Hierarchical System Development for the Unified Forecast System

-long-term vision and plan for its progressive implementation

Xia Sun[a,b,d], Tracy Hertneky[c,d], Ligia Bernardet[b,d] (Co-lead), Tara Jensen[c,d] (Co-lead), Yi-Cheng Teng[e,f], Stelios Flampouris[e,f], Lulin Xue[c,d], Weiwei Li[c,d], Louisa Nance[c,d], Mike Ek[c,d]

a. *Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, CO*
b. *NOAA Global Systems Laboratory, Boulder, CO*
c. *NCAR Research Applications Laboratory, Boulder, CO*
d. *Developmental Testbed Center, Boulder, CO*
e. *Tomorrow.io, Boston, MA*
f. *Earth Prediction Innovation Center, Riverdale, MD*

June 2023

Developmental Testbed Center is a distributed facility where the NWP community can test and evaluate new models and techniques for use in research and operations.

Contents

I. Introduction

Despite significant improvement of model development and computational resources, model forecast issues still exist in numerical weather prediction (NWP) and Earth system models (ESMs), including the Unified Forecast System (UFS). Examples of forecast challenges can be referred to the [Model Issues and Forecasters' Request list](#) from the UFS Forecasters' workshop (Sims et al., 2021). The Hierarchical System Development ([HSD](#)) (Ek et al., 2019) approach stands out as an efficient way for model development, enabling the community to have multiple entry points for research efforts spanning from simple processes to complex systems, as illustrated in Figure 1. Note that process-level assessments are made at all HSD testing steps, and that HSD provides understanding of both processes and systems. To accelerate research to operations (R2O) activities for UFS, a strategic plan for HSD in UFS is needed.

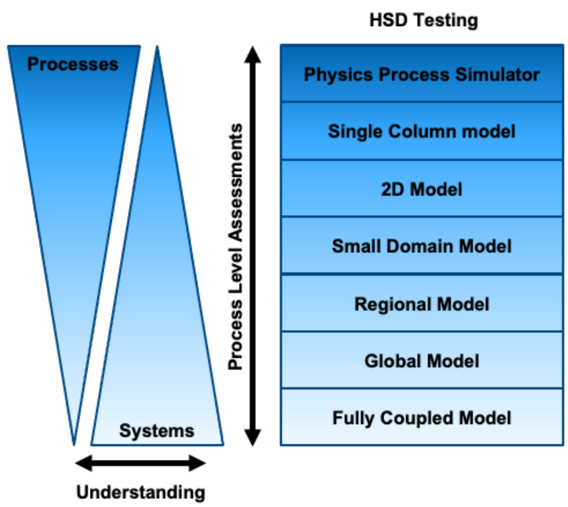

Figure 1. Schematic of the Hierarchical testing concept, from simple to complex.

The scientific evaluation for specific forecast problems of the model components is a prerequisite for the evolution of forecasting capabilities through higher model resolution or a more accurate representation of the physical processes (e.g. the High-Resolution Rapid Refresh (HRRR); James et al., 2022). But, as the complexity of the forecast systems increases, the code complexity causes the increase of computer resources for testing, evaluation, and operational implementation. The necessity for adopting hierarchical testing to include the functionality of the code and fundamental principles of best practices at all levels, from individual components to end-to-end systems tests, has been established and has become the golden standard in numerical forecasting (Teng and Flampouris, 2023). Also, the numerical aspect of prediction models has a significant impact on the operational implementation of any scientific feature; having a standardized assessment mechanism across all the stages of the development for the continuous evaluation of the computational performance is a critical requirement for community development and eventually acceleration of the Innovation-to-Operations (I2O, Hoffman et al., 2018). Therefore, the expansion of the relevant infrastructure of the HSD framework to include software testing on top of the scientific advancement becomes part of the systematic development of forecast models.

These two testing attributes have been adopted as developmental principles among the UFS model components, with several of them being implemented in various independent testing infrastructures. For instance, the "regression test" capabilities of the WAVEWATCH III Framework (WW3DG, 2019) are significantly broader than the name indicates and include the whole spectrum of testing or the testing infrastructure of the MOM6 (Adcroft et al., 2019). Despite these significant efforts, there is a lack of systematic approach and engineering infrastructure within the UFS. In addition, the survey results accompanying this document in Appendix A indicate that less than 40% of participants are aware of the hierarchical testing capabilities of the UFS; the UFS community has minimal experience with the software tools that support HSD, and the limited adoption of hierarchical testing indicates the narrow or maybe superficial understanding of the testing technologies, how it is to be implemented, and the value of HSD in accelerating the research.

No single unique model hierarchy exists. The definitions of HSD vary in different communities and depend on the scientific questions of interest. We are conscious of the different recognitions in HSD in terms of scientific research and software engineering testing. Here we mainly focus on the HSD contribution to scientific advancement. Available atmospheric models have been categorized based on complexity (Bony et al. 2013) and model configurations (Jeevanjee et al. 2017). Distribution of models can be defined based on model complexity, from conceptual models to fully-coupled ESMs, and system complexity, from particle systems to Earth & Human Systems. In terms of model configurations, Jeevanjee et al. (2017) describes the climate model hierarchies based on aspects of dynamics, boundary forcing, and bulk forcing. An alternative description is proposed in Maher et al. (2019), who organizes model hierarchies based on three principles in research areas of large-scale atmospheric circulations. The first principle is a dynamical hierarchy of atmospheric fluid flow, which allows investigation of the importance of different scales on the governing equations. The second is a process hierarchy of the atmosphere. Process hierarchies include two aspects: a) stepwise integration of atmospheric processes into the governing equations; and b) construction of model hierarchy by increasing the complexity of the boundary conditions. The third principle is hierarchy of scale, pertaining to the configurations of model domains and grid resolutions.

The hierarchies discussed lays the foundation for understanding of the underlying principles and structures for HSD in UFS. These hierarchies serve as a stepping stone for the subsequent section, where the axes for HSD for UFS will be proposed. The forthcoming proposed aspects for HSD for UFS aim to refine our understanding and offer perspectives for a hierarchical testing framework that fosters model innovations and facilitates transitions from R2O. The axes, or aspects, of hierarchical system development are discussed in Section II. The existing capabilities in UFS are described in Section III. Proposed plans with recommendations using results from a community survey are presented in Section IV.

II. Axes of Hierarchical System Development for UFS

In this section, we describe the axes of HSD that could be applied to the UFS framework by introducing the axes first and then highlight selective cases that advance the understanding of atmospheric processes. The examples presented herein serve as an illustration of the HSD axes and do not represent a complete list of preferences. They also focus primarily on the

Atmospheric component but it should be noted that there are analogous axes, or aspects, of HSD for component models of a coupled system as well. The prioritized rankings from a community survey will be provided subsequently.

A. Sample Sizes

To identify a model problem and later test model innovations, case studies are an indispensable component during this process. During the model issue identification stage, one common approach is to start from one case study and expand to multiple similar cases to determine any systematic biases. Those cases that illustrate the model biases provide entry points for researchers and developers to initiate their own work towards improving models. Another approach used in the UFS community during the model development stage is to conduct retrospective runs lasting from weeks, months, and/or years, to demonstrate improved statistical diagnostics compared to the older model version.

B. Hierarchy of Scales

Global models are usually configured at relatively coarse resolutions due to the substantial computational costs when run at higher resolution. In order to capture regional features, high resolution modeling is required to better capture the fine-scale atmospheric features, processes and interactions, especially over complex terrain. This coarse-to-fine resolution and global-to-regional domain evolvement is one component of HSD axes. Downscaling techniques are needed to model regional climate and weather at high spatial resolutions over a specific region. Two dynamic downscaling and grid refinement strategies, including nested modeling and variable-resolution modeling, will be discussed here.

Nested limited area models (LAMs) are usually forced by output from global models or reanalysis datasets. Nested model simulations combine the strengths of fine resolution to capture small-scale variability of atmospheric dynamics and physics, and relatively realistic lateral boundary conditions provided by the outer domain for real-world cases. High-resolution nested LAM simulations can have one-way nesting or two-way nesting. Interactions between outer and inner domains are turned on in a two-way nesting configuration or absent in a one-way configuration. The performances of nested modeling have been rigorously assessed and its advantages have been demonstrated in many previous works (Moeng et al., 2007; Talbot et al., 2012; Hagos et al., 2013; Mohan and Sati, 2016). Nested modeling are widely employed in research areas of land-atmosphere exchange (Sun et al., 2020; Alexander et al., 2022), wind energy application (Liu et al, 2011), cloud-radiation interactions (Archer-Nicholls et al., 2016), convection (Heath et al., 2017), *etc*.

The most recent technique, variable resolution modeling (VRM), refers to the capability to use a global model in a relatively coarse resolution with high resolution over a specific region. The variable-resolution-enabled global models allow for fine resolution in a global modeling framework to capture the fine-scale atmospheric features and processes at a reduced computational cost compared to global high-resolution simulations. The fidelity of VRM has been demonstrated in simulating climatological quantities (Huang et al., 2016), marine air penetration (Wang and Ullrich, 2018), tropical cyclones (Zarzycki et al., 2014), and snowpack projections (Rhoades et al, 2018) using the variable-resolution of Community Earth System Model (CESM). Challenges that need to be addressed in this uniform modeling framework include scale-aware physics parameterizations, differential topography smoothing (Zarzycki et al., 2015), and grid generation. The tool of SQuadGen (Ullrich 2014) has been used to generate VR grids for usage in CESM. Fig. 2 shows an example of a VR grid.

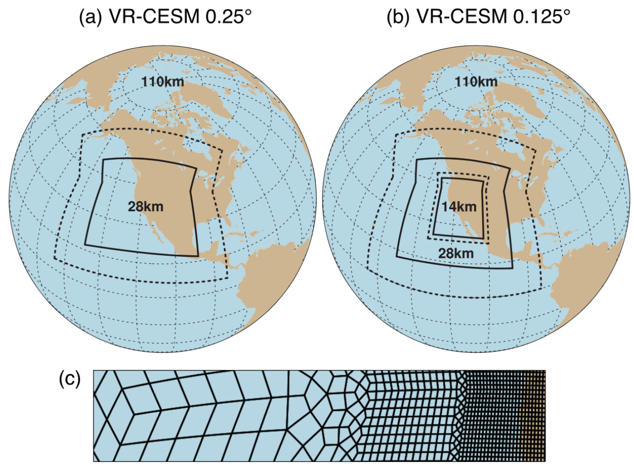

Figure 2. Illustration of grid spacing in the (a) VR-CESM 0.25° and (b) VR-CESM 0.125° meshes. (c) A depiction of the transition from the global 1° resolution mesh through two layers of refinement to 0.25° and again to 0.125° (Huang et al., 2016).

C. Simulation Realism

Idealized simulations refer to the use of simplified models that can help improve theoretical understanding of atmospheric processes and identify how the atmospheric features and processes of interest interact with others. The idealization of complex modeling systems is an important component of HSD to help better understand and potentially address model biases. Here we highlight several idealized simulations that have been used extensively and have led to progress in key scientific research areas related to the Earth system.

Radiative-convective equilibrium (RCE) is a conceptual model of the tropical environment that can help address a variety of scientific questions related to tropical meteorology and climate. RCE refers to a state where the radiative cooling of the atmosphere counterbalances the convective heating. The balanced state can be achieved by prescribed constant solar radiation and uniform boundary conditions with fixed values of sea surface temperature or a slab ocean model (Wing et al. 2018). RCE simulations have been conducted to investigate convection organization (Muller & Bony, 2015), coupling of clouds and convection to the climate system (Becker et al. 2017), the predictivity of precipitation extremes (Pendergrass et al. 2016), *etc*.

Aquaplanet is a simplified version of a global model, which idealizes the earth as an all water-covered surface without land, topography, sea-ice, and seasons. Aquaplanet configurations can be achieved by setting up a time-invariant and zonally symmetric sea-surface temperature (SST) profile. The SST profile is an idealized approximation to the Earth's surface temperature, with a maximum of 27°C at the equator and drops off to 0°C at 60° latitude (Blackburn et al. 2013). Aquaplanet simulations are an important tool to advance understanding of atmospheric processes and improve their representations in global models. For example, Landu et al. (2014) used an Aquaplanet simulation to investigate the dependence of the intertropical convergence zone (ITCZ) structure on resolution and dynamic core. Their results suggest that structure of the ITCZ depend on the feedback between convection and large-scale circulations. Rios-Berrios et al. (2020) used a community model, the Model for Prediction Across Scales-Atmosphere (MPAS-A), to conduct Aquaplanet experiments. They found that the simulated tropical precipitation variability depends on both the cumulus parameterizations and the coupling between physics and dynamics in climate and weather models. Aquaplanet has also been used to understand responses of comprehensive climate models to forcing (Medeiros et al. 2015). The role of land-surface effects on damping the adjusted forcing was demonstrated by the slightly larger adjusted forcing produced in aqua planets with quadrupling atmospheric CO_2 vs Earth-like configurations. Fig. 3 illustrates an Earth-like configuration and aquaplanet configuration.

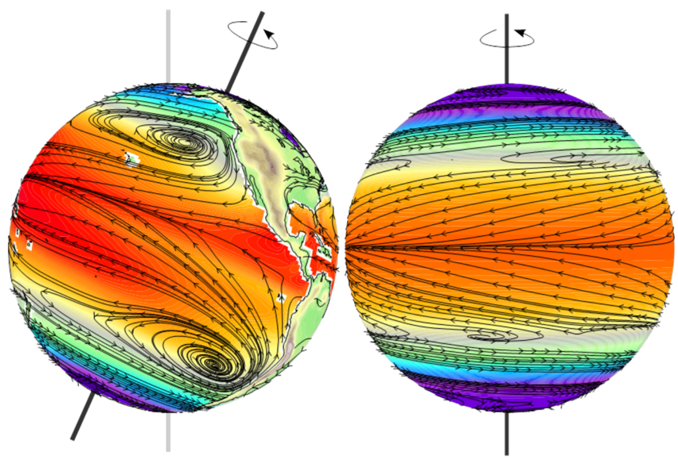

Figure 3. Illustration of the (left) Earth-like configuration in the Atmospheric Model Intercomparison Project (AMIP) and (right) aquaplanet configuration. Color shadings are SST and topography. Streamlines represent the annual mean flow at 925hPa (Figure from Medeiros et al. 2015).

Warm bubble is another test that is applied in idealized simulations to initiate convection. It is achieved by modulating the initial conditions with a horizontal and vertical radius setup, temperature perturbation, and/or moisture perturbation for the warm bubble. Warm bubbles are usually much larger than the thermals within a convective boundary layer, and thus differ from real world convective initiation. Warm bubbles have been used for research related to convection (Costantino and Heinrich 2014; Morrison et al. 2020; Mulholland et al. 2021).

Idealized mountain wave simulations are used to increase understanding of the effects of wind flow over simplified terrain, orographic lifting and convection, turbulence structure such as wave breaking, and blocking effects. Idealized mountain flow has been configured in 2D and 3D space, with varying shapes, such as the *witch of Agnesi* profile or two hills with a valley, and varying roughness. The flow itself can be defined by its moisture content (moist vs dry), variation in wind speed and direction, and stability (Mayr and Gohm 2000; Miglietta and Rotunno 2005; Fuhrer and Schär 2005; Kirshbaum and Durran 2005)

D. Mechanism/Interaction Denial

The impacts of each mechanism or multiple mechanisms can be explored with sensitivity experiments by switching a given mechanism(s) off. A good example of mechanism denial study is provided by a study on the Madden-Julian Oscillation (MJO) using an atmospheric general circulation model. Kim et al. (2011) explored the impacts of wind induced surface heat exchange (WISHE), cloud radiation interactions (CRI), and frictional wave-CISK (conditional instability of second kind) (FWC) mechanisms. The mechanisms were turned off by using prescribed daily climatological seasonal cycles of surface latent heat flux, net radiation heating rate, and surface wind stress, correspondingly. They found CRI and WISHE are both important mechanisms for MJO amplitude and propagation speed, while FWC has less systematic impacts.

Mechanism denial has also been applied in large-eddy simulations to advance understanding of the underlying mechanisms that are important for specific atmospheric processes of interest. Stratocumulus-to-cumulus transitions (SCT) (often found in the subtropics) have substantial impacts on the radiative budget and have been notoriously difficult to simulate in either NWP models or EMSs for climate. Zheng et al. (2021) conducted a mechanism-denial experiment to investigate the role of surface latent heat flux (LHF) on the SCT. A large-eddy simulation model, the System for Atmospheric Modeling (SAM) v 6.11.6 (Khairoutdinov and Randall, 2003), was used to simulate SCT during the Atlantic Stratocumulus Transition Experiment (ASTEX, Albrecht et al. 1995). The mechanism-denial was achieved by setting the LHF to a fixed value, versus a control experiment in which the LHF in tandem with an increasing sea surface temperature is used to mimic the impact of cold-air advection. Their research confirms the key role of LHF for SCT.

A simple change in physics parameterization schemes would involve complex interactions between the thermodynamics and large-scale dynamics. Interaction denial is a useful tool to untangle impacts of model physics of interest. One approach to apply interaction denial is through piggybacking, where this approach uses two sets of thermodynamic variables in a single simulation (Grabowski 2019). The two sets differ in the specific element related to the model physics being investigated, such as a parameter change in the microphysics scheme. In Set 1 at the 1st simulation, thermodynamics affects the buoyancy and thus is coupled to general dynamics while Set 2 piggybacks/uses the flow but does not impact dynamics. In the 2nd simulation, the two thermodynamic sets are switched. Fig. 4 shows the schematic of the piggybacking method. The (driver-piggybacker) differences between the two simulations represents the impacts of model physics being investigated, excluding the impacts of the scheme on dynamics. It was first proposed by Grabowski (2014) and illustrated by simulations of shallow convection using the anelastic Eulerian–semi-Lagrangian (EULAG) model with a single-moment cloud microphysics scheme. The approach was later extended to the research area of microphysics impacts on deep convection (Grabowski, 2015, and Grabowski and Morrison, 2016). Key findings from their studies include: no convective invigoration exists above the freezing level and the anvil cloud fractions strongly depend on the small cloud condensation nuclei. Lately, piggybacking was implemented in the widely-used WRF model. Refer to Appendix A in Sarkadi et al. (2022) for details on the implementation of the piggybacking technique into WRF.

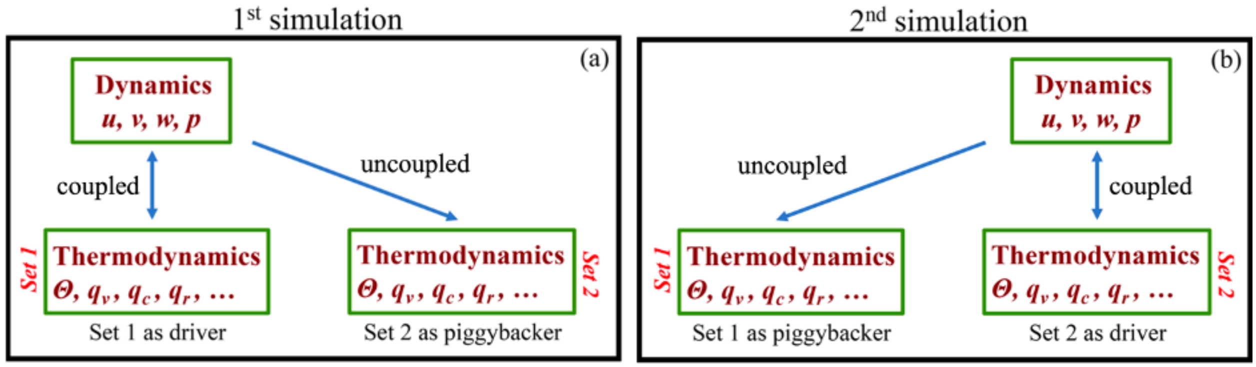

Figure 4. Schematic illustrating the piggybacking approach (Adapted from Grabowski 2019).

Another well-known application of interaction denial is the Single Column Model (SCM), which simulates a single atmospheric column as it evolves through time (Betts and Miller 1986; Randall et al, 1996). While the physics remain identical to those used in GCMs, the dynamical interaction from the large-scale circulation is absent and instead replaced by large-scale and advective forcing. This setup creates an ideal method in which to study the small scale processes that are entirely determined by the internal process balance and large scale forcing. One drawback is that SCM simulations are sensitive to differing large-scale forcing datasets (Xie et al. 2003). Aside from SCMs being initialized from observations or an idealized model, they can be forced with a column from a GCM, demonstrating the "replay" capability. The replay mode is a powerful tool that is useful in investigating systematic model biases, especially in locations that lack large-scale forcing observations. SCMs provide an easy, extremely computationally efficient way to study a subset of processes or a single process, assess interactions between physical processes, and even compare different physics packages. Oftentimes, SCMs are evaluated in parallel to Large Eddy Simulation (LES) or other idealized model hierarchies. Dal Gesso (2015) used a hierarchy of models, including SCMs, to assess the dependence of stratocumulus clouds on the free-tropospheric thermodynamic conditions. Zheng et al (2020) and Neggers et al. (2017) utilized SCMs to evaluate low-level marine stratocumulus-to-cumulus transition, which is not well understood or accurately modeled, and provided critical insights for model developers related to cloud-top entrainment issues and precipitation processes.

III. Existing Capabilities

A. UFS Applications

The UFS Weather Model can be run in several configurations (Table 1), from applications including only a single model component, to sub-seasonal to seasonal fully coupled model applications. Publicly-released UFS Applications (Apps) consist of end-to-end systems of the Short-Range Weather (SRW, UFS Development Team, 2022) and Medium-Range Weather (MRW) Apps for running the UFS Weather Model. These applications enable UFS users different pathways for community research that can contribute to testing, evaluating and improving operational forecast models (R2O). The SRW App version 2.1.0, released on 17 November 2022, is meant to address atmospheric predictions on limited spatial domains for timescales of sub-hourly to several days. The MRW version 1.1.0, released on 6 October 2020, is a global model supporting several grid resolutions for timescales of about 2 weeks. Development is underway for the next MRW App release, to adopt the global workflow that has historically been used for development and operations of the Global Forecast System (GFS) and Global Ensemble Forecast System (GEFS). Both the SRW and MRW comprise a workflow and build system that includes preprocessing utilities, the Common Community Physics Package (CCPP), the forecast model, and a post processor. Additionally, the SRW components include the enhanced Model Evaluation Tools (METplus, Brown et al., 2021) verification system for forecast evaluation, with plans to expand this capability to all UFS applications. Development for hurricane application is on the forefront to be integrated into the UFS using CCPP and includes a high-resolution storm-following nest, multi-scale data assimilation, and 3D ocean coupling.

Table 1. UFS component models used in each application (retrieved from https://ufscommunity.org/science/aboutapps/). These components include the FV3 atmosphere, the Noah and Noah-MP land, the Modular Ocean Model 6 (MOM6), the Los Alamos sea-ice model 6 (CICE6), the Goddard Chemistry Aerosol Radiation and Transport (GOCART) aerosol model, the Ionosphere-Plasmasphere Electrodynamics (IPE) model, the Advance Circulation (ADCIRC) model for storm surge, tides, and coastal circulation, and the WAVEWATCH III wave model.

	Atm	Land	Ocean	Sea Ice	Aerosol	Ionosphere	Storm Surge	Wave
Medium-Range Weather	1	0						0
S2S	0	0	0	0	0			
Hurricane	0	0	0					0
Short-Range Weather	0	0						
Space Weather	0	0				0		
Coastal							0	0
Air Quality	0	0			0			

B. CCPP and CCPP SCM

Model biases are often attributed to inadequate representation of sub-grid scale processes, and regardless of improvements made over the past years, many issues remain. An analysis of model development by Jakob (2010; 2014) showed that to improve this process, connections between process-level studies and model developers should be improved. One way to accelerate the research and development capabilities of physical parameterizations is using the Common Community Physics Package (CCPP) to connect the levels of the HSD. The CCPP consists of physics parameterizations and suites, and a framework which connects the physics to model dynamics. This infrastructure permits the research community to test and develop parameterizations to support the transition from R2O.

Currently, the CCPP is a component of the Unified Forecast System (UFS) ranging from limited area to global models, is packaged with the CCPP Single Column Model (Firl et al. 2022), utilized by the U.S. Naval Research Laboratory in their experimental NEPTUNE (Navy Environmental Prediction System utilizing the NUMA Engine), and has been selected for use in the NCAR models. Several cases are supplied with the SCM in CCPP-SCM format. Additionally, the SCM can be run with case input data in the Development and Evaluation of Physics in atmospheric models (Développement et Évaluation PHYsiques des modèles atmosphériques, i.e., DEPHY) format, an internationally-adopted data format for use with SCMs and LESs. A recent addition to the CCPP SCM is the ability to run using an arbitrary physics subset of a given physics suite, i.e. "on/off" switches for a particular physics parameterization or set of parameterizations, with data models replacing those parameterizations switched off; this capability allows the impacts of changes to be isolated for better understanding when examining systematic biases; this capability still needs thorough testing, further refinements and expanded documentation to fully meet the needs of the community.

C. METplus

A critical element of the HSD and R2O strategy is performing verification using well-defined quantitative metrics to assess a system's strengths and weaknesses before implementation. The METplus framework, developed at the Developmental Testbed Center, provides an array of verification and diagnostic tools available to the community. METplus has been adopted by research institutes world-wide and is a component of the UFS SRW App, and will be added to MRW, HAFS, Seasonal Forecast System (SFS), Space Weather, and other UFS applications.

D. UFS Case Studies Platform

The UFS Case Studies Platform (https://ufs-case-studies.readthedocs.io/en/develop/) provides 10 cases that reveal major forecast challenges in GFS. The case collection covers atmospheric phenomena of winter storms, hurricanes, extreme temperature, convection, and low-level inversions (Sun et al. 2021). This platform serves as a repository for the research community to gather the necessary resources and instructions to conduct model runs using different UFS weather applications including the public releases of the UFS Medium-Range-Weather (MRW) App, UFS Short-Range-Weather (SRW) App and the UFS Weather Model. Specifically, the platform provides initial condition datasets hosted on cloud storage, model configuration, and setup, and high-level preliminary simulation results compared with reanalysis data. It also includes the examples of applying hierarchical testing framework to investigate the 2020 July CAPE case and the 2020 Cold Air Damming (CAD) case. It should be pointed out that there are many other HSD steps that are not fully captured by use of the MRW and SRW alone, where these other HSD steps can provide more in-depth evaluation of e.g. model physics not easily assessed by using the primary UFS applications (SRW, MRW). Relevant cases studies could be contributed by the UFS community.

IV. Proposed Plan

The goal of this document is to help establish plans for future development of HSD for UFS. A survey was designed and sent out to the community in April 2023 to gather insights and feedback to help shape the future of HSD for UFS. Detailed survey results are summarized in Appendix A. The following proposed plan is built upon the views of the authors and contributors of this white paper, as well as the survey results that were collected from across the NWP and Earth system research and modeling community. The action plan contained herein focuses on first steps, which encompasses the use of the HSD approach for model improvement, including testing of individual ESM elements (e.g. physics schemes). The HSD framework should be developed in a way that once the first steps are completed, it can sustain progressively connecting elements with increased coupling between ESM components at the different HSD steps, all the way up to a complex fully-coupled ESM.

A. Recommendations and action plan

Recommendations and actions are suggested below for the four UFS HSD axes mentioned in the above section. An illustration of the proposed HSD axes with key recommendations is presented in Fig. 5 with estimated level of effort/perceived necessity shown at the top of each panel. The scales range from 1 to 5, representing least effort/necessity to most effort/necessity, based on existing capabilities and priority ranking from the survey results. While there is nominal difference in the perceived necessity, there is a delineation in levels of effort ranging from Sample Size to Simulation Realism. The estimations for effort/necessity under each axis are provided with the intention to enable software engineers to make informed decisions about HSD support.

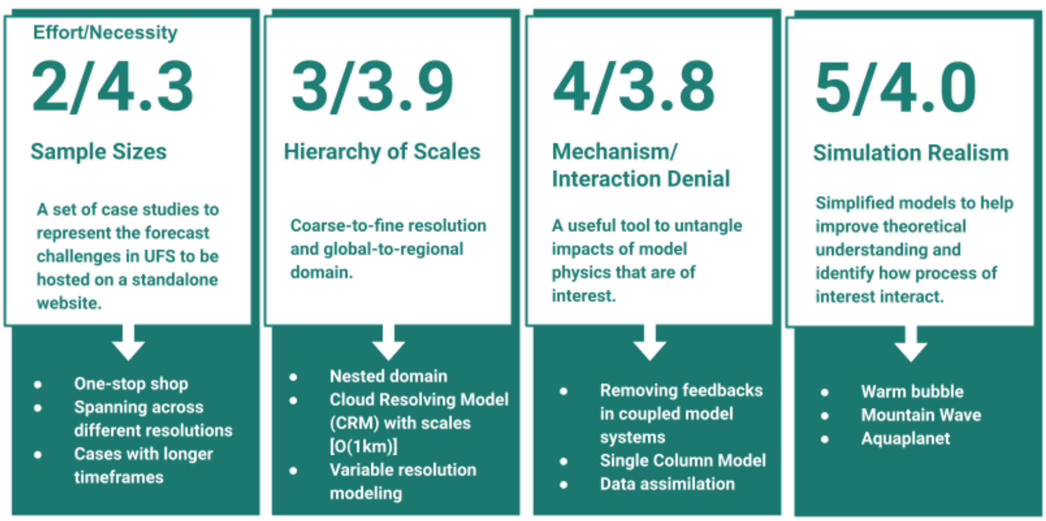

Figure 5. Axis for UFS HSD with proposed recommendations. The numbers listed above each axis represent the effort/necessity, correspondingly. The efforts are given subjectively based on the existing capabilities from the community or developers. The necessities are evaluated based on the priority rankings from the survey results.

1) **Sample Sizes**: A suite of case studies relevant for studying model shortcomings should be provided to the community for running UFS Applications. These should be maintained on a separate website and include detailed documentation and description of the case, initial and lateral boundary condition files or scripts for retrieving them or forcing files, and example scripts for visualization.

   This capability could be built upon the existing UFS Case Studies Platform (https://ufs-case-studies.readthedocs.io/en/develop/) hosted on GitHub and Read the Docs website. This platform was initially developed under a Hurricane Supplement Project. It has the infrastructure to provide resources for representative case studies for UFS required by the community. The current platform covers three applications including the UFS MRW App, global workflow, and SRW App with 10 cases. Case studies should be expanded to represent all of the other applications as the UFS evolves, such as hurricane, coastal, marine and cryosphere, land, space weather, air quality and biogeocheimcal/ecosystem applications. These cases should be updated synchronously with public releases of UFS applications, with the possibility of adding more cases for each UFS forecast challenge, and "retiring" cases that might not be relevant any longer. It is recommended that the cases include convection-resolving configurations (dx=1-4km), cases with forecast ranges spanning coarser resolutions and longer lead times, and cases across the multiple HSD steps/levels (global/regional/idealized/SCM).

2) **Hierarchy of Scales**: Moving nest capability is already supported in the UFS hurricane application with nested and stretched grids capabilities, and thus are available within FV3. The implementation of the capabilities of nested domain and variable resolution modeling in the publically-released versions of UFS applications could be accomplished by coordinating with the FV3 team and integrating those existing capabilities within the MRW, SRW and Hurricane apps.

   The current UFS limited area domain application, SRW App, only supports grid resolutions up to 3km. For sub-3km simulation capabilities, input data, such as high-resolution topography is needed. A subgrid turbulence scheme needs to be integrated within FV3 to obtain the LES capability for UFS, or preferably adapting an existing moist-turbulent boundary-layer scheme to have horizontal mixing in an scale-adaptive way. Corresponding efforts have been initiated by the FV3 team; EPIC should engage with the FV3 team for coordinated efforts in developing these capabilities.

3) **Mechanism/Interaction Denial**: The CCPP SCM is an existing capability that supports mechanism/interaction denial in UFS. Continued development of CCPP SCM is needed, such as the capability to include more supported physics suites, and to expand the current SCM catalog by including cases for different atmospheric conditions and field campaigns. A SCM automation workflow including the ability to derive forcing datasets from UFS high resolution runs and conduct SCM runs is currently being developed through the DTC's contributions to the Model Uncertainty Model Intercomparison Project (MUMIP), an international effort under the World Weather Research Programme. EPIC can leverage this ongoing development of SCM automation workflow as part of the full-spectrum hierarchical workflow.

   Existing capabilities for removing feedbacks in coupled systems within UFS are limited at this time. One possibility to achieve this capability is to use the Community Data Models for Earth Predictive Systems (CDEPS), which is Earth System Modeling Framework (ESMF) based and enables capabilities to selectively eliminate feedbacks in a coupled model system. CDEPS basically performs the tasks of reading external data files, making modifications to the data, and subsequently returning the data to the Community Mediator for Earth Prediction Systems (CMEPS), which is a National Unified Operational Prediction Capability (NUOPC)-compliant mediator that uses ESMF to couple earth grid components. The data sent back to the mediator is similar to the fields that would be sent back by an active model component in a coupled model. The selective feedbacks in coupled systems are thus removed by replacing the fully active component with the data component.

4) **Simulation Realism**: The aquaplanet capability is available from the FV3 team. This could serve as the starting point for Earth Prediction Innovation Center (EPIC) to develop simpler models covering other idealized simulation scenarios. The capabilities for idealized simulations can be generalized for UFS to include features such as periodic boundary conditions, 2D simulation, idealized initializations using sounding or specified perturbations, a dry atmosphere without microphysics, *etc*.

   The ability to run idealized simulations through a highly configurable and well-documented framework is desired. Currently, those capabilities are not available in public released versions of UFS applications, and it is not straightforward for the community to configure the UFS Weather Model to run in an idealized environment. FV3 capabilities do exist for running idealized cases, such as those developed and made publicly available by NOAA-GFDL, which could be added to the UFS HSD framework, with minimal effort. These reside in the SHiELD_build Github repository as a part of their automated CI system, along with python Jupyter notebooks detailing the available cases.

B. Other Aspects Highlighted in the Survey

1) **Expand initial condition datasets supported by UFS applications:** Currently, UFS applications allow for a limited selection of datasets for initial and boundary layer conditions. In order for NOAA to participate in international collaborative projects that use common initializations for model comparisons, it is desirable to expand the current capabilities for UFS to read additional datasets such as European Centre for Medium-Range Weather Forecasts (ECMWF), ECMWF Reanalysis v5 (ERA5), Unified Model (UM; UK Met Office), Icosahedral Nonhydrostatic Model (ICON; German Weather Service, DWD), and climate model output. Priorities should be given to ECMWF or ERA5, which are widely used in the international community. While functionality does exist for reading in dynamics from ERA5 in FV3, full capability, including reading in surface temperature should be accomplished in the *chgres* utility (https://github.com/ufs-community/UFS_UTILS). Data management and accessibility should be improved to allow for consistent and reliable data for scientific testing and model comparison.

2) **Develop a hierarchical workflow**: A hierarchical workflow that adequately supports HSD should be developed in the long run. Multiple configurations within a hierarchy should be available within this single modeling framework. These configurations could be controlled by a simple on/off switch in the configuration or namelist parameter. For example the CESM2.0 can be configured to run several dynamical core and aquaplanet simulations. This hierarchical workflow should include acquiring the input dataset, supporting different HSD applications, post-processing the output dataset, and applying metrics from one level of testing to the next level of testing by leveraging the vast

statistics and diagnostics supported by METplus. The community defined metrics can be found on the 2021 DTC UFS Metrics Workshop page (DTC, 2021). This workflow could be developed using a unified workflow across UFS applications.

3) **Development of a testing infrastructure:** To facilitate physics innovations and model development in the UFS, a common testing infrastructure for all the UFS components and applications that can accelerate the transfer of innovations and model improvements into operations is necessary. A CTest (i.e., the CMake test driver program)-based hierarchical testing infrastructure for UFS is currently under development at the EPIC with the potential to serve as a platform for component tuning, unit testing, end-to-end testing (E2E), model sensitivity analysis (Johnson et al., 2023), sanity checks of model subsystems, and multiscale case studies.

   This prototype testing infrastructure has been implemented into the recent release of the UFS Land Data Assimilation (DA) System V1.0.0 (UFS Development Team, 2023) as a demonstration. The adoption of the CTest permitted the deployed testing framework to cover unit testing and baseline checks for each sub-component within the system. Additionally, this prototype HSD has been integrated with continuous integration and development pipelines using a containerized approach to support code management for continuous release practices.

V. Summary

The vision and ongoing development for the testing framework by the EPIC team, in collaboration with the UFS community, is (1) to accelerate the research and development capabilities of the UFS by providing a standardized, shared toolkit to conduct systematic experiments for scientific and engineering testing and evaluation; and (2) to integrate the testing infrastructure with the Continuous Integration (CI) framework, to facilitate the day-to-day development work for the UFS, which eventually will lead to the continuous release and deployment of the UFS Weather Model and its applications. While much of this paper is focused on the Atmospheric Component of a coupled ESM framework, it is recommended that the HSD framework be built out in a similar manner for each component of the UFS system.

In summary, the use of the HSD approach for model improvement includes testing of individual ESM elements (e.g. physics schemes), then progressively connecting elements with increased coupling between ESM components at the different HSD steps, all the way up to a complex fully-coupled ESM. The time this takes is an important measure of success, and can greatly accelerate the R2O process, as well as maximize compute efficiency.

References

Adcroft, A., W. Anderson, V. Balaji, C. Blanton, M. Bushuk, C. O. Dufour, et al. 2019: Model description and simulation features. *Journal of Advances in Modeling Earth Systems*, **11**, 3167– 3211. https://doi.org/10.1029/2019MS001726

Albrecht, B. A., C. S. Bretherton, D. Johnson, W. H. Scubert, and A. S. Frisch, 1995: The Atlantic Stratocumulus Transition Experiment—ASTEX. *Bulletin of the American Meteorological Society*, **76,** 889-904, 10.1175/1520-0477(1995)076<0889:Taste>2.0.Co;2.

Alexander, G. A., H. A. Holmes, X. Sun, D. Caputi, I. C. Faloona, and H. J. Oldroyd, 2022: Simulating land-atmosphere coupling in the Central Valley, California: Investigating soil moisture impacts on boundary layer properties. *Agricultural and Forest Meteorology*, **317,** 108898, https://doi.org/10.1016/j.agrformet.2022.108898.

Archer-Nicholls, S., D. Lowe, D. M. Schultz, and G. McFiggans, 2016: Aerosol–radiation–cloud interactions in a regional coupled model: the effects of convective parameterisation and resolution. *Atmos. Chem. Phys.*, **16,** 5573-5594, 10.5194/acp-16-5573-2016.

Becker, T., B. Stevens, and C. Hohenegger, 2017: Imprint of the convective parameterization and sea-surface temperature on large-scale convective self-aggregation. *Journal of Advances in Modeling Earth Systems*, **9,** 1488-1505, https://doi.org/10.1002/2016MS000865.

Betts, A. K., and M. J. Miller, 1986: A new convective adjustment scheme. Part II: Single column tests using GATE wave, BOMEX, ATEX and arctic air-mass data sets. *Quarterly Journal of the Royal Meteorological Society*, **112,** 693-709, https://doi.org/10.1002/qj.49711247308.

Blackburn, M., and Coauthors, 2013: The aqua-planet experiment (APE): Control SST simulation. *Journal of the Meteorological Society of Japan. Ser. II*, **91,** 17-56, https://doi.org/10.2151/jmsj.2013-A02.

Bony, S., and Coauthors, 2013: Monograph on Climate Science for Serving Society: Research, Modelling and Prediction Priorities.

Brown, B. B., Jensen, T., Gotway, J. H., Bullock, R., Gilleland, E., Fowler, T. L., Newman, K. M., Adriaansen, D. R., Blank, L., Burek, T., Harrold, M., Hertneky, T., Kalb, C., Kucera, P. A., Nance, L., Opatz, J., Vigh, J. L., & Wolff, J. K., 2021: The Model Evaluation Tools (MET): More than a Decade of Community-Supported Forecast Verification. *Bulletin of the American Meteorological Society*, *102*(4), E782–E807. https://doi.org/10.1175/bams-d-19-0093.1

Costantino, L., and P. Heinrich, 2014: Tropical deep convection and density current signature in surface pressure: comparison between WRF model simulations and infrasound measurements. Atmospheric Chemistry and Physics, 14, 3113-3132, https://doi.org/10.5194/acp-14-3113-2014.

Dal Gesso, S., Van der Dussen, J. J., Siebesma, A. P., De Roode, S. R., Boutle, I. A., Kamae, Y., et al., 2015. A single-column model intercomparison on the stratocumulus representation in present-day and future climate. *Journal of Advances in Modeling Earth Systems*, 7(2), 617–647, https://doi.org/10.1002/2014ms000377.

Developmental Testbed Center (DTC), 2021: 2021 DTC UFS Evaluation Metrics Workshop. https://dtcenter.org/events/2021/2021-dtc-ufs-evaluation-metrics-workshop

Ek, M., and coauthors, cited 2023: Hierarchical System Development for the UFS. [Available online at https://ufscommunity.org/articles/hierarchical-system-development-for-the-ufs/.]

Firl, G., D. Swales, L. Carson, L. Bernardet, D. Heinzeller, and M. Harrold, 2022: Common Community Physics Package Single Column Model v6.0.0 User and Technical Guide, 40 pp, https://dtcenter.org/sites/default/files/paragraph//scm-ccpp-guide-v6-0-0.pdf.

Fuhrer, O., and C. Schär, 2005: Embedded cellular convection in moist flow past topography. Journal of the Atmospheric Sciences, 62, 2810-2828, https://doi.org/10.1175/jas3512.1

Grabowski, W. W., 2014: Extracting Microphysical Impacts in Large-Eddy Simulations of Shallow Convection. *Journal of the Atmospheric Sciences*, **71,** 4493-4499, 10.1175/jas-d-14-0231.1.

——, 2015: Untangling Microphysical Impacts on Deep Convection Applying a Novel Modeling Methodology. *Journal of the Atmospheric Sciences*, **72,** 2446-2464, 10.1175/jas-d-14-0307.1.

——, 2019: Separating physical impacts from natural variability using piggybacking technique. *Adv. Geosci.*, **49,** 105-111, 10.5194/adgeo-49-105-2019.

——, and H. Morrison, 2016: Untangling Microphysical Impacts on Deep Convection Applying a Novel Modeling Methodology. Part II: Double-Moment Microphysics. *Journal of the Atmospheric Sciences*, **73,** 3749-3770, 10.1175/jas-d-15-0367.1.

Hagos, S., R. Leung, S. A. Rauscher, and T. Ringler, 2013: Error Characteristics of Two Grid Refinement Approaches in Aquaplanet Simulations: MPAS-A and WRF. *Monthly Weather Review*, **141,** 3022-3036, 10.1175/mwr-d-12-00338.1.

Heath, N. K., H. E. Fuelberg, S. Tanelli, F. J. Turk, R. P. Lawson, S. Woods, and S. Freeman, 2017: WRF nested large-eddy simulations of deep convection during SEAC4RS. *Journal of Geophysical Research: Atmospheres*, **122,** 3953-3974, https://doi.org/10.1002/2016JD025465.

Hoffman, R. N., V. K. Kumar, S.-A. Boukabara, K. Ide, F. Yang, and R. Atlas, 2018: Progress in forecast skill at three leading global operational NWP centers during 2015–17 as seen in Summary Assessment Metrics (SAMs). *Weather and Forecasting*, **33**, 1661–1679. https://doi.org/10.1175/WAF-D-18-0117.1

Huang, X., A. M. Rhoades, P. A. Ullrich, and C. M. Zarzycki, 2016: An evaluation of the variable-resolution CESM for modeling California's climate. *Journal of Advances in Modeling Earth Systems*, **8,** 345-369, https://doi.org/10.1002/2015MS000559.

Jakob, C., 2010: Accelerating Progress in Global Atmospheric Model Development through Improved Parameterizations: Challenges, Opportunities, and Strategies. *Bulletin of the American Meteorological Society*, **91,** 869-876, 10.1175/2009bams2898.1.

——, 2014: Going back to basics. *Nature Climate Change*, **4,** 1042-1045, 10.1038/nclimate2445.

James, E. P., C. R. Alexander, D. C Dowell, S. S.Weygandt, S. G. Benjamin, G. S Manikin, et al., 2022: The High-Resolution Rapid Refresh (HRRR): An hourly updating convection-allowing forecast model. Part II: Forecast performance. *Weather and Forecasting*, **37(8)**, 1397– 1417. https://doi.org/10.1175/WAF-D-21-0130.1

Jeevanjee, N., P. Hassanzadeh, S. Hill, and A. Sheshadri, 2017: A perspective on climate model hierarchies. *Journal of Advances in Modeling Earth Systems*, **9,** 1760-1771, https://doi.org/10.1002/2017MS001038.

Johnson, A., X. Wang, B. T. Blake, E. Rogers, Y. Wang, J. R. Carley, et al., 2023: Verification and model configuration sensitivity of simulated ABI radiance forecasts with the FV3-LAM model. *Earth and Space Science*, **10**, e2022EA002651. https://doi.org/10.1029/2022EA002651

Khairoutdinov, M. F., and D. A. Randall, 2003: Cloud Resolving Modeling of the ARM Summer 1997 IOP: Model Formulation, Results, Uncertainties, and Sensitivities. *Journal of the Atmospheric Sciences*, **60,** 607-625, https://doi.org/10.1029/2011ms000081

Kim, D., A. H. Sobel, and I. S. Kang, 2011: A mechanism denial study on the Madden‐Julian Oscillation. *Journal of Advances in Modeling Earth Systems*, **3**, https://doi.org/10.1029/2011ms000081

Kirshbaum, D. J., and D. R. Durran, 2005: Observations and modeling of banded orographic convection. Journal of the Atmospheric Sciences, 62, 1463-1479, https://doi.org/10.1175/jas3417.1

Landu, K., L. R. Leung, S. Hagos, V. Vinoj, S. A. Rauscher, T. Ringler, and M. Taylor, 2014: The Dependence of ITCZ Structure on Model Resolution and Dynamical Core in Aquaplanet Simulations. *Journal of Climate*, **27,** 2375-2385, 10.1175/jcli-d-13-00269.1.

Liu, Y., and Coauthors, 2011: Simultaneous nested modeling from the synoptic scale to the LES scale for wind energy applications. *Journal of Wind Engineering and Industrial Aerodynamics*, **99,** 308-319, https://doi.org/10.1016/j.jweia.2011.01.013.

Maher, P., and Coauthors, 2019: Model Hierarchies for Understanding Atmospheric Circulation. *Reviews of Geophysics*, **57,** 250-280, https://doi.org/10.1029/2018RG000607.

Mayr, G., and A. Gohm, 2000: 2D airflow over a double bell-shaped mountain. Meteorology and Atmospheric Physics, 72, 13-27, https://doi.org/10.1007/s007030050002

Medeiros, B., B. Stevens, and S. Bony, 2015: Using aquaplanets to understand the robust responses of comprehensive climate models to forcing. *Climate Dynamics*, **44,** 1957-1977, 10.1007/s00382-014-2138-0.

Miglietta, M., and R. Rotunno, 2005: Simulations of moist nearly neutral flow over a ridge. Journal of the Atmospheric Sciences, 62, 1410-1427, http://dx.doi.org/10.1175/JAS3410.1

Moeng, C.-H., J. Dudhia, J. Klemp, and P. Sullivan, 2007: Examining Two-Way Grid Nesting for Large Eddy Simulation of the PBL Using the WRF Model. *Monthly Weather Review*, **135,** 2295-2311, 10.1175/mwr3406.1.

Mohan, M., and A. P. Sati, 2016: WRF model performance analysis for a suite of simulation design. *Atmospheric Research*, **169,** 280-291, https://doi.org/10.1016/j.atmosres.2015.10.013.

Morrison, H., J. M. Peters, A. C. Varble, W. M. Hannah, and S. E. Giangrande, 2020: Thermal chains and entrainment in cumulus updrafts. Part I: Theoretical description. Journal of the Atmospheric Sciences, 77, 3637-3660, https://doi-org.colorado.idm.oclc.org/10.1175/JAS-D-19-0243.1

Mulholland, J., J. Peters, and H. Morrison, 2021: How does LCL height influence deep convective updraft width? Geophysical Research Letters, 48, e2021GL093316, https://doi-org.colorado.idm.oclc.org/10.1029/2021GL093316,

Muller, C., and S. Bony, 2015: What favors convective aggregation and why? *Geophysical Research Letters*, **42,** 5626-5634, https://doi.org/10.1002/2015GL064260.

Neggers, R. A. J., Ackerman, A. S., Angevine, W. M., Basile, E., Beau, I., Blossey, P. N., Boutle, I. A., de Bruijn, C., Cheng, A., van der Dussen, J., Fletcher, J., dal Gesso, S., Jam, A., Kawai, H., Kumar, S., Larson, V. E., Lefebvre, M.-P., Lock, A. P., Meyer, N. R., de Roode, S. R., de Rooy, W., Sandu, I., Xiao, H., and Xu, K.-M., 2017: Single-column model simulations of subtropical marine boundary-layer cloud transitions under weakening inversions, *J. Adv. Model. Earth Syst.*, 9, 2385–2412, https://doi.org/10.1002/2017MS001064.

Pendergrass, A. G., K. A. Reed, and B. Medeiros, 2016: The link between extreme precipitation and convective organization in a warming climate: Global radiative-convective equilibrium simulations. *Geophysical Research Letters*, **43,** 11,445-411,452, https://doi.org/10.1002/2016GL071285.

Randall, D. A., K.-m. Xu, R. J. C. Somerville, and S. Iacobellis, 1996: Single-Column Models and Cloud Ensemble Models as Links between Observations and Climate Models. *Journal of Climate*, **9,** 1683-1697, 10.1175/1520-0442(1996)009<1683:Scmace>2.0.Co;2.

Rhoades, A. M., P. A. Ullrich, and C. M. Zarzycki, 2018: Projecting 21st century snowpack trends in western USA mountains using variable-resolution CESM. *Climate Dynamics*, **50,** 261-288, 10.1007/s00382-017-3606-0.

Rios-Berrios, R., B. Medeiros, and G. H. Bryan, 2020: Mean Climate and Tropical Rainfall Variability in Aquaplanet Simulations Using the Model for Prediction Across Scales-Atmosphere. *Journal of Advances in Modeling Earth Systems*, **12,** e2020MS002102, https://doi.org/10.1029/2020MS002102.

Sarkadi, N., and Coauthors, 2022: Microphysical Piggybacking in the Weather Research and Forecasting Model. *Journal of Advances in Modeling Earth Systems*, **14,** e2021MS002890, https://doi.org/10.1029/2021MS002890.

Sims, J., D. Koch, H. Tolman, and D. Achuthavarier, 2021: UFS Forecasters Workshops 2020 & 2021, [Available online at https://www.weather.gov/media/sti/Final%20Consolidated%20Forecasters%20requests%20April%202021.pdf.]

Sun, X., H. A. Holmes, and H. Xiao, 2020: Surface Turbulent Fluxes during Persistent Cold Air Pool Events in the Salt Lake Valley, Utah. Part II: Simulations. *Journal of Applied Meteorology and Climatology*, **59,** 1029-1050, 10.1175/jamc-d-19-0250.1.

Sun, X., D. Heinzeller, L. Bernardet, L. Pan, and J. M. Brown, 2021: Case Studies That Exemplify Known Biases of the Unified Forecast System (UFS) Weather Model. *101st American Meteorological Society Annual Meeting*, AMS.

Talbot, C., E. Bou-Zeid, and J. Smith, 2012: Nested Mesoscale Large-Eddy Simulations with WRF: Performance in Real Test Cases. *Journal of Hydrometeorology*, **13,** 1421-1441, 10.1175/jhm-d-11-048.1.

Teng, Y., and S. Flampouris, 2023: The Hierarchical Testing Framework for the Unified Forecast System (UFS). *103st American Meteorological Society Annual Meeting*, AMS.

The WAVEWATCH III Development Group (WW3DG), 2019: User manual and system documentation of WAVEWATCH III version 6.07. Tech. Note 333, NOAA/NWS/NCEP/MMAB, 465 pp.

UFS Development Team, 2022. Unified Forecast System (UFS) Short-Range Weather (SRW) Application (Version v2.1.0). Zenodo. https://doi.org/10.5281/zenodo.7277602.

UFS Development Team, 2023. Unified Forecast System (UFS) Land Data Assimilation System (Version v1.1.0). Zenodo. https://doi.org/10.5281/zenodo.7675721.

Ullrich, P., 2014: SQuadGen: Spherical quadrilateral grid generator. *University of California, Davis, Climate and Global Change Group software. [Available online at http://climate. ucdavis. edu/squadgen. php.]*,

Wang, M., and P. Ullrich, 2018: Marine Air Penetration in California's Central Valley: Meteorological Drivers and the Impact of Climate Change. *Journal of Applied Meteorology and Climatology*, **57,** 137-154, 10.1175/jamc-d-17-0089.1.

Wing, A. A., K. A. Reed, M. Satoh, B. Stevens, S. Bony, and T. Ohno, 2018: Radiative–convective equilibrium model intercomparison project. *Geoscientific Model Development*, **11,** 793-813, https://doi.org/10.5194/gmd-11-793-2018.

Xie, S., R. T. Cederwall, M. Zhang, and J. J. Yio, 2003: Comparison of SCM and CSRM forcing data derived from the ECMWF model and from objective analysis at the ARM SGP site. *J. Geophys. Res.*, **108**, 4499, doi:10.1029/2003JD003541.

Zarzycki, C. M., and C. Jablonowski, 2014: A multidecadal simulation of Atlantic tropical cyclones using a variable-resolution global atmospheric general circulation model. *Journal of Advances in Modeling Earth Systems*, **6,** 805-828, https://doi.org/10.1002/2014MS000352.

——, 2015: Experimental Tropical Cyclone Forecasts Using a Variable-Resolution Global Model. *Monthly Weather Review*, **143,** 4012-4037, 10.1175/mwr-d-15-0159.1.

Zheng, X., Klein, S. A., Ghate, V. P., Santos, S., McGibbon, J., Caldwell, P., Bogenschutz, P., Lin, W., & Cadeddu, M. P., 2020. Assessment of Precipitating Marine Stratocumulus Clouds in the E3SMv1 Atmosphere Model: A Case Study from the ARM MAGIC Field Campaign, *Monthly Weather Review*, *148*(8), 3341-3359.

Zheng, Y., H. Zhang, and Z. Li, 2021: Role of surface latent heat flux in shallow cloud transitions: A mechanism-denial LES study. *Journal of the Atmospheric Sciences*, **78,** 2709-2723, https://doi.org/10.1175/JAS-D-20-0381.1.

Appendix A

HSD Survey Results

The Hierarchical System Development (HSD) Survey for Unified Forecast System (UFS) was conducted by the Developmental Testbed Center (DTC) in collaboration with the Earth Prediction Innovation Center (EPIC). The survey aimed to gather insights and feedback from the community to help shape the future development of HSD for UFS (Ek et al. 2019). It was designed to collect information on the respondents' backgrounds and experience with using an HSD approach, opinions on gaps in the existing HSD system for UFS, rank the significance of potential HSD tools, and collect other recommendations. The online survey included a total of 13 questions, with a mix of multiple-choice and short answers. It was sent to the broad community via several email lists and newsletters in April 2023. A total of 55 responses were received by May 2023. The survey responses were analyzed using a combination of quantitative and manual methods. Results are summarized as follows.

A. Respondents' background and experience with HSD:
   - **The diverse backgrounds of the respondent pool ensure that the survey results encompass a wide range of perspectives and insights**. The majority of participants (61.8%) identified themselves as Research Scientists/Postdoctoral Researchers. Other participants indicated that they hold the role of Professor/Faculty Member (11%), Software Engineer (11%), Manager (7.3%), and Model Developer (5.4%). The remaining percentage identified themselves under the 'Others' category.

- **55% of respondents have experience conducting hierarchical testing at multiple levels of model complexity while developing a new capability or innovation**. 45% of respondents conducted testing only on one level of model complexity.
- **An equal percentage (14.3%) of respondents chose infrastructures of ctest, PyTest, and Unittest, respectively, when designing tests required for model development.** Other respondents (14.5%) stated that they use custom-made or existing verification packages (*e.g.*, Verification Statistics DataBase, VSDB) to develop tests for model development. Note that the remaining large percentage of respondents indicated that they were not familiar with the options provided. This is not unexpected, since the majority of respondents are not software engineers.
- **Sample sizes, Hierarchy of scales, and Mechanism/Interaction denial emerged as equally prevalent HSD tools chosen by the subset of respondents who are currently using HSD-related capabilities or tools (26 out of 55).**

B. Opinions on Gaps in the UFS HSD research-to-operations:
There is a need to improve the hierarchical testing capabilities for UFS components, as in Earth system Earth system components, including atmosphere, land, ocean, *etc*. A total of 53% of respondents expressed partial agreement, uncertainty, and disagreement on the readiness of the UFS components' capabilities to support hierarchical testing. 38.2% of respondents agreed that the UFS components they use support hierarchical testing. The remaining percentage of respondents are not in the UFS community.

The opinions of the respondents about gaps in the UFS HSD R2O are summarized below:

- Access to repositories is not easy for the academic/university community. It is currently too difficult for anyone outside of the immediate scope of UFS (mostly federal labs) to contribute to UFS development without having to jump through a lot of hoops.
- Data availability in general for running operational workflow is problematic outside NOAA.
- Capabilities of high-resolution limited area domain setup, nested domain, and idealized simulations are not available in the publicly-released versions of UFS applications, or they exist in the UFS Weather Model but a lot of setup is required to run them.
- More case studies across the multiple HSD levels would be useful, i.e. global/regional/idealized/SCM.
- Workflows/Frameworks that adequately support UFS HSD are not currently available. This includes getting an input dataset, supporting different HSD applications, post-processing an output dataset, and the criteria (such as statistical benchmarks) for moving from one level of testing to the next level of testing.
- Communications between researchers and operational users about how to transform the outcomes from R2O are not well established. This is worth further discussion.

C. Priority Rankings on Potential HSD Tools:

The survey team provided seven options for potential HSD tools corresponding to the proposed UFS HSD axes. The priority rankings were calculated using a weighted average, Each option was given a scale of 1 to 5, representing the range from "not important at all" to "extremely important". For some options, additional sub-options were given to prioritize specific HSD tools in more detail. For the sub-options, the highest weight corresponds to the number of options available. For example, if 6 options of specific HSD tools are provided, the highest weight will be 6.

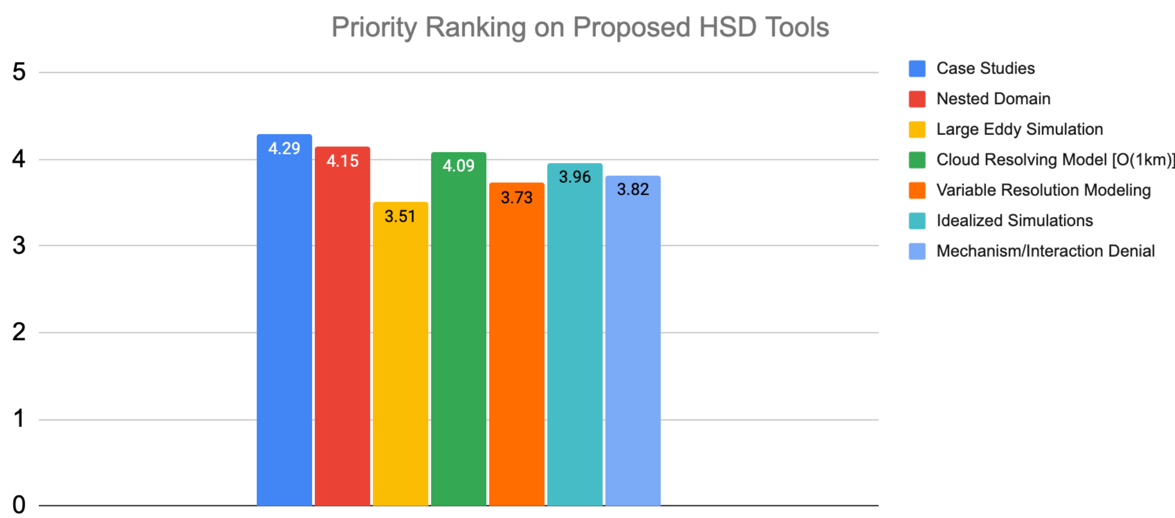

Fig. 1

The average ranking for the 7 proposed HSD tools is 3.94. Providing a set of case studies has the highest ranking (4.29), followed by Nested Domain (4.15) and Cloud Resolving Model capability [O(1km)] (4.09). The LES capability was ranked as the lowest priority (3.51) in this survey. Priorities for more specific HSD tools under some categories in Fig. 1 are summarized below.

- **The platform for hosting Case Studies**: 60% of the respondents stated that they prefer a set of case studies to represent the UFS forecast challenges to be hosted on a standalone website. The platform should function as a one-stop shop providing initial datasets and descriptions of the cases. It is recommended that the cases include convection-resolving cases (dx=1-4km), as well as cases with forecast ranges spanning coarser resolutions and longer timeframes.
- **Configurations of nested domain**: Two-way nesting is identified by 60% of respondents as a configuration that is necessary for their research or application. One-way nesting has a close percentage (52.5%) to two-way nesting, followed by moving nests (43.6%). Multiple nests and nesting on ocean grids are also recommended by the respondents.

Note that this was a multiple answer question, which allows the total percentage to surpass 100%.

- **Kilometer and Subkilometer-scale applications**: The two options provided, urban scale and fire weather, received 21.8% and 29.1%, respectively. Various responses were received for this question under the ‘Other’’ option. Storm predictions were mentioned several times, as well as warn-on-forecast-like capabilities. It is not applicable to categorize these answers due to their variability. Interested readers are referred to Question 9.2 in Appendix A.
- **Idealized simulations**: The top three ranked options for idealized simulations are warm bubble, mountain wave, and aquaplanet. Idealized cases of the diurnal cycle over land, stable boundary layer (weakly and strongly stable), marine stratocumulus, urban heat island, *etc*., are mentioned in the ‘Other’ option. It is suggested that the capabilities for idealized simulations can be generalized for the UFS to include features such as periodic boundary conditions, 2D simulation, idealized initializations using sounding or specified perturbations, a dry atmosphere without microphysics, *etc*.

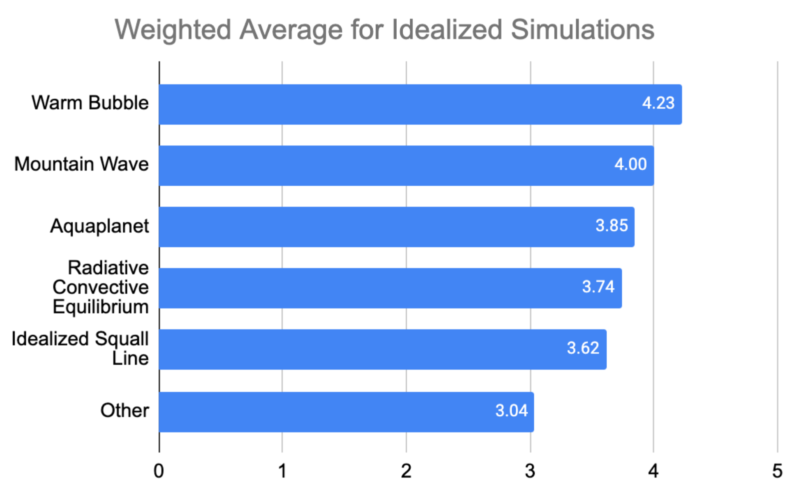

Fig. 2

- **Mechanism/Interaction Denial**: The top three ranked options for Mechanism/Interaction Denial are removing feedback in coupled model systems, single column models, and data assimilation.

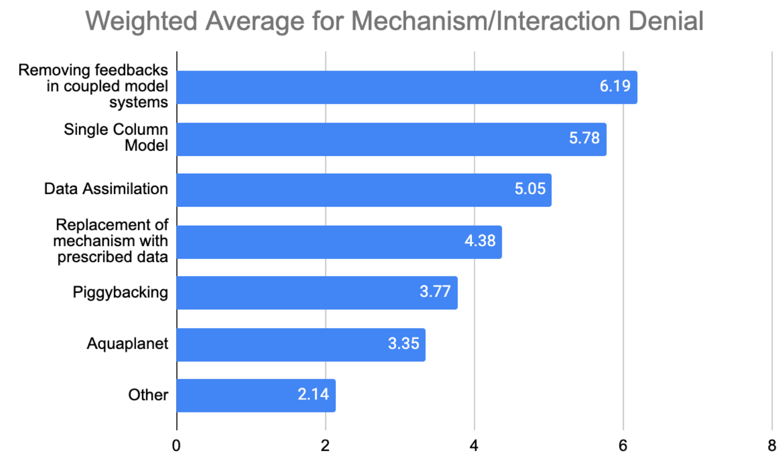

Fig. 3

D. Other Aspects of HSD:

There were a number of comments related to the fact that the survey did not address questions specific to HSD for other model components other than the atmosphere, where complex fully-coupled ESMs include (or should include) components for atmosphere/chemistry/aerosols, as well as for ocean/waves/sea-ice, land-hydrology/snow/land-ice, and biogeochemical cycles/ecosystems, a subset of which (i.e. atmosphere+land and specified ocean conditions) has traditionally addressed NWP needs. In fact, HSD capabilities **do** exist for other model components, such as the MOM6 community ocean model, land models (e.g. Noah) with and without coupling to the atmosphere (for different UFS applications and model resolutions), and Wave Watch III. The HSD tools/capabilities should be applicable and connected across the various components of the UFS, and across a wide array of regimes (tropics/mid-latitudes/poles, terrestrial/maritime, global/regional, etc).

The community expressed the need to easily configure and test relevant case studies by means of a one-stop-shop for different UFS capabilities, tools, and software, making them readily available and pre-installed or easy to install. They also emphasized the need for a simple, highly configurable, and well-documented testing workflow, especially for idealized case studies. Within the workflow for testing implementations for R2O, establishing methodologies and criteria across the different levels of HSD testing would be useful.

The community desires UFS HSD capabilities that are well-documented, easy to use, and accessible at all levels of the UFS and general Earth system research and modeling community. The components should be easily ported to non-operational platforms.

The community highlighted the importance of being able to initialize the model with a larger variety of datasets (e.g. ERA5/ECMWF/UM/ICON/etc). There were also comments based on data management and availability, where users wish to have access to reliable and consistent data for scientific testing and model comparison.