Review of GMD-2023-28 titled "Optimized Stochastic Representation of Soil States Model Uncertainty of WRF (v4.2) in the Ensemble Data Assimilation System" by S. Lim, S. K. Park and C. Cassardo

This study used a micro-genetic algorithm to perturb soil moisture and temperature in Noah-LSM within the WRF model (V4.2) to improve ensemble spread that can lead to the short-term prediction in the planetary boundary layer (PBL). Tuning parameters such as the amplitude and the horizontal decorrelation length/time scale of random forcing applied to soil states are evaluated based on 6-h forecasts in temperature and moisture in the boundary layer. While authors claimed that they improved ensemble forecasts by tuning parameters for random perturbations of soil states, this study only worked on an initial ensemble spread, which is different from the actual ensemble spread that grows with cycles. Also, it is not clear how the random perturbations added to soil temperature and moisture can represent the uncertainty of Noah LSM since the actual model uncertainty or error is never examined. The concept of ensemble cycling presented in this draft is either incorrect or vague to make the experiment design and figures not supportive of authors' main points. Numerous fundamental or structural issues are found throughout the manuscript. Unfortunately, authors might need to advance their understanding of the ensemble data assimilation system, as specified in my major comments below. I would recommend authors to spend enough time revisiting the issues, performing ensemble cycling experiments to examine the effect of perturbing land states on the weather prediction, with a clear target time in mind (as six hours is not the characteristic time scale where soil states are expected to significantly affect atmospheric forecasts). For that, I would recommend "Rejection" of the manuscript for now.

Major comments:

1. Lack of innovation: Various stochastic perturbation techniques have been already introduced in the WRF model and widely used to increase ensemble forecast skills (See the references in https://www2.mmm.ucar.edu/wrf/users/docs/user_guide_v4/v4.2/users_guide_chap5.html#stochastic). Authors should recognize all those related efforts specific to the WRF model (including the version 4.2 used in this study) and should justify why we need another perturbation algorithm despite all the corresponding options having been fully supported in the WRF system for a decade, already. For example, users can easily create a 3-D Gaussian random perturbation by simply turning on a namelist parameter (e.g., rand_perturb=1), and the capability of stochastically perturbing parameters also exists for RUC LSM, which can be easily expanded or applicable to Noah LSM. Because these are basically doing the same thing as what authors try to do here, with almost the same tuning parameters, it is mandatory to clarify the need of another algorithm for the same system.
2. Inappropriate references: Along the line, the micro-genetic algorithm should be introduced and understood as an alternative to the existing options available in the WRF model, but in the Introduction section, authors did not include any of the previous work specific to the WRF implementation. It is not convincing if the algorithm introduced in this study or the study *per se* could make any meaningful contributions to improving our understanding or ensemble forecast skills. One should not ignore others' decade-long efforts on the same system for the same problem (e.g., inflating ensemble spread).
3. Inappropriate title: Authors called it optimized representation, but I would expect a much more generic approach to optimize ensemble configurations in the context of a coupled system, not based on a single case study. To me, the presented study rather seems to be one of many ad-hoc tuning practices.
4. Clarity issues: The manuscript needs an extensive work on clarifications. It took me a while to figure out that this study dealt with initial ensemble spread, not the spread during cycling because authors mixed up the two different things throughout the manuscript and from the very beginning. The first statement in the Introduction, for instance, is incorrect: "The ensemble data assimilation

(EDA) describes both initial conditions (ICs) and model uncertainties represented by the flow-dependent background error covariance (BEC).” => In fact, EDA only “requires” initial ensembles to start cycling, and the initial ensembles are not described by EDA because they can be generated separately as this study shows. Also, EDA does not need to describe model uncertainties since there are different ways to construct ensembles without taking model uncertainties into account (e.g., perturbing observations). On the other hand, the general description of EnKF in GSI (e.g., Eqs. (9)-(15)) may not be necessary unless authors made any changes for this application.

5. Improper goal setting with poor experiment design and methodology: From my view, this is one of the most fundamental problems this work has.
    a. The motivation of this study seems to tackle the insufficient ensemble spread that can lead to poor forecast skills. When it comes to under-dispersive ensemble systems, however, the actual problem lies on the reduction of ensemble spread with cycles, not the spread in the initial ensemble. As it takes at least dozens of cycles to saturate the ensemble spread in the regional cycling system, the initial ensemble construction certainly matters to the efficiency (e.g., how quickly the spread grows), but that is a fundamentally different problem from the filter divergence issue where observations are gradually rejected due to the lack of spread.
    b. Moreover, this study focuses on the uncertainties of soil states, it is thus critical to have the land states that are well spun up before the initial time. Considering the imperfect land surface model with no land data assimilation, the characteristic time and spatial scale of soil states, and the initialization from the NCEP-FNL at coarse resolution (1°x1°), a three-day spin-up used in this study is not even close to the very minimum requirement (say, a month).
    c. The experiment design is not well described, but if only 27 ensemble members were used, covariance localization must have been used for such a small ensemble size. In that case, the localization would certainly affect ensemble spread and additive perturbations, but are not found anywhere in this script.
    d. The construction of an initial ensemble needs clarification: Was the random CV option in WRFDA used to perturb atmospheric variables for a 5-member ensemble (L232-235) while the micro-genetic algorithm was used only for soil perturbations in a 27-member ensemble (L239)? How were the two different ensembles combined in your experiment, then? The choice of ensemble size is critical to ensemble spread, but the description of the ensemble system is unclear.
    e. Table 1: How did you decide the optimized ranges for soil moisture and temperature? That range used in this study seems to be ad-hoc, not necessarily representing either model or observation uncertainty.
    f. Authors defined a fitness function in Eq. (8) to determine the best parameter values among several candidates. Although it is true that meteorological variables in the boundary layer are closely tied to land states, atmospheric fields are characterized at different time and spatial scale from that of soil states. Considering the response time of atmospheric variables to soil perturbations as well as many other potential contributors to the boundary layer structure (such as advection, convection, radiation, and clouds), it is questionable if the fitness function based on 6-h forecasts can fully capture the actual impact of soil perturbations on boundary layer forecasts or can be used as a proxy to the optimal parameter settings.

    This is indeed one of the most complex issues to disentangle clearly, but it is my concern that the approach used in this study seems overly simple to resolve such a challenging issue. Anyhow, given that the WRF system already provides various perturbation techniques, another perturbation algorithm cannot guarantee a noble work, and the only way I can see to make this type of work meaningful is to examine how the initial ensemble affects the ensemble forecast skills for a long period of time in a statistical manner (e.g., not in a single case).

6. Figures not supportive of main points:
   - Figure 3: If this study is all about inflating spread, it is expected to show ensemble spread, not a single member of soil states.
   - Note that an initial ensemble is only a start of cycling and is not supposed to represent the saturated ensemble spread. Hence, Fig. 6 is not needed.
   - Figures 7, 8, and 10 only show the sensitivity to the initial ensemble, not the optimal ensemble spread.
   - Figure 9: Again, we do not expect the saturation of ensemble spread at the initial time.
   - Figure 11: Analysis increments at the initial time are not indicative of the system performance, and the GFS analysis is not quite trustworthy near the surface.
7. Section 2.1 is named WRF-Noah LSM Coupled System. As far as I know, Noah-LSM is just one of the physics options available in WRF. Did you develop/change anything to enhance the coupling part either in the model or in your analysis step? As all the physics parameterization schemes are interacting with each other within the WRF framework, it is not clear why authors emphasized the “coupled” system here. How does your Noah LSM work differently from all other studies using the same option, again from the modeling or DA aspect?