynter, D., Paulot, F., Rammaswamy, V., Reichl, B., Robinson, T., Rosati, A., Seman, C., Silvers, L. G., Underwood, S., & Zadeh, N. (2019). Structure and performance of GFDL's CM4.0 climate model. *Journal of Advances in Modeling Earth Systems*, 11(11), 3691– 3727. https://doi.org/10.1029/2019MS001829