Dear reviewer,

Thank you for your interest in our paper and your feedback. In the following, we will directly respond to your comments.

Best regards,
Jan Gärtner

**Reviewer Comment**
The manuscript compares a demonstration case by Mehlmann et al (2021) with a stationary wind field and states that narrow features are observed. Are these right or wrong?

**Author Response**
As described in section 2.2, we conducted a comparative analysis between Veris and the MITgcm under a variety of initial conditions and forcing fields, including those used to produce the sea ice field shown in Fig. 1. The resulting differences between the models were on the order of $10^{-3}$ and did not accumulate over time. These findings demonstrate that Veris reproduces results consistent with the MITgcm, thereby validating the correctness of the simulated features.

We will clarify our comparison by additional text in the figure caption to indicate that the MITgcm produces the same features, as well as by adding a difference plot that demonstrates the randomness of the differences.

**Reviewer Comment**
Then it simulates Antarctic in order to show that the coupled system works. But again no real validation or discussion.

**Author Response**
Fig. 3 serves primarily as a proof of concept, showing that Veris can be integrated with Veros to form a fully Python-based coupled sea ice-ocean model. The accuracy of Veris itself was established through the direct comparison with the MITgcm presented in section 2.2. The simulation result of the coupled Veris-Veros model is mostly determined by the ocean model veros. Consequently, a thorough analysis of the Antarctic sea ice would primarily evaluate Veros rather than Veris.

**Reviewer Comment**
The largest focus is on optimization and they show scaling based on increased domain size. This Scaling is normally shown with increasing number of processors (including the timing of running on 1 processor). For climate models the limitation should be usage of the full bandwidth, which is not necessarily reached.

**Author Response**
We understand that the scaling is typically assessed with respect to the number of processors, rather than with domain size. However, due to limitations of the JAX backend, Veris does not yet support parallel execution when JAX is used. A scaling analysis with the number of processors is therefore not possible. Nonetheless, we believe that the presented domain-size scaling is still informative, as it demonstrates that Veris exhibits similar scaling behavior of computational cost with domain size similar to the MITgcm, demonstrating its usefulness in large scale simulation. This is especially relevant in the context of GPU-based applications, where Veris shows its most efficient performance.

**Reviewer Comment**
Line 15: This is a subjective thing. I don't think that Python is easier to read or maintain. I think that this depends on the programmer. I agree that Python is easier accessible as it does not require a compiler and that more people have used this.

**Author Response**
In the introduction, we highlight that the advantages of Python over Fortran, particularly in terms of ease of use and maintainability, are especially beneficial for new researchers or those with limited experience in low-level programming.

**Reviewer Comment**
Line 77: I assume this is theoretical speaking and if all resources are used.

**Author Response**
We will revise the manuscript as follows, with the changes marked in italics:

Line 77: Due to their higher core counts, GPUs are much better suited for parallel computing than CPUs*, provided that the available computational and memory resources are fully utilized.* As a result, they are more energy-efficient for parallel tasks such as climate simulations, as they complete these simulations faster than CPUs.

**Reviewer Comment**
Section 2.2 Validation: What is the setup for the dynamic test? 1000 iterations is a few days Is this long enough?

**Author Response**
We will revise the manuscript as follows, with the changes marked in italics:

Line 185: For the dynamics component, a two-dimensional benchmark was conducted *using rectangular grids at varying resolutions, with side lengths of 128, 256, 512, and 1024 grid cells.*
Line 191: The resulting relative differences between the models were on the order of $10^{-3}$ and did not accumulate over time, indicating that the variations were consistent and within acceptable numerical limits. *Observing no significant changes in these differences up to iteration 1000 is considered sufficient evidence that this number of iterations is adequate for model verification.*

**Reviewer Comment**
2.3 Parallelization. EVP cannot use JAX. How does the model speed up the dynamics in the JAX parallelization cases in the figure if it only do dynamics? This is normally the most expensive part.

**Author Response**
Veris and its solver EVP can use JAX. While Veris and its EVP solver support the use of JAX, data parallel execution is not available when using JAX as the backend. All benchmarks performed were therefore conducted using JAX on a single CPU.

**Reviewer Comment**
Section 3.1. Is this the same test as the dynamics? Described in section 2.2?

**Author Response**
In section 2.2, several tests were conducted using varying initial and forcing conditions, including — but not limited to — the simulation presented in section 3.1, Fig.1.