Response to RC1

This manuscript describes a unique model configuration of a global model run with regional refinement down to 100m. It is generally well written and appropriate for GMD. The work does a lot of detailed discussion of the software engineering and configuration of the model at the front, and then a simulation performance description at the back. I think a lot of the software discussion could be usefully made a supplement or appendix to improve readability of the manuscript and make it a good description of model simulations. This should be publishable in GMD with some minor revisions as I note below.

Thank you for your careful reading and insightful suggestions, as well as for your recognition of the significance of our work! In response to your comments, we have revised the figures and the main text. Five figures have been updated, four figures and two tables have been added in the main text.

General Points:

1. I appreciate the step by step description in the first sections of how the model is configured, but I think this is too detailed for the main text and could be put in an appendix.

Thank you for your valuable feedback! As the primary goal of this study is to provide a first proof of concept and document the associated challenges, rather than to pursue scaled-up deployment or a comprehensive evaluation at this stage, the methodology forms the core of our contribution. The key objective of this work is to document the process in full detail to facilitate reproducibility for other interested users. Recognizing that many readers may not be interested in highly technical details, we have reorganized the Methods section. Detailed content has been moved into deeper-level subsections, and we have added two summary sentences at the start of the Methods:

> “The section includes all necessary steps and technical details that made these simulations possible, organized by level of detail. Readers primarily interested in results and discussion may choose to skip the tertiary subsections.”

2. There are some inconsistencies in the figure labeling, numbering and referencing that need to be corrected as noted below.

Thank you for pointing this out! We have revised the text accordingly based on the suggestions provided below.

3. It would be nice to make a few more comments about turbulence across scales. It is hinted at in a few places (with some contradictions that need to be clarified as noted below), but clarification would be good. Specifically (as the text notes): the conventional wisdom says at 100m you need 3D turbulence, but you are using a unified bulk closure with SHOC and this 'seems to work'. That's great, but can you show a figure of turbulence or KE or something that does a bit more to convince a skeptic about it? That would be a great addition.

Thanks for your nice suggestions! Here contains two distinct concepts.

First, the conventional view is that the gray zone – defined as the regime in which the energy-containing turbulence scale is comparable to the scale of the spatial filter – necessitates the use of 3D turbulence parameterizations. This is distinct from the LES regime, which is able to resolve the dominant turbulent structures. The gray zone is typically between the upper bound of LES and the lower bound of mesoscale models. Mesoscale models do not require explicit turbulence resolution because the dominant fluxes arise from eddy ensemble-averaged statistics, which are largely governed by mean gradients in the vertical. On the other hand, within high-resolution LES, the role of tensor diffusion schemes is limited: most turbulent fluxes are resolved, and subfilter-scale processes contribute minimally; so long as energy is positively removed from the resolved scales at the filter level, which a simple scalar eddy diffusivity often achieves effectively (Wyngaard, 2004). If the simulation is run at 800 m horizontal resolution, it is best to use a 3D turbulence version of SHOC. Indeed, even at resolutions as fine as 100 m or finer, developing a 3D SHOC implementation will not be wasted effort, as many studies suggest the filter scale should be interpreted as the effective resolution.

Second, SHOC can be used at LES-scale resolutions without explicit tuning largely due to its built-in scale awareness. That is, as the grid spacing (dx) decreases, the contributions from SGS TKE and SGS fluxes become smaller, while resolved TKE and fluxes become increasingly dominant. In contrast, a turbulence scheme that lacks scale awareness (i.e., one that exerts a similar influence at dx = 100 m and dx = 3 km) would require manual adjustment of parameters or a deliberate weakening of its influence. In the simplest case, this might involve disabling SHOC entirely at dx = 100 m + nudging outside the high-resolution domain.

The first point is conceptual, whether a 3D turbulence implementation would significantly improve simulations at 100 m resolution remains to be verified once such a scheme is implemented into SCREAM. Notably, 100 m sits close to the lower boundary of the gray zone, but the exact transition between modeling regimes can vary in practice, depending on the dominant turbulent length scale which might be case specific. The second point, however, can be demonstrated directly. Although we didn't output resolved TKE, we have SHOC TKE, moisture flux/variance, w variance, and the third moment of w. As shown in the added Figs. 18 and 19 (also shown below), the magnitudes of SGS terms are substantially smaller in BA-100m simulations compared to CA-3km. Note that the shading colorbars differ in each variable to better highlight the spatial structure; the absolute values can be compared through the line plots on the right side of each shading plot.

We have revised the structure of the Results section accordingly. The new analysis of SGS TKE/flux/variance has been added in a separate subsection of "Sub-grid-scale flux" after the in situ evaluation. We also have clarified the concept of the gray zone:

> "\subsection{Sub-grid-scale flux}
>
> Building on the resolution sensitivity study of DP-SCREAM in \citet{Bogenschutz2023}, SCREAM exhibits characteristics of a scale aware model. As horizontal resolution increases, the partitioning between SGS and resolved turbulence diminishes. This scale awareness, inherent to the SHOC parameterization \citep{Bogenschutz_Krueger2013}, enables SCREAM to operate effectively at 100 m resolution without the need for parameter tuning. Specifically, Fig. 9 and Fig. 16 in \citet{Bogenschutz2023} show that in the marine stratocumulus case, maritime shallow cumulus case, and mixed-phase Arctic stratocumulus case, as the horizontal grid spacing Δx decreases, the contribution of SGS moisture flux becomes smaller while the resolved flux becomes increasingly dominant. In the 100 m DP-SCREAM simulations, above 0.2 km, the proportion of SGS moisture flux is negligible, and the resolved flux is nearly unity.
>
> In our simulations, although we did not output high-frequency resolved moisture flux or TKE, Fig.~\ref{SHOCtimeplevStorm2008} and Fig.~\ref{SHOCtimeplevStratocumulus2023} shows substantial reductions in SGS TKE, moisture flux/variance, \emph{w} variance, and the third moment of \emph{w} in the BA-100m simulations compared to the CA-3km simulations. For clarity, the plotted ranges in these figures differ between the two resolutions, with the CA-3km values being much larger.

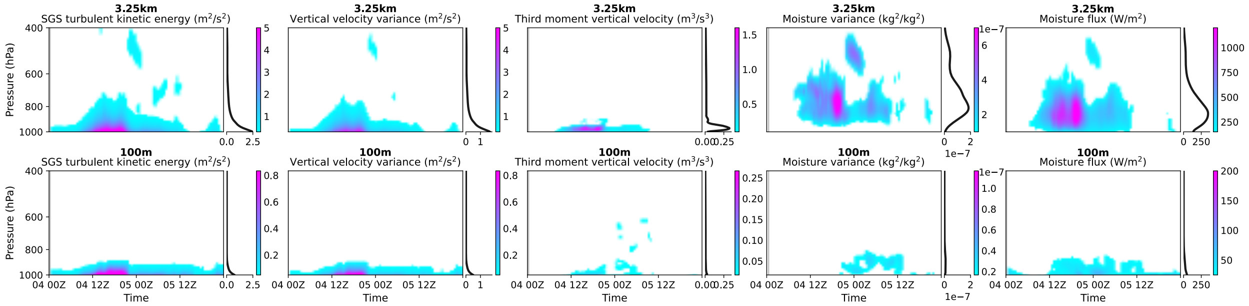


Figure 18. Simulated Sub-grid-scale (SGS) variables in the 3.25 km California RRM (top) and 100 m Bay Area RRM (bottom) during the Storm2008 event. From left to right: TKE, vertical velocity variance, third-moment vertical velocity, moisture variance, and moisture flux. Each panel consists of a time–evolution shading plot on the left and a vertical profile averaged over the simulation period on the right. For clarity, the colorbars differ between the 3.25 km and 100 m simulations, with the former being much larger.

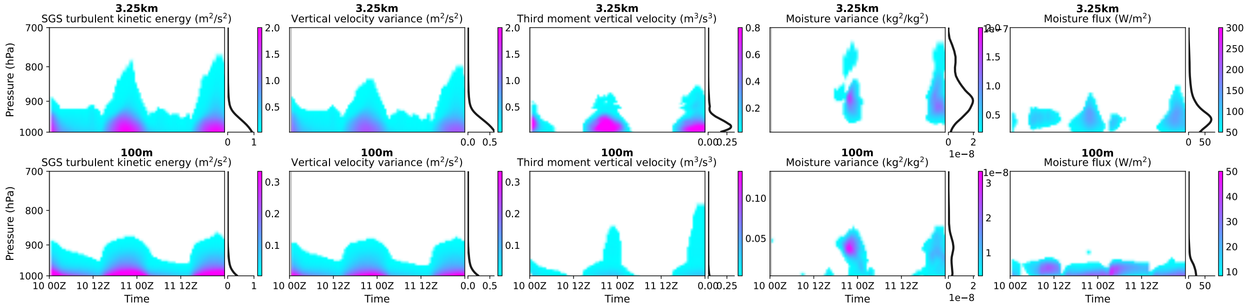


Figure 19. Same as Fig. 18 but for the Stratocumulus2023 event.

At 100 m resolution, simulations are close to, but largely below, the turbulence gray zone, where the grid spacing becomes comparable to the dominant eddy scale. The gray zone typically spans the transition from mesoscale models, which rely on ensemble-averaged vertical fluxes, to LES, where most turbulent motions are explicitly resolved and subgrid closures play only a minor dissipative role \citep{Wyngaard2004}. In this transitional regime, subgrid transport is best treated with 3D turbulence schemes that represent the full stress tensor \citep[e.g.,][]{Wyngaard2004,Chow2019,Honnert2020}, whereas SHOC currently parameterizes only vertical mixing. Thus, at coarser resolutions such as 800 m, a 3D implementation of SHOC would likely be beneficial. At 100 m, near the lower edge of the gray zone, the need for 3D turbulence remains uncertain and case-dependent, pending the implementation and testing of such a scheme in SCREAM.”

Specific Comments:

Page 4, L124: so except where noted, this is SCREAMv0 FORTRAN code? Would be good to be explicit about this.

This entire paragraph refers to the SCREAMv1 C++ codebase (EAMxx), whereas other parts of the paper are based on the SCREAMv0 Fortran version. We have added this clarification to the sentence:

> “When this study began, we only had CPU resources, EAMxx was not yet fully operational, and all simulation results presented in this paper were conducted using the SCREAMv0 Fortran version. “

Page 5, L136: this is not a ‘nest’ however right? Please be clear because ‘level’ might imply there are multiple grids underneath.

Yes, it is not a nested grid, but a seamless single-layer mesh. We have changed “level-1 refinement” to “first-order refinement”, and “second layer of refinement” to “second-order refinement”.

Page 6, L149: same refined grid correct?

Yes, the land and atmospheric physical grids are identical. In CA-3km land, the resolution transitions from global 100 km to 3 km over California. In BA-100m land, it transitions from global 100 km to 800 m over California, and further down to 100 m over the Bay Area.

Page 7, L186: so what is the actual resolution of the topography you can resolve? It seems like you go from 500m —> 800m and then down to 100m? Is there more information or is this smoothed at 800m. I assume there is higher resolution topography than 500m available for Northern California?

Although high-resolution DEMs with resolutions at O(1)m are available for California, there are two limitations. First, since a RRM is a subset of a global high-resolution grid, it is most efficient and sustainable to build topography from a global high-resolution DEM. This ensures that new RRMs can be generated for any region of interest without requiring users to rerun the entire topography workflow each time (a process that would otherwise be redundant, cumbersome, and resource-intensive). Second, the current toolchain requires a global lat-lon DEM (referred to as “terr_latlon_glb”). Even if high-resolution DEMs are available for specific regions, they still need to be stitched and interpolated into a global DEM of matching resolution. As a result, the size of terr_latlon_glb remains extremely large, and unless the toolchain is upgraded, the similar OOM issues will persist. One of the key conclusions of this study is thus the importance of the toolchain upgrades. For topography specifically, this would ideally involve developing MPI-enabled tools and simplifying the mapping algorithms.

Page 8, L203: I appreciate the step by step description, but I think this is too detailed for the main text and could be put in an appendix. The source of the topo data should be identified, then the method can be a page in an appendix.

Thank you again for your valuable suggestions! As noted above, we have reorganized the structure of the Methods section by moving technical details into deeper-level subsections (i.e., under tertiary headings), and we have added a reading guide at the end of the Introduction. The source of the raw DEM data for topography generation has also been described in detail in the Methods section.

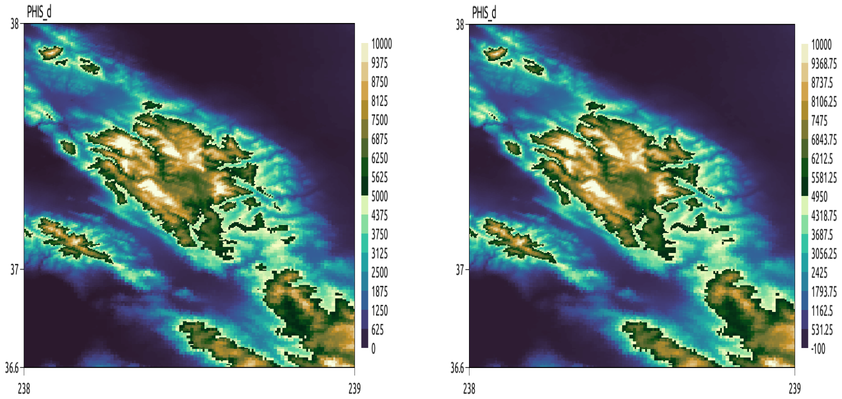
Page 8, L214: unclear what the difference is between you two 100m test runs. How was the simulation without the 'steep topographic gradients' specified? Also, a bit too much info perhaps, and maybe just note that smoothing is important. But doesn't it affect the height of the topography?

This refers to two separate points.

First, the statement "a sensitivity test using a finer 100 m mesh – which did not include the steep topographic gradients present in the RRM configuration examined in this study" refers to an earlier test conducted prior to the current simulations. In that test, we designed a different 100 m mesh with a smaller domain that was shifted eastward, excluding many of the mountainous regions present in the current BA-100m configuration, hence the reference to the absence of "steep topographic gradients". This earlier simulation used the default smoothing level and other default dynamical core settings, and it successfully ran for one hour. This preliminary test served as supporting evidence that the model instability encountered in the current BA-100m run was likely related to the topography, especially due to the inability to generate sufficiently high-resolution RRM topography, which led to overly steep slopes at the mountain grid boundaries.

Second, regarding the two smoothing configurations tested in the current BA-100m simulations: while it is true that smoothing affects the height of the topography, the difference between 6 and 12 iterations of smoothing is not apparent. Notably, the E3SM v2 topography toolchain recommends 12 smoothing passes as standard practice.

The figure below (Fig. R1.1) compares the topography after 6 (left) and 12 (right) smoothing iterations. Visually, the difference is minimal:



R1.1 BA-100m topography generated using 6 (left) and 12 (right) smoothing iterations.

The Methods section has been reorganized, with the technical details moved under tertiary-level headings.

Page 10, L239: so what timestep does an LES model with 100m resolution typically use? This seems VERY short (I'm guessing the LES is more like 1sec). So what does that say about the quality of the dynamics or the utility of this configuration? Seems like you could get quite a speed up (e.g. 20x) if you got the timestep to 0.5s)....

The dynamical core timestep was set to one-tenth of the default value, and the physics timestep was set to 1 second. This configuration reflects underlying issues with the topography, which were somewhat expected. The hypothesis that topo-related problems are contributing to the instability is supported by the comparison between the current BA-100m and the "without the steep topographic gradients" 100m mesh as mentioned in the previous question. In more detail, due to OOM limitations, we were unable to use more realistic high-resolution topography – such as the 250 m global dataset or the 100 m regional data (which would require stitching and interpolation with other regions). As a result, the model dx exceeds the effective dx of RRM topo on 800 m cubed-sphere, which can lead to unrealistic blocky mountain phenomenon with flat centers and overly steep edges. These artificial slopes can introduce excessive high-wavenumber energy into the simulation. Once the current toolchain limitations are resolved, we expect the dynamical core timestep to return to its default value, allowing the simulation to run approximately 10 times faster.

Page 10, L242: This paragraph states both “one simulated hour per wall clock hour” (i.e. 1SDPD) then “0.16 SDPD”. Please clarify.

Thanks for pointing this out! The first sentence was a typo; the second sentence is correct. The intended statement was “one simulated hour per 6.4 wall-clock hours”. This has now been corrected.

Page 12, L269: The 100m and 3.25km are two different simulations right? What are the number of grid cells and the timestep in the 3.25km simulation? I assume this is just the 100m grid without the additional ‘level’ or refinement?

Yes, these are two separate simulations. The 3.25 km configuration includes 67,872 physical grid cells, with a dynamical core timestep of 9.375 seconds and a physics timestep of 72 seconds. As shown in Fig. 1, the 100 m grid is based on a first-order of 800 m California refinement.

If the concern is whether the differences between the 100 m and 3.25 km simulations might be primarily driven by the surrounding 800 m region outside the Bay Area – rather than by the 100 m mesh itself – then the possibility is very low. One line of evidence comes from previous studies: the AR evaluation paper by Bogenschutz et al. (2024) conducted a detailed comparison between 3.25 km and 800 m CARRM simulations; Zhang et al. (2025) evaluated multi-year energy generation at both resolutions; and Zhang et al. (2024) assessed the resolution sensitivity at 800 m vs. 3.25 km on simulating extreme floods in northern China, where complex terrain also plays a key role. These studies found only marginal improvements at 800 m compared to 3.25 km. A second line of evidence is shown in Fig. 11 of this study: the Stockton station, which lies just outside the 100 m mesh and is located the 800 m region, exhibits much smaller improvements than other stations located within the 100 m mesh.

References:

1. Bogenschutz, P. A., Zhang, J., Tang, Q., & Cameron-Smith, P. (2024). Atmospheric-river-induced precipitation in California as simulated by the regionally refined Simple Convective Resolving E3SM Atmosphere Model (SCREAM) Version 0. Geoscientific Model Development, 17(18), 7029-7050.
2. Zhang, J., Caldwell, P. M., Bogenschutz, P. A., Ullrich, P. A., Bader, D. C., Duan, S., & Beydoun, H. (2024). Through the lens of a kilometer-scale climate model: 2023 Jing-Jin-Ji flood under climate change. Authorea Preprints.
3. Zhang, J., Golaz, J., Signorotti, M. V., Lee, H., Bogenschutz, P., Monteagudo, M., Ullrich, P. A., Arthur, R. S., Po–Chedley, S., Cameron–smith, P., & Watson, J.: Simulation of wind and solar energy generation over California with E3SM SCREAM regionally refined models at 3.25 km and 800 m resolutions, EGUsphere [preprint], https://doi.org/10.5194/egusphere-2025-3947, 2025.

Page 15, L324: what is the timestep of the land model on this grid? And also while I am thinking about it, what is the timestep of the land model when run interactively?

In the land-only simulations, the land model timestep is 1800 seconds, consistent with the 6-hourly atmospheric forcing from ERA5. In Zhang et al. (2024), we tested a shorter timestep of 150 seconds and found it had negligible impact on the land initial conditions. In the atm–land coupled simulations, the land model runs with the same timestep as the atmosphere's physics timestep, which is 75 seconds.

We have added two sentences here:

> “The land model uses a timestep of 1800 seconds in the land-only simulations, consistent with the 6-hourly atmospheric forcing from ERA5. In the atmosphere–land coupled simulations, the land model runs with the same timestep as the atmospheric physics, which is 75 seconds.”

Page 17, L367: I assume the soundings vary in time and space as they go up as well (different model grid boxes with height)? Or does this not matter.

The raw sounding data were vertically interpolated to the reference pressure levels of SCREAM’s 128 reference pressure levels, using a log(pressure) to log(pressure) interpolation method. A description of the pressure-level interpolation for the IGRA data has been updated in the following sentence:

> “The IGRA soundings were vertically interpolated using a log(pressure)–log(pressure) method to SCREAM’s 128-layer reference pressure levels, which are defined by midpoint pressures ranging from 998.5 hPa near the surface to 2.6 hPa at the model top.”

Page 18, L368: are you using a fixed pressure for each level? Shouldn’t it vary with surface pressure and not using just the reference pressure? I assume you have a 3D pressure field from the model. The errors for the storm case (surface pressure well below 1000hPa) would be considerable in the lower troposphere.

The IGRA data include pressure records, so they can be directly interpolated to fixed pressure levels. For the model data, pressures at each level are computed from the surface pressure PS and the hybrid coordinate coefficients (hyam, hybm, hyai, hybi, P0), and thus can also be interpolated to the same fixed pressure levels. This procedure inherently accounts for variations in PS. The purpose of this interpolation is to facilitate plotting, since MetPy's SkewT function requires pressure as the vertical coordinate.

We have added the following clarification:

> "For sounding comparisons, model data were interpolated to the same reference pressure levels, computed from the simulated surface pressure and hybrid sigma-pressure coefficients."

Page 20, L396: what vertical level are you referring to for the Venturi effect.

Thank you for bringing this up! Please refer to Video A1: around the midpoint of the simulation, following the passage of the AR front, a narrow band of vertical motion develops from north to south on the lee side of the coastal ranges. This feature is most pronounced at 850 hPa. Initially, we noted that it differed notably from the mountain wave pattern observed prior to the frontal passage. However, upon further inspection of U850 and V850, we found that the wind speed enhancement extended far downstream from the gap exits (whereas gap winds typically peak near the exit region). Moreover, due to the lack of vertical profile of wind output, we could not confirm whether the flow weakened with height (as the topo gap widens). As a result, we decided to remove the original description of this feature.

Page 20, L399: again, what level? Where is this to be seen on Figure 7? Might need panel labels (A-L).

Sorry for the confusion. The phenomenon also appears in Video V1, beginning around the last third of the simulation and persisting until roughly the final one-sixth.

Because the videos contain far more information than can be fully conveyed in a static figure, it is difficult to capture this feature comprehensively in the main-text plots. We have therefore added more explanation in the text and again directed readers to the videos provided in the Appendix:

> "During the final third of the simulation, multiple suspected cold pools propagate successively inland. Their gust fronts are evident in the 850 hPa vertical velocity field. We recommend readers consult the animations in Video Supplement, which contains far more information than can be captured in static snapshots."

Page 24, L437: For fig 11, is the timeseries coarse because it's a single observation? If so, then maybe showing the variation or average around the time of the observation from the model would be useful. E.g. If it is a point measurement, the model temporal variability could be within that envelope.

The timeseries is coarse because the time sampling in this observation product is rough. We can implement what you suggested for Fig. 11, since the sampling frequency of Meteomanz (6 h) and ISD (irregular, preprocessed to 3 h) is much lower than the model's output frequency (5 min). We decided not to apply this to Tides and Currents data as its sampling frequency (6 min) is already comparable to that of the model.

While revising this figure, we also noticed that the model data had previously been averaged in time to match the obs time sampling, whereas the observations were instantaneous measurements. When the observational frequency is low, this mismatch can introduce a systematic phase lag in the diurnal cycle. In the revised manuscript, the model time sampling has therefore been switched to instantaneous output. This change improved the agreement of the diurnal cycle with Meteomanz observations in both simulations, and also yielded better consistency between the Meteomanz and ISD time series.

Thanks for the nice suggestion! We have updated the method description and the timeseries figures for Meteomanz and ISD accordingly:

> "The model's 5 min instantaneous outputs were resampled to match the observational time sampling. For Meteomanz and ISD, we extracted instantaneous values aligned with observation times and calculated the standard deviation within each observational window. For Tides and Currents, the 5 min outputs were averaged to 6 min intervals."

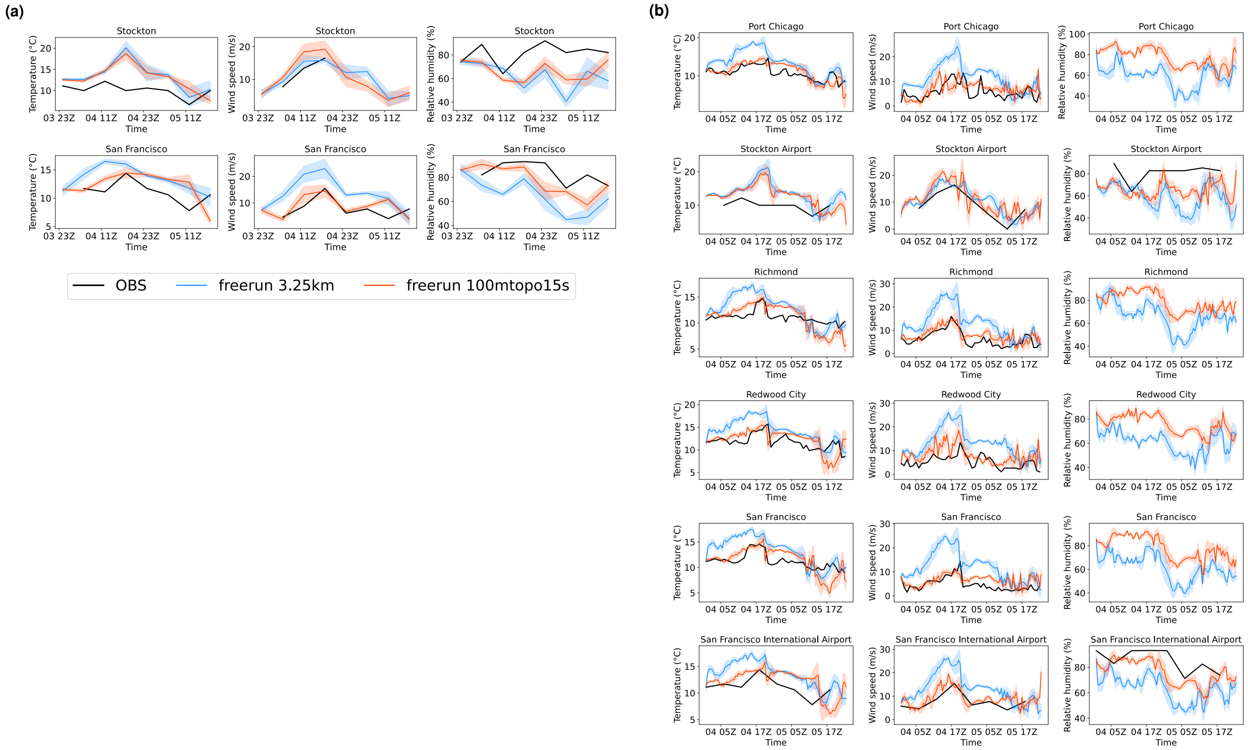


Figure 11. (a) Time series at each station for the Storm2008 event, with black, blue, and orangered lines representing Meteomanz observations, the 3.25 km California SCREAM-RRM simulation, and the 100 m Bay Area SCREAM-RRM simulation, respectively. (b) Same as (a), but for ISD observations. From left to right, the columns show near-surface air temperature, wind speed, and relative humidity. The bold line shows model output resampled from 5 min resolution to match the observational intervals as instantaneous values; shading denotes the standard deviation within each observational window.

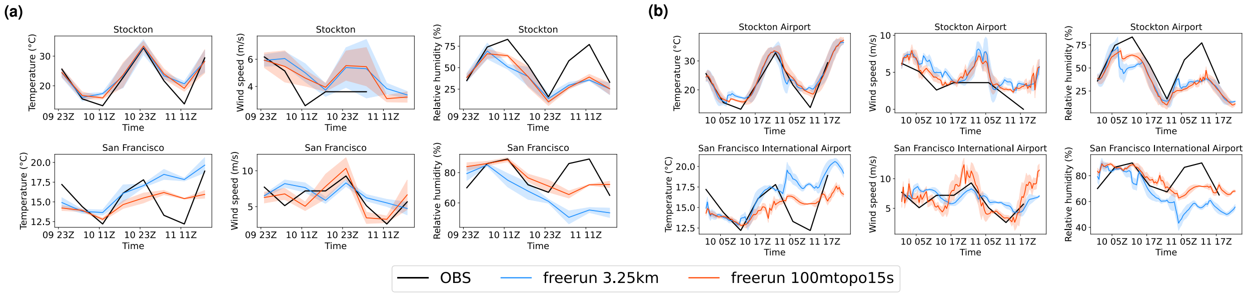


Figure 16. Same as Fig. 12 but for the Stratocumulus2023 event.

Page 26, L448: where is wind speed shown? Fig 11 and 12 are temperature and surface pressure.

The middle columns of Fig. 11 (Meteomanz, ISD) and Fig. 12 (Tides and Currents) show the wind speed.

Page 28, L471: reduces the wind bias from the 3km simulation? (Please be explicit).

Thanks! We have added “of the 3.25 km simulation”.

Page 28, L474: is the improvement specifically around orography?

This seems to be a wording issue.. what we intend to express is that, for the San Francisco Bay Area – a region that *as a whole* is characterized by complex terrain – precipitation processes related to orography may be better represented.

We have updated it to “with local orographic features in complex terrain”.

Page 28, L476: but increased turbulence (TKE) noted above would increase mixing and damp accumulation of cooler air masses.

What we intended to emphasize is that when more TKE is explicitly resolved, small-scale turbulent mixing processes are represented more realistically. However, more effective mixing does not necessarily imply that mesoscale phenomena like cold pools would be damped. A possibly unrelated but thought-provoking analogy is the sensitivity of aggregation to numerical filtering discussed in Silvestri (2004), which showed that SAM, without any scale-selective filter, contains much more energy at small scales and exhibits stronger aggregation (although these two phenomena are not necessarily causally related). In other words, enhanced small-scale TKE may also facilitate upscale energy transfer to the mesoscale, while, our current understanding is very limited, and we view this only as speculation.

References:

1. Silvestri, L., Saraceni, M., & Bongioannini Cerlini, P. (2024). Numerical diffusion and turbulent mixing in convective self-aggregation. Journal of Advances in Modeling Earth Systems, 16(5), e2023MS004151.

Page 28, L478: reference figure 13? Or similar for this case.

Yes, we have added the figure reference here.

Page 29, L487: Fig 14: is this a correlation over space? Can you put any variability on it? Say by correlating in space at different times?

Fig. 14 shows station-level metrics, e.g., the average correlation between observed and modeled timeseries at each station. It is difficult add variability at different times to this figure, because both Meteomanz and ISD data are relatively coarse. If correlations were computed at different times, the sample size within each subgroup would be too small (with the number of stations on the order of O(10)).

Nevertheless, we later performed 10 ensemble runs for each case using CA-3km, which allows us to add ensemble variability to the CA-3km bar in this figure. The updated metrics for the two events are shown below. The CA-3km simulation exhibits a non-trivial ensemble spread for the Storm2008 case, while the Stratocumulus2023 case shows almost no spread. More details on the CA-3km ensemble are provided in our response to Reviewer #2, where it is evident that the ensemble spread for Storm2008 begins to grow at the 34th simulation hour, while the moisture transport and overall precipitation patterns are highly robust.

The summary metrics plots have been updated in the main text:

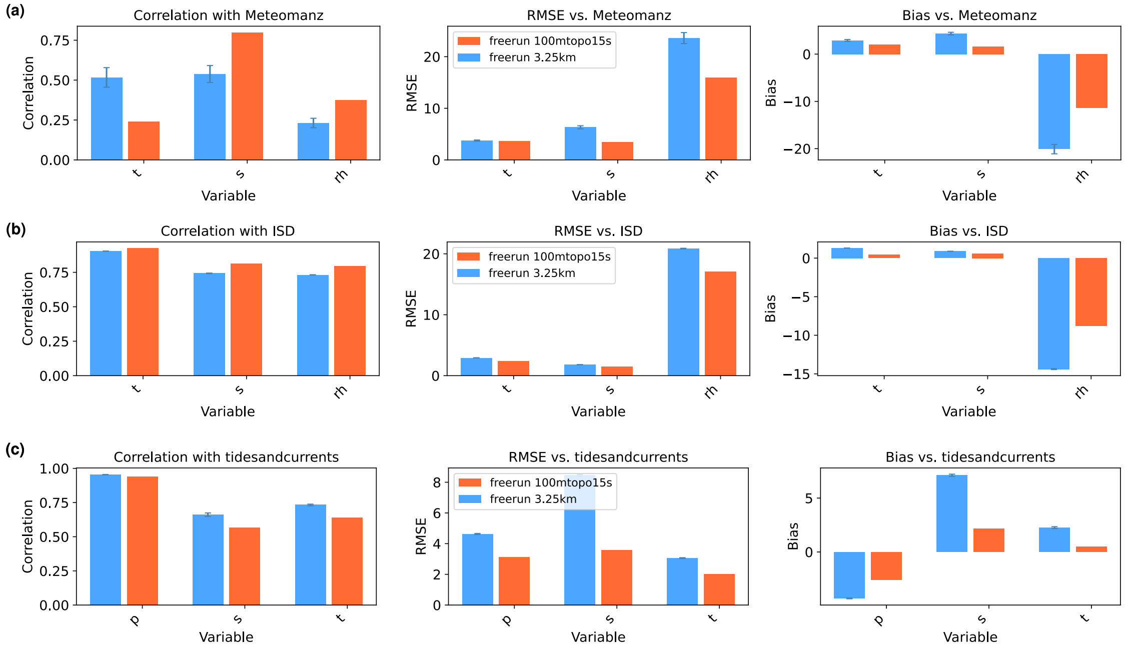


Figure 10. Skill scores for the Storm2008 event are shown for near-surface temperature (t), wind speed (s), relative humidity (rh), and surface pressure (p). These are compared against observations from (a) Meteomanz, (b) ISD, and (c) Tides and Currents, and presented as three overall metrics: Pearson correlation coefficient (left), root-mean-square error (RMSE, middle), and bias (right). The blue and orangered bars indicate simulation results from the 3.25 km California SCREAM-RRM and the 100 m Bay Area SCREAM-RRM, respectively. The ensemble spread in the 3.25 km simulation is represented by standard deviation bars.

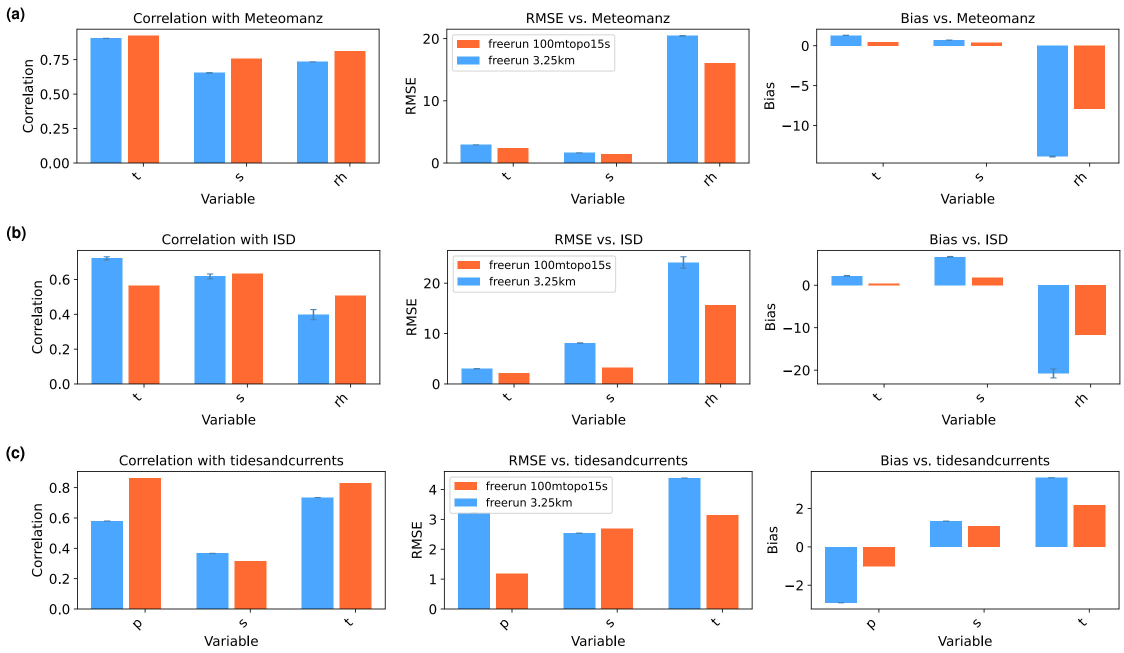


Figure 14. Same as Fig. 10 but for the Stratocumulus2023 event.

Page 29, L488: Do you mean Figure 15 here? If not, you refer to figure 16 before figure 15.

Thank you for pointing this out! We have corrected the order accordingly.

Page 31, L510: but these discrepancies would be expected right? If you averaged the radiosonde over the SCREAM or ERA5 levels you probably would not see the features right?

This is somewhat expected, but just to note that we interpolated both the IGRA soundings and SCREAM output to the same pressure levels for plotting, since MetPy’s SkewT requires pressure as the vertical coordinate.

Page 32, L525: is it really turbulence or just that the larger scale pattern is easier to initialize correctly and has less forcing than the other cases?

The initial conditions are the same in the CA-3km and BA-100m simulations, but the 100 m simulation shows improvement, which at least suggests that the improvement is not due to initialization? As you noted in your later comment regarding kinetic energy spectra (please see the response below), our supplementary analysis confirms the expected behavior: in the 100 m simulation, the kinetic energy is generally well resolved at scales smaller than 1 km.

Page 32, L531: can you show this partitioning?

As shown in our earlier response, we did not output resolved TKE. Fortunately, however, we did output SHOC TKE, moisture flux/variance, w variance, and the third moment of w. As shown in the added Fig. 18 and Fig. 19, the magnitudes of the SGS terms are much smaller in the BA-100m simulation compared to CA-3km. This similar partitioning is well documented in Fig. 9 and Fig. 16 of Bogenschutz (2023), which evaluated the SGS/resolved partitioning in a marine stratocumulus case, a maritime shallow cumulus case, and a mixed-phase Arctic stratocumulus case.

References:

1. Bogenschutz, P. A., Eldred, C., & Caldwell, P. M. (2023). Horizontal resolution sensitivity of the simple convection-permitting E3SM atmosphere model in a doubly-periodic configuration. Journal of Advances in Modeling Earth Systems, 15(7), e2022MS003466.

Page 32, L533: Your statement about avoiding the turbulence gray zone is contradicted by the statement below that turbulence should be modeled in 3D. What are the potential errors in this assumption?

This statement was intended to clarify that the gray zone (defined as scales larger than LES and smaller than mesoscale) requires a 3D turbulence scheme. Thus, at dx=100m, SHOC should largely avoid the gray zone issue (in contrast, at 800 m resolution a 3D turbulence scheme would be best added). As shown in our earlier response, we have revised this section in the main text accordingly.

Page 33, L554: You state that SHOC can smoothly transition without tuning. Is there a way to show this? KE spectra at different resolutions? It would be nice to show.

Based on this great comment from you and Reviewer #2, we have added KE and *w* spectra analyses. Note that spectra analysis is not a direct test of scale awareness, whereas the marked differences in SGS TKE/flux/variance between the CA-3km and BA-100m simulations (Fig. 18 and Fig. 19 in the updated manuscript) provide a more straightforward demonstration. As grid spacing decreases, the role of SGS terms naturally diminishes, which is precisely the manifestation of scale awareness, i.e., “transitioning smoothly from kilometer to LES scales without tuning”. The scale awareness of SCREAM was first quantified in Bogenschutz (2023)'s Fig. 9 and Fig. 16.

The added energy spectra analysis in the revised manuscript is as follows:

> “\subsection{Energy spectra} → in Methods
>
> For global models, spherical harmonics are a natural method for spectral decomposition, but they are not suitable for limited-area regional models and RRMs. For regional outputs, Discrete Fourier Transforms (DFT) and Discrete Cosine Transforms (DCT) are commonly used. DFT requires detrending or windowing, while detrending can artificially remove large-scale gradients, and windowing can distort spectra for already periodic fields \citep{Errico1985, Denis2002}. DCT mirrors the field to ensure symmetry before applying a Fourier transform, and is reliable for fields with spectral slopes between –4 and 1. DCT was originally developed for digital image, audio, and video compression (e.g., JPEG), but it has also been used to diagnose energy spectra in numerical simulations \citep[e.g.,][]{Denis2002, Selz2019, Prein2022}.
>
> We used the scipy.fft package (\url{https://docs.scipy.org/doc/scipy/tutorial/fft.html}, last access: 17 September 2025) for Discrete Cosine Transforms. Since we only output high-frequency relative vorticity, divergence, and \emph{w} profiles, we computed rotational and divergent KE spectra as well as \emph{w} spectra at every level using 10 min instantaneous outputs. The raw outputs were on the dynamical GLL grid; they were horizontally interpolated using the NCO-native first-order conservative algorithm to 0.03° (3.25 km California RRM) and 0.001° (100 m Bay Area RRM) over the small domain of the 100 m mesh (237.5E–238.5E, 37.3N–38.3N), and vertically interpolated to SCREAM’s 128 reference pressure levels using NCO’s default method. The two-dimensional spectra were projected onto the zonal and meridional directions, then averaged in time (from the 6th simulation hour to the end of the simulation) and in the vertical (within 100 hPa around 200, 500, and 850 hPa, respectively).
>
> \subsection{Energy spectra} → in Results
>
> Energy spectra provide a benchmark for evaluating the transition from large-scale quasi-2D motions to small-scale 3D turbulence. The canonical k^{-3} slope at synoptic scales and $k^{-5/3}$ slope at mesoscale wavelengths \citep{Nastrom_Gage1985} are widely used to assess effective resolution and numerical diffusion in atmospheric models \citep[e.g.,][]{Skamarock2004, Jablonowski_Williamson2011, Caldwell2021}. Numerous studies have examined KE spectra in global and regional models \citep[e.g.,][]{Bierdel2012, Skamarock2014, Durran2017, Menchaca_Durran2019, Prein2022, ZhangY2022,

Khairoutdinov2022, Silvestri2024}, with some studies have emphasized the rotational and divergent components \citep[e.g.,][]{Hamilton2008, BlažicaN2013, Selz2019}. Spectra of vertical velocity (\emph{w}) are also informative, as they emphasize divergent motions and typically peak at mesoscale wavelengths \citep{Bryan2003, Schumann2019}. Figures~\ref{spectraStorm2008}–\ref{spectraStratocumulus2023} show KE and \emph{w} spectra for the Storm2008 and Stratocumulus2023 cases.

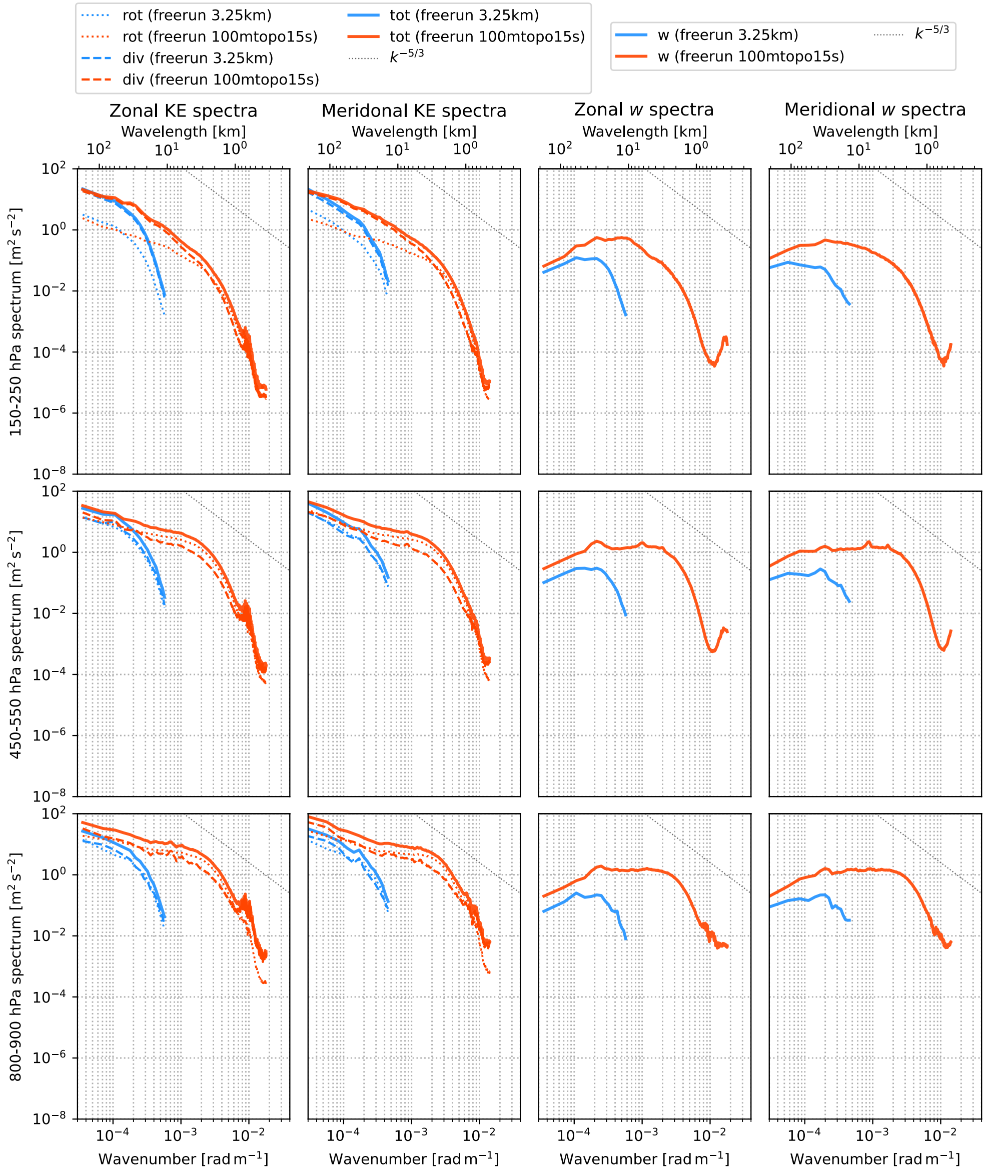


Figure 20. Energy spectra simulated by the 3.25 km California RRM (blue) and the 100 m Bay Area RRM (orangered) for the Storm2008 case. From left to right: zonal kinetic energy (KE), meridional KE, zonal vertical velocity (w), and meridional w spectra. From top to bottom: averages centered at 200 hPa, 500 hPa, and 850 hPa, each using a 100 hPa vertical window. In the KE spectra, the total, rotational, and divergent components are shown as thick solid, thin dotted, and thin dashed lines, respectively. A reference $k^{-5/3}$ slope line is shown in the top-right corner.

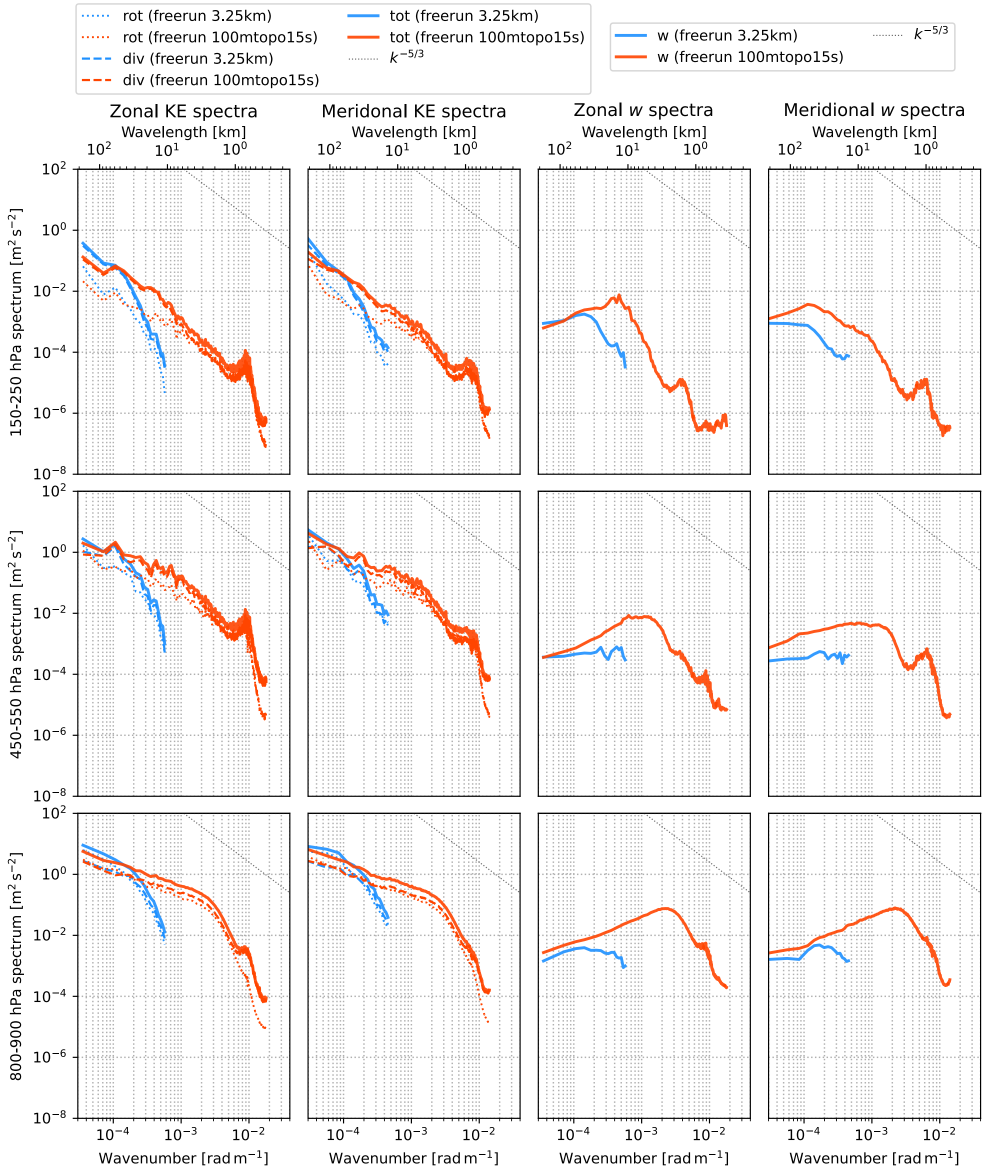


Figure 21. Same as Fig. 20 but for the Stratocumulus2023 event.

First, the CA-3km simulations roll off sooner than the global spectra in \citet{Caldwell2021}; however, we caution that the two differ in important ways (DCT over a region vs. spherical harmonics globally; two two-day events vs. 40-day statistics). For BA-100m, the effective resolution in Stratocumulus2023 is shorter than in Storm2008 if the roll-off standard is applied. However, within the mesoscale regime (10 km–100 m), the Storm2008 KE spectra flatten relative to $k^{-5/3}$ before steepening again, which corresponds to the earlier roll-off. The mesoscale flattening of the KE spectra in this case may reflect a genuine accumulation of mesoscale energy, given that this was a record-breaking extreme event. In both cases, it is robust that BA-100m contains much more small-scale energy than CA-3km.

A notable feature in both events is a sharp KE increase between 1 km and 500 m (approximately) in BA-100m. We suspect this results from the blocky mountain effect, caused by the mismatch between the 800 m cubed-sphere topography (the highest-resolution global dataset available) and the model's 100 m grid spacing. Toolchain memory limits prevented higher-resolution topography, so the effective topographic resolution lags behind the model Δx, producing unnaturally flat peaks and steep slopes. These slopes can generate excess high-wavenumber energy, consistent with the abrupt KE rise below 1 km. An alternative is contamination by inflow of coarser-resolution energy from the surrounding 800 m mesh via lateral boundaries. However, this is inconsistent with: 1) the effect being stronger in Stratocumulus2023 than in Storm2008, whereas boundary advection should amplify it in the latter; and 2) the 800 m mesh itself having an effective resolution of \textasciitilde4.8 km, with KE decaying rapidly beyond that, making a rise near 500 m hard to explain. To confirm the blocky mountain hypothesis, sensitivity tests with 100 m cubed-sphere topography are needed. To rule out lateral boundary effects, larger RRM domains need be tested \citep[cf.][]{Bogenschutz2024}. In either case, toolchain memory upgrades are essential.

The \emph{w} spectra exhibit a mesoscale peak, consistent with \citet{Bryan2003, Schumann2019}, with a cutoff near 1–3 km in BA-100m. Below 1 km, the ratio of \emph{w} to KE spectra approaches unity in Storm2008 but remains below one in Stratocumulus2023, except in the lower troposphere. The temporal evolution further shows rapid development of small- and mesoscale \emph{w} spectra, which are well developed within the first simulation hour (not shown).

Overall, KE spectra vary substantially across these two events, reflecting the influence of different forcing mechanisms and dynamical regimes. This aligns with \citet{Menchaca_Durran2019} and \citet{Selz2019}, and does not support a universal mesoscale KE spectrum."

Page 33, L557: It's probably better to say that you ignore the turbulence gray zone problems and it seems to work in these cases. That's an interesting point (and useful). Again, would be nice to show a figure that illustrates this.

Thanks for asking this! We are indeed operating at a critical resolution, i.e., based on the criterion that determines whether a scale lies in the turbulence gray zone regime by comparing the energy-containing turbulence length scale and the spatial filter scale. When the two are of similar magnitude, a 3D turbulence scheme may be required (or more precisely, its added value becomes significant). Compared to Storm2008, where the energetic vortices are much larger in scale, the Stratocumulus2023 case lies closer to the gray zone boundary. Some papers suggest using effective resolution instead of nominal grid spacing to assess this. Of course, developing a 3D turbulence scheme would be beneficial, and once implemented, it will allow us to quantify whether its added value is substantial at 100 m resolution or not.

As noted earlier, we have already added the discussion in the subsection titled “Sub-grid-scale flux”.

Response to RC2

The study presents the SCREAM/E3SM model run at global-scale to 100 m horizontal resolution over the San Francisco Bay Area in the US for two selected weather forcing periods.

Overall, the manuscript is written in much detail which is appreciated but feels lengthy. I strongly suggest moving the technical sections of the manuscript to either supplemental or appendix. That space could be used to discuss more about the capabilities of the model. The use of varying grid resolution for refining over regions of interest is quite challenging and is important as authors noted. Demonstrating that the model can deliver measurable skill gains relative to the baseline run at 3.25 km while remaining numerically stable is an advancement for the modeling community. To my knowledge, this is the first LES-scale within a global model framework.

Having said that, some scientific and technical issues need further treatment before the manuscript is publishable, especially the generalization of the model performance based on two case studies. I understand the computational cost associated with each case, however, to accept the SCREAM model for routine use in the LES community, it would be better to have either a small ensemble, or another synoptically driven flow (perhaps Diablo winds!) simulated, or a different region to answer few questions such as - How does the model perform for mid or high-latitudes? - What would be the sensitivity to the initial and lateral boundary conditions? I therefore recommend a major revision.

Thank you for your insightful, balanced comments, and for recognizing the significance of this work! We would like to emphasize that the primary contribution of this study is to demonstrate and document the feasibility of running E3SM SCREAM at 100 m resolution – as you noted, this represents the first such attempt within a global model framework. Prior to performing the simulations, we were unsure whether such a setup would even be viable. Thus, during the early design phase of the project (especially while applying for computing resources), it was not feasible to plan for a broader set of simulations after a successful proof-of-concept, since success was far from guaranteed.

Frankly, the simulations and analyses presented in the current manuscript have already exhausted (actually exceeded) the resources approved. Much of the allocated computing time was consumed during the trial-and-error phase, including months of development, test runs, and debugging prior to launching the full simulations. The additional 100 m simulations you suggested are exciting and we deeply appreciate your recognition of their potential, however, they go far beyond what this work is intended to cover.

We would like to take this opportunity to clarify the scope and contribution of this work:

First, this is a proof-of-concept study, focused on demonstrating feasibility rather than conducting a scaled-up, systematic evaluation of model performance for routine use. The latter, in our view, should be carried out by larger collaborative teams with expertise across various scientific processes and evaluation metrics.

Second, beyond demonstrating feasibility, this study identifies key limitations and opportunities for improvement in the current toolchains. For instance, we were unable to use 100 m topography consistent with the model resolution because the existing topography tool runs only in serial interpolation from the 250 m source DEM to model grids triggered out-of-memory issues on all available machines. A practical solution would involve developing an MPI interpolation tool and simplifying the mapping algorithm. Such development needs fall into several categories. Until they are implemented, additional simulations would remain prohibitively expensive. Moreover, proceeding without fixing these known bottlenecks is impractical. Clear improvements exist that could greatly reduce resource demands; otherwise, new simulations would again require 10× the reasonable cost which is currently not feasible. As far as we know, these toolchain developments are already part of broader community plans.

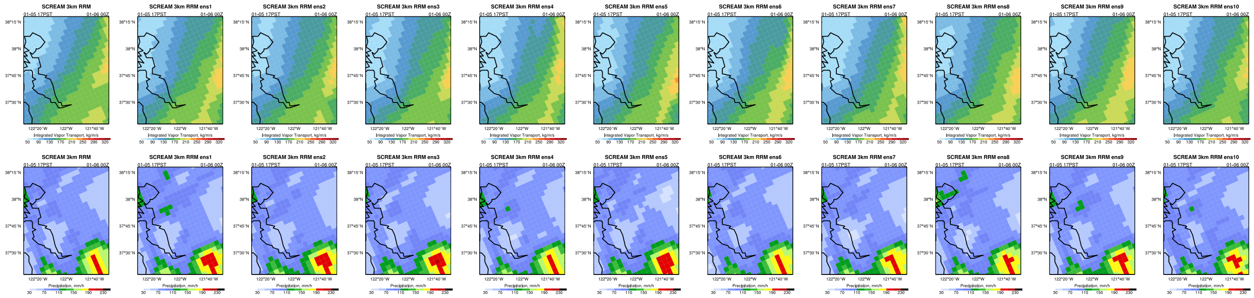
Third, because this work serves as a first demonstration of feasibility and a documentation, the methodology itself is a central contribution. A key mission of this paper is to record every step as clear as possible, so that interested users can replicate the process. Given that many readers may not be interested in highly technical details, we have reorganized the Methods section: detailed content has been moved into deeper-level subsections, and summarizing sentences were added at the start of the Methods.

All in all, we hope this clarifies the role and scope of the study, the practical challenges of designing/conducting new 100 m simulations, and the broader community efforts already in progress. We are equally eager to see more E3SM SCREAM simulations at 100 m resolution and more in-depth analyses in the future, which would undoubtedly expand the model’s scientific impact. Once current toolchain limitations are resolved, additional simulations will become much more feasible – especially with the support of GPU resources.

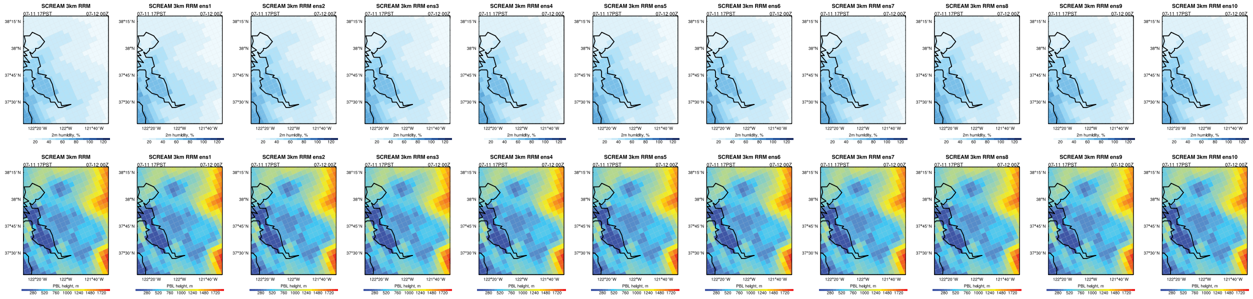
We sincerely thank you again for your careful review and constructive suggestions! In response, we have incorporated three major updates to the analysis: 1) two sets of ensemble simulations for CA-3km for each case, 2) energy spectra analysis, and 3) a comparison of SGS TKE, fluxes, and variances in CA-3km vs. BA-100m. Following your specific comments, we have updated five figures, added four figures and two tables in the main text.

=============

To quantify the sensitivity to small perturbations in the initial conditions, we conducted two small ensembles (10 members) for the CA-3km simulations of both cases. The ensembles were generated by applying random perturbations to the initial temperature profiles at all grid points. Fig. R2.1 and Fig. R2.2 shows the domain-averaged vertically integrated moisture transport and accumulated precipitation at the final timestep for the Storm2008 case, and 2 m relative humidity and online-diagnosed boundary layer height for the Stratocumulus2023 case. Apart from small uncertainty in the location of the precipitation maximum in Storm2008, the moisture transport and overall precipitation patterns are highly robust. Differences in the Stratocumulus2023 case are almost unable to distinguish visually.

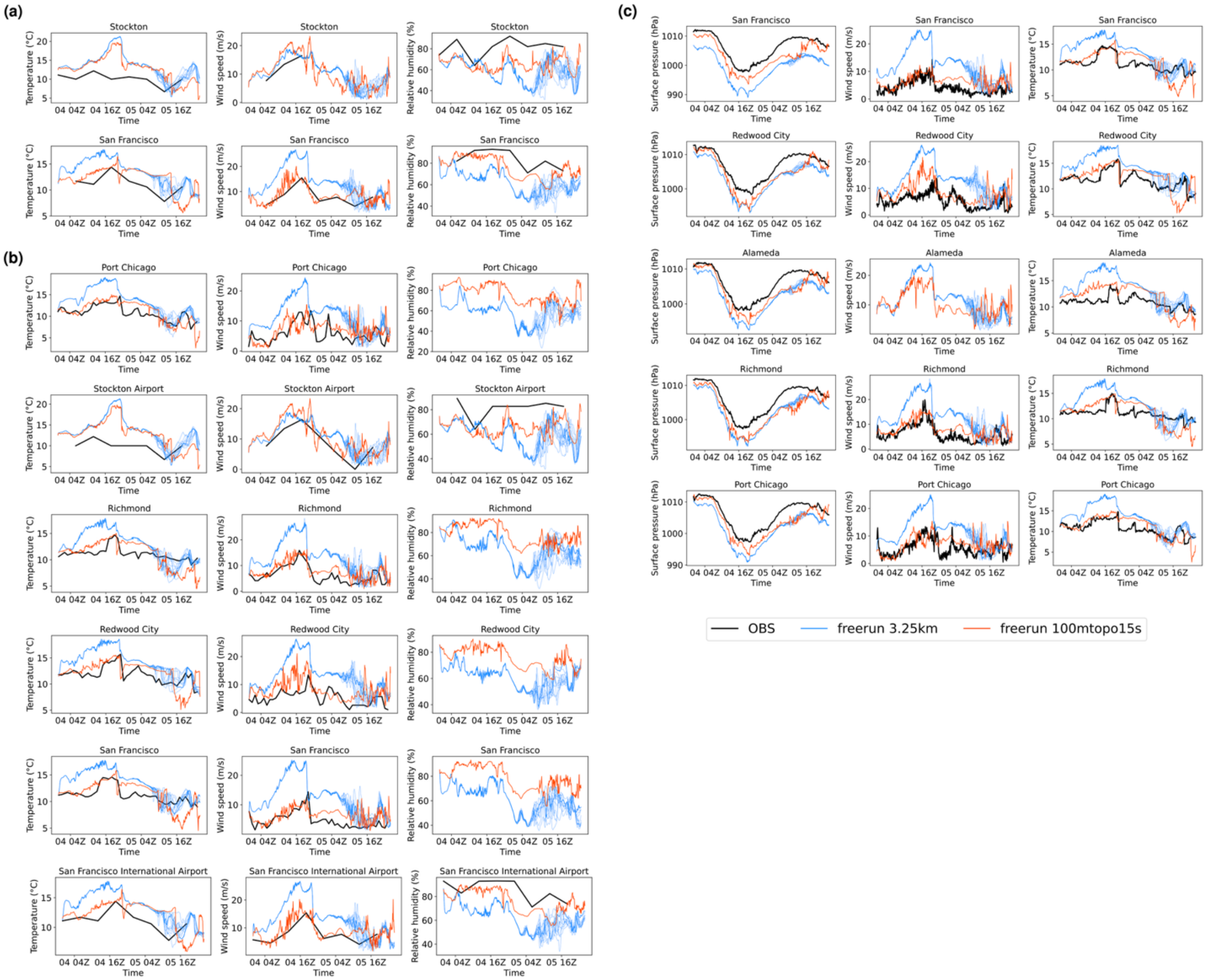


R2.1 vertically integrated vapor transport (top) and accumulated precipitation (bottom) at the final timestep of the Storm2008 event, for the single-realization control simulation (left column) and ensembles 1–10 (2nd to 11th columns).

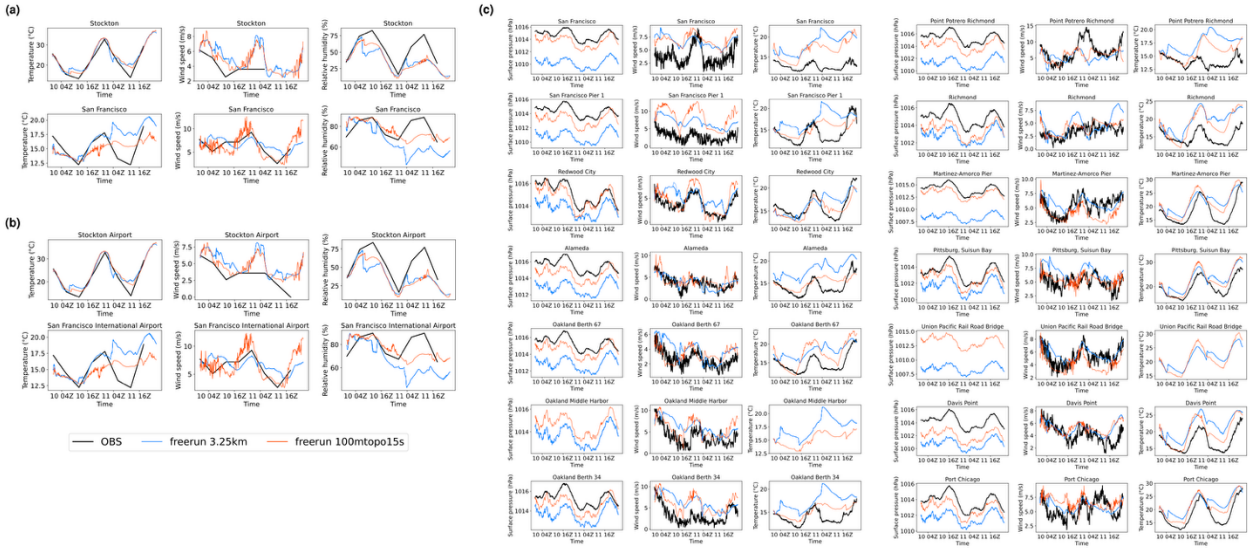


R2.2 Same as R2.1 but for 2 m relative humidity (top) and boundary layer height (bottom) for the Stratocumulus2023 event.

Fig. R2.3 and Fig. R2.4 show timeseries of in situ timeseries evaluation for all ensemble members over the simulation period. Thick lines represent the control runs, and thin lines represent ensemble members. Note that these plots differ slightly from their counterparts in the main text: here, to better show high-frequency variability, we did not resample the raw 5 min instantaneous model outputs to match the obs, nor did we compute variability based on the model outputs within each observational window (as was done in the Meteomanz and ISD plots in the updated manuscript). Instead, we directly plot the 5 min timeseries in Fig. R2.3 and Fig. R2.4.



R2.3 (a) Time series at each station for the Storm2008 event, with black, blue, and orangered lines representing Meteomanz observations, the 3.25 km California SCREAM-RRM simulation, and the 100 m Bay Area SCREAM-RRM simulation, respectively. (b) and (c) are the same as (a), but for ISD and Tides and Currents observations. For the 3.25 km simulations, the thick line indicates the single-realization control run, and the thin lines represent ensemble members 1–10.



R2.4 Same as R2.3 but for the Stratocumulus2023 event.

For Storm2008, ensemble spread in wind speed and relative humidity begins to emerge after the 34th simulation hour, with temperature showing moderate spread and surface pressure showing very little. Nevertheless, the ensemble spread does not alter the first-order comparisons. For Stratocumulus2023, ensemble spread remains negligible throughout the two simulation days.

The metrics plots in the main text have been updated to include standard deviation bars for CA-3km, representing ensemble spread. The relatively large bias and RMSE in CA-3km are maintained, reinforcing that resolution sensitivity played a dominant role, rather than random variability:

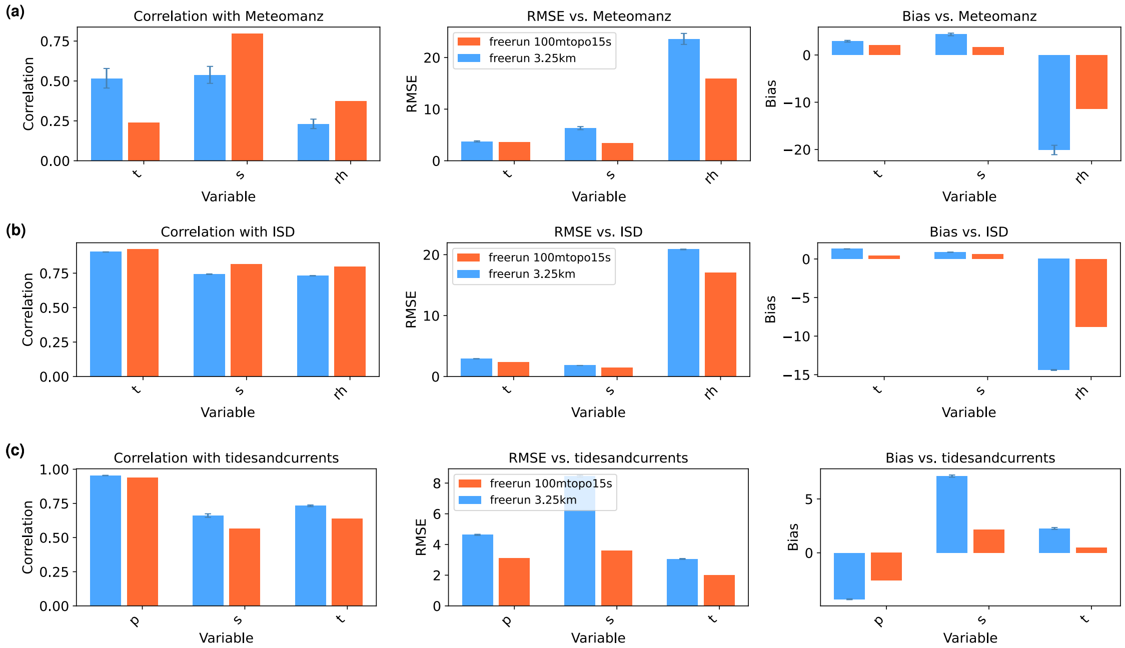


Figure 10. Skill scores for the Storm2008 event are shown for near-surface temperature (t), wind speed (s), relative humidity (rh), and surface pressure (p). These are compared against observations from (a) Meteomanz, (b) ISD, and (c) Tides and Currents, and presented as three overall metrics: Pearson correlation coefficient (left), root-mean-square error (RMSE, middle), and bias (right). The blue and orangered bars indicate simulation results from the 3.25 km California SCREAM-RRM and the 100 m Bay Area SCREAM-RRM, respectively. The ensemble spread in the 3.25 km simulation is represented by standard deviation bars.

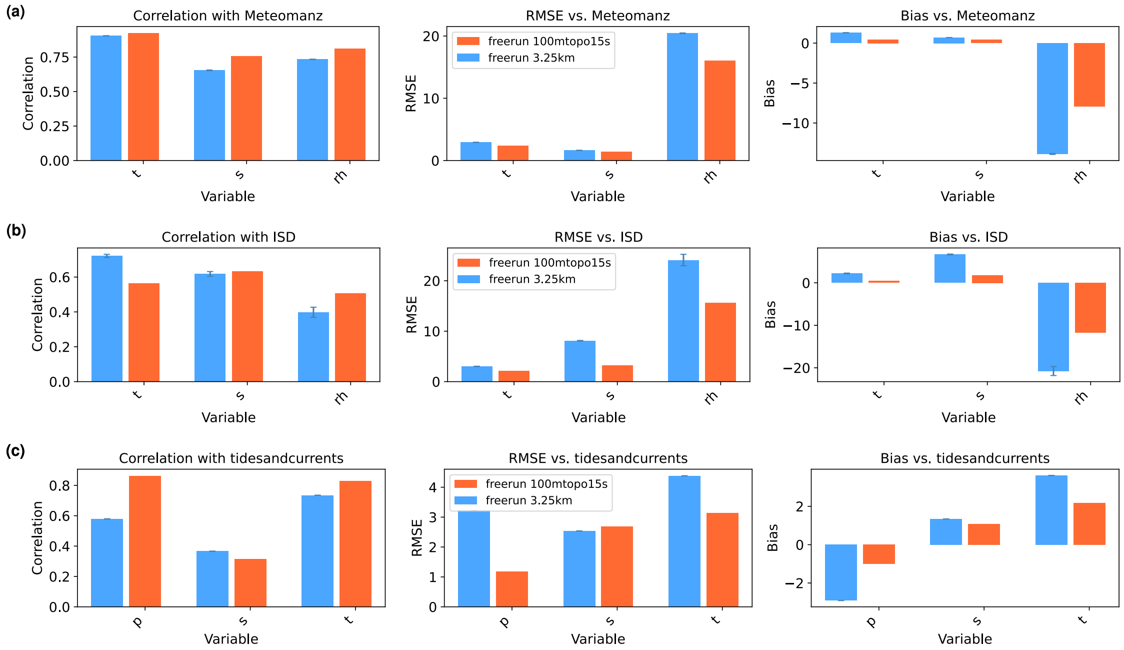


Figure 14. Same as Fig. 10 but for the Stratocumulus2023 event.

We have added the following descriptions to the Methods and Results sections of the manuscript:

> “\subsubsection{Sensitivity to initial condition} → in Methods
>
> In addition to the single-realization control runs, we conducted small ensembles (10 members each) for both events in the 3.25 km California RRM to quantify sensitivity to small perturbations in the initial conditions. Ensembles were generated by adding random perturbations to the initial temperature profiles across all grid points. Due to computational resource constraints, we were unable to perform ensemble simulations for BA-100m. Ensemble spread is represented by standard deviation bars in the 3.25 km California RRM metrics.
>
> \subsubsection{Storm2008} → in Results
>
> Ten ensemble members were run for CA-3km to assess sensitivity to initial condition perturbations. In Fig.~\ref{metricsStorm2008}, the vertical bars on CA-3km values represent the standard deviation across ensemble members. The relatively large bias and RMSE remain, highlighting the key role of resolution sensitivity rather than random variability. Ensemble spread appears after hour 34 of the simulation, most prominently in wind speed and relative humidity, but does not alter the first-order comparisons with observations or BA-100m. Spatially, except for small uncertainty in the location of the precipitation maximum, the moisture transport and precipitation patterns are highly robust (not shown).
>
> \subsubsection{Stratocumulus2023} → in Results
>
> The CA-3km ensemble shows virtually no ensemble spread throughout the two-day simulation (Fig.~\ref{metricsStratocumulus2023}).”

Specific comments:

1. Sec 2.1: For a horizontal resolution of 100 m, the 30 m near surface layer thickness seems high. Are you sure you are resolving the surface shear with such an aspect ratio?

Comparison with IGRA sounding observations shows that near-surface wind shear is generally well captured in the 100 m simulations, except for a slight underestimation in Storm2008 on 01-05 23PST and Stratocumulus2023 on 07-11 05PDT. Mirocha et al. (2010) found that an aspect ratio between 2 and 4 yielded the best agreement with similarity solutions in their boundary-layer LES simulations using WRF. The near-surface aspect ratio in SCREAM falls within this range.

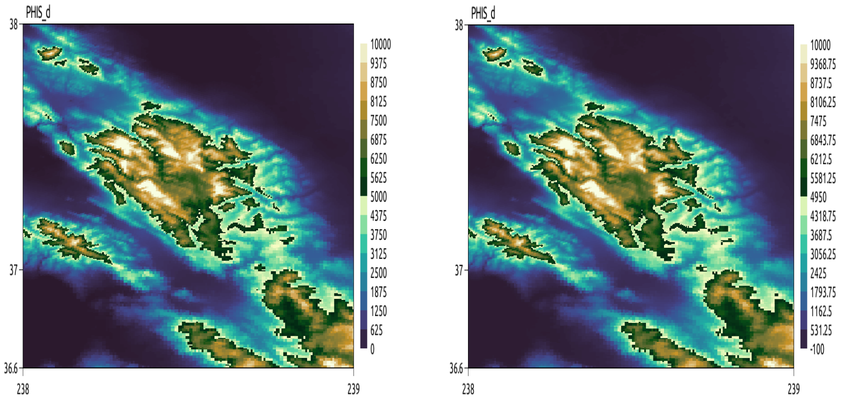
We have added this discussion to the IGRA results section.

References:

1. Mirocha, J. D., Lundquist, J. K., & Kosović, B. (2010). Implementation of a nonlinear subfilter turbulence stress model for large-eddy simulation in the Advanced Research WRF model. Monthly Weather Review, 138(11), 4212-4228.

2. Sec 2.3: The topography generation was discussed in much detail, however, no information was provided whether the extra smoothing has any effect on the terrain induced flows.

We were unable to compare the simulation “effects” between 6× and 12× smoothing because the run with 6× smoothing crashed.. We note that 12× smoothing is the recommended reference value in the E3SM v2 Topography Toolchain workflow. The figure below compares the topography after 6× (left) and 12× (right) smoothing, showing only minor visual differences:



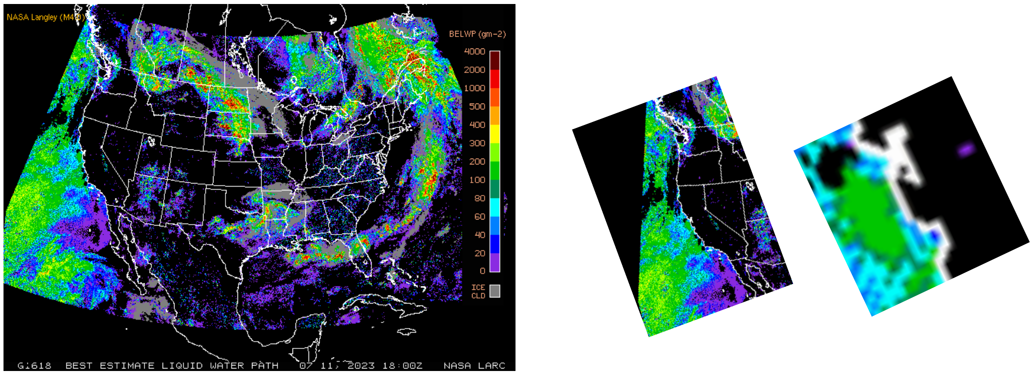
R2.5 BA-100m topography generated using 6 (left) and 12 (right) smoothing iterations.

3. Sec 2.4: What is the justification for increasing the hyper viscosity coefficient?

Looking back, we tested the sensitivity of several dycore parameters in the early setup tests, and the hyperviscosity coefficient was one of the commonly adjusted ones: raising it helps mitigate numerical instability. However, in retrospect, it appears that topography played a dominant role than this parameter. However, the hyperviscosity coefficient was not reverted to its default value when the simulations were eventually run.

4. Figure 5: Replace the GOES LWP map with a zoomed-in version so that it is easy to qualitatively compare against the model plots.

Since only the GOES images were available (not the raw data), we relied on a screenshot for a rough qualitative comparison. As you can see, after zooming in, the coarse resolution of GOES leaves very few pixels within the domain of interest:



R2.6 GOES-East/West Merged CONUS LWP Best Estimate, shown at three spatial scales: full North America (left), California (middle), and the Bay Area (right).

5. Even though the observations are sparse, the selection of the Storm2008 case warrants a quick comparison between the model simulated precipitation amounts and the observed ones. Capturing the extreme precip amounts has been a challenge for many fine resolution models. Showing that SCREAM at 100m does a reasonable job would make the discussion more robust.

During our initial evaluation, we did search for in situ precipitation records, but found that all available sources (Meteomanz, ISD, tidesandcurrents) either lacked precipitation observations entirely or reported maximum rates no higher than 3.2 mm/h, with most time steps missing data altogether. However, Storm2008 brought record-breaking extreme rainfall to the region (https://en.wikipedia.org/wiki/January_2008_North_American_storm_complex).

We appreciate your thoughtful and balanced comment! And we would like to reiterate that this study is not intended as a comprehensive evaluation of the model. Such an effort would require the involvement of experts across multiple areas. As noted in the discussion, we acknowledge that more complete information prior to the simulation might have guided us toward selecting an event with denser observational coverage. These are valuable lessons and experiences that will inform future works.

6. Figure 6: Some station names are overlapped, making it difficult to identify them.

Thank you for pointing this out! We have adjusted the figure to improve the readability of station names.

7. Figures 10 and 11 could be converted to tables.

Did you mean the original Fig. 10 and Fig. 14? We are quite fond of the detailed view in Fig. 11, as it provides information that summary metrics alone cannot convey (such as the high-frequency variability in observed winds). For the summary metrics figures, we've actually received suggestions in other works to convert tables into plots for clearer visual interpretation.

In response to your comment, we've added corresponding tables for the summary metrics while retaining the original plots, so readers can refer to whichever format they prefer:

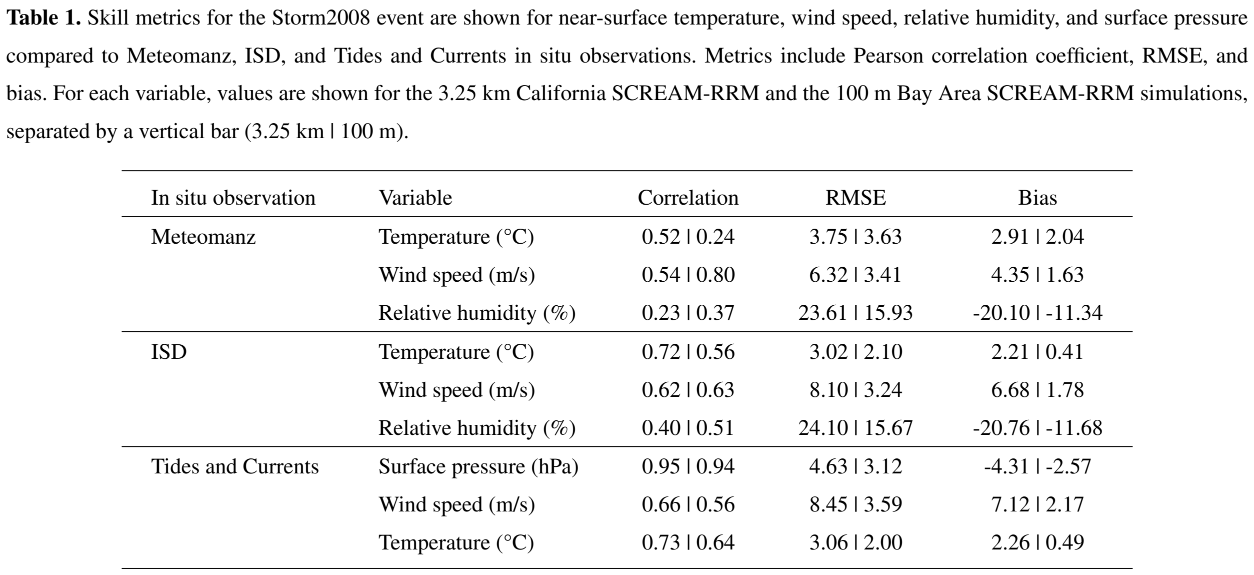


Table 1. Skill metrics for the Storm2008 event are shown for near-surface temperature, wind speed, relative humidity, and surface pressure compared to Meteomanz, ISD, and Tides and Currents in situ observations. Metrics include Pearson correlation coefficient, RMSE, and bias. For each variable, values are shown for the 3.25 km California SCREAM-RRM and the 100 m Bay Area SCREAM-RRM simulations, separated by a vertical bar (3.25 km | 100 m).

In situ observation	Variable	Correlation	RMSE	Bias
Meteomanz	Temperature (°C)	0.52 | 0.24	3.75 | 3.63	2.91 | 2.04
	Wind speed (m/s)	0.54 | 0.80	6.32 | 3.41	4.35 | 1.63
	Relative humidity (%)	0.23 | 0.37	23.61 | 15.93	-20.10 | -11.34
ISD	Temperature (°C)	0.72 | 0.56	3.02 | 2.10	2.21 | 0.41
	Wind speed (m/s)	0.62 | 0.63	8.10 | 3.24	6.68 | 1.78
	Relative humidity (%)	0.40 | 0.51	24.10 | 15.67	-20.76 | -11.68
Tides and Currents	Surface pressure (hPa)	0.95 | 0.94	4.63 | 3.12	-4.31 | -2.57
	Wind speed (m/s)	0.66 | 0.56	8.45 | 3.59	7.12 | 2.17
	Temperature (°C)	0.73 | 0.64	3.06 | 2.00	2.26 | 0.49

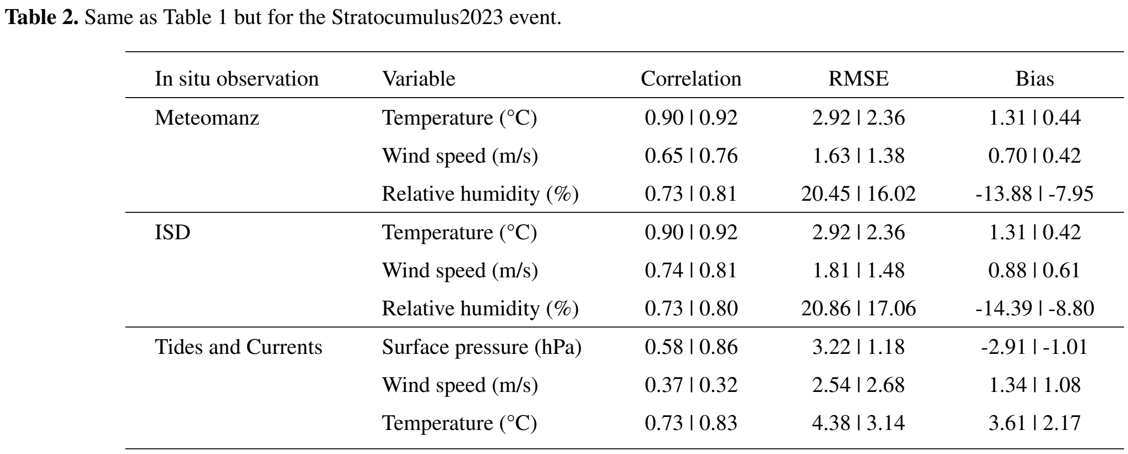


Table 2. Same as Table 1 but for the Stratocumulus2023 event.

In situ observation	Variable	Correlation	RMSE	Bias
Meteomanz	Temperature (°C)	0.90 | 0.92	2.92 | 2.36	1.31 | 0.44
	Wind speed (m/s)	0.65 | 0.76	1.63 | 1.38	0.70 | 0.42
	Relative humidity (%)	0.73 | 0.81	20.45 | 16.02	-13.88 | -7.95
ISD	Temperature (°C)	0.90 | 0.92	2.92 | 2.36	1.31 | 0.42
	Wind speed (m/s)	0.74 | 0.81	1.81 | 1.48	0.88 | 0.61
	Relative humidity (%)	0.73 | 0.80	20.86 | 17.06	-14.39 | -8.80
Tides and Currents	Surface pressure (hPa)	0.58 | 0.86	3.22 | 1.18	-2.91 | -1.01
	Wind speed (m/s)	0.37 | 0.32	2.54 | 2.68	1.34 | 1.08
	Temperature (°C)	0.73 | 0.83	4.38 | 3.14	3.61 | 2.17

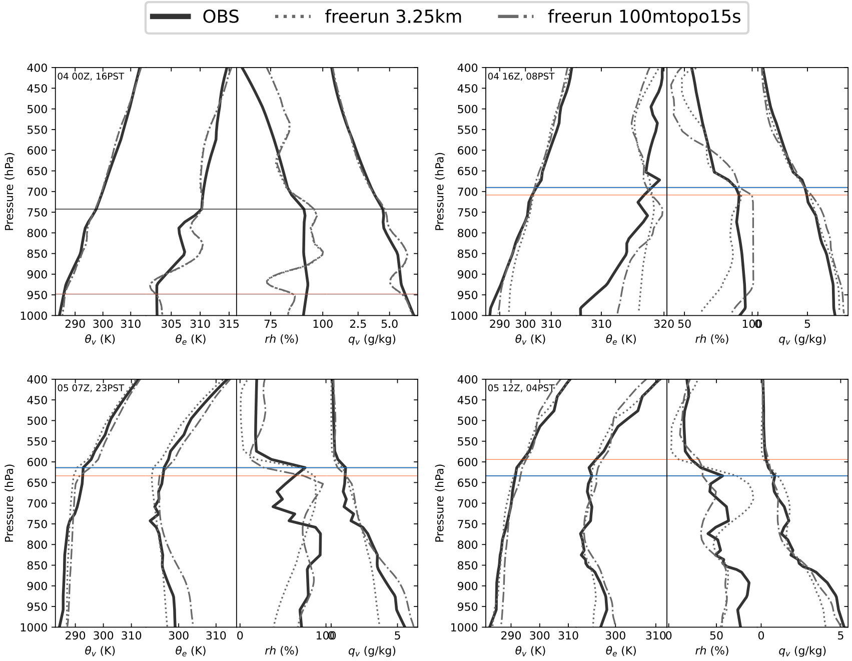
8. Figure 13 shows a significant difference in the mixed layer profiles at Oakland. Especially at 23PST on 2008-01-05, the dew point difference got a lot better between the 3.25 km and 100 m runs. This is a considerable difference and needs a comment or two in the manuscript. Also, it would be better to compare the boundary layer height estimations against the soundings to see if the LES is capturing all the turbulent motions.

Thank you for pointing this out! We have now explicitly referenced this figure in the sentence discussing the improved dew point in the 100 m run.

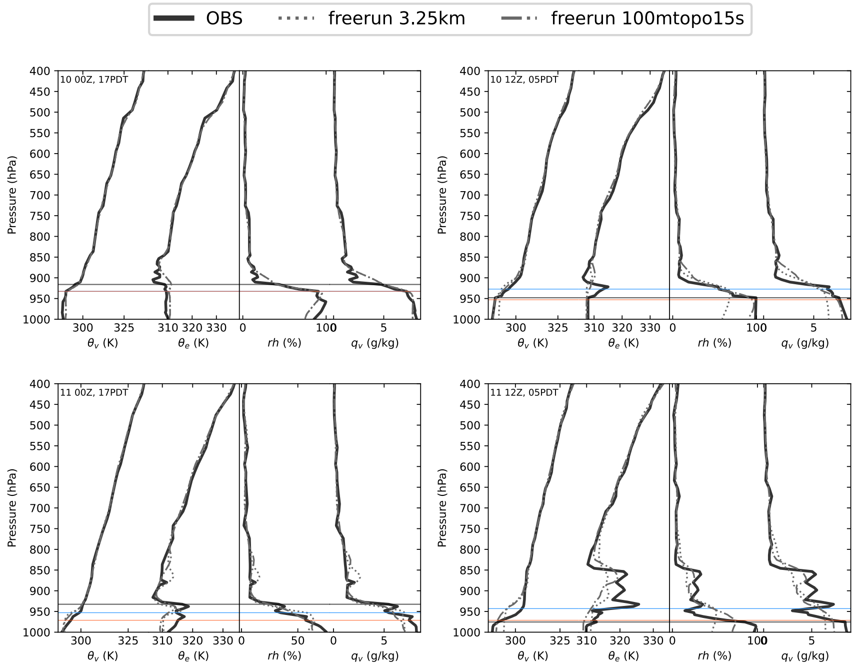
Also thanks for the nice suggestion about PBLH! While SCREAM includes an online diagnostic PBLH based on the Richardson number in SHOC, this method relies on surface friction velocity, which is not available from observations. We attempted to infer observational PBLH using the same approach by assuming a friction velocity, but the resulting estimates corresponded poorly with subjectively identified PBLH from the sounding plots.

Therefore, we adopted a simple RH-gradient method with the following steps: 1) Compute the forward finite difference of RH with respect to pressure (dRH/dp). 2) Identify RH inflection points that meet the following criteria: a) dRH/dp is positive at the current level, b) dRH/dp is larger at the current level than at the next lower level, c) dRH/dp becomes negative at the next lower level or the ratio between current and the next lower levels exceeds 2. 3) Among all detected RH inflection points, the one with the steepest gradient is taken as the PBL top.

Fig. R2.7 and Fig. R2.8 show the PBLH estimates from IGRA (black), CA-3km (blue), and BA-100m (orangered), respectively. Note that All profiles and sounding data were interpolated to the reference pressure levels before analysis. In the Storm2008 case, except at the initial time, the PBLH from CA-3km is visually indistinguishable from the IGRA estimate. The PBLH in the BA-100m simulation is generally lower. In the Stratocumulus2023 case, except at the initial time and at 07-11 00Z, the PBLH from BA-100m is closer to IGRA, whereas CA-3km tends to overestimate the stratocumulus boundary layer height.



R2.7 Vertical profiles of virtual potential temperature, equivalent potential temperature, relative humidity, and specific humidity at Oakland International Airport for the Storm2008 event. IGRA, the 3.25 km California RRM, and the 100 m Bay Area RRM are shown as thick solid, dotted, and dotted-dashed lines, respectively, with their corresponding estimated PBL heights indicated by black, blue, and orangered lines.



R2.8 Same as R2.7 but for the Stratocumulus2023 case.

A brief discussion of the PBLH comparison has been added to the main text:

> "\subsubsection{Storm2008} → in Results
>
> Estimated planetary boundary layer height (PBLH) at Oakland International Airport using a RH-gradient method agrees closely between CA-3km and IGRA, with differences that are visually indistinguishable. The BA-100m simulation generally underestimates PBLH.
>
> \subsubsection{Stratocumulus2023} → in Results
>
> We note that the estimated PBLH using a RH-gradient method from BA-100m aligns better with IGRA, while CA-3km tends to overestimate the stratocumulus boundary layer height."

9. Lack of turbulence characterization in the study. Given how much of the manuscript relies on the LES capability, the turbulence metrics such as velocity power spectra, resolved to sub-grid turbulent kinetic energy, flux-gradient relations are expected. LES models are often judged by these metrics, and providing at least one such metric

would convince readers about the SCREAM model LES capabilities. Authors have hinted that improved performance could be linked to capturing turbulent mixing (L525) but did not provide any supporting evidence.

Thank you and Reviewer #1 for this valuable suggestion! In response, we have added two analyses: 1) SGS turbulent kinetic energy, fluxes, and 2) KE and w spectra. These additional analyses provide a more explicit characterization of turbulence and help support the LES-scale capabilities of SCREAM.

This content has been added into two subsections in the main text:

> “\subsection{Sub-grid-scale flux}
>
> Building on the resolution sensitivity study of DP-SCREAM in \citet{Bogenschutz2023}, SCREAM exhibits characteristics of a scale aware model. As horizontal resolution increases, the partitioning between SGS and resolved turbulence diminishes. This scale awareness, inherent to the SHOC parameterization \citep{Bogenschutz_Krueger2013}, enables SCREAM to operate effectively at 100 m resolution without the need for parameter tuning. Specifically, Fig. 9 and Fig. 16 in \citet{Bogenschutz2023} show that in the marine stratocumulus case, maritime shallow cumulus case, and mixed-phase Arctic stratocumulus case, as the horizontal grid spacing Δx decreases, the contribution of SGS moisture flux becomes smaller while the resolved flux becomes increasingly dominant. In the 100 m DP-SCREAM simulations, above 0.2 km, the proportion of SGS moisture flux is negligible, and the resolved flux is nearly unity.
>
> In our simulations, although we did not output high-frequency resolved moisture flux or TKE, Fig.~\ref{SHOCtimeplevStorm2008} and Fig.~\ref{SHOCtimeplevStratocumulus2023} shows significant reductions in SGS TKE, moisture flux/variance, \emph{w} variance, and the third moment of \emph{w} in the BA-100m simulations compared to the CA-3km simulations. For clarity, the plotted ranges in these figures differ between the two resolutions, with the CA-3km values being much larger.

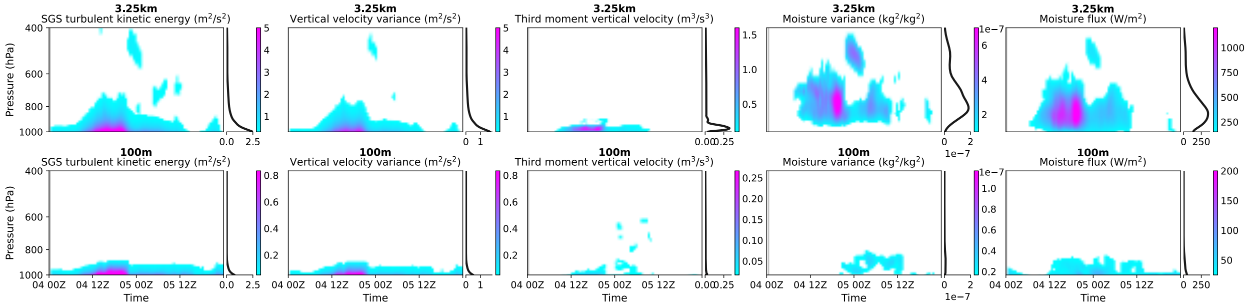


Figure 18. Simulated Sub-grid-scale (SGS) variables in the 3.25 km California RRM (top) and 100 m Bay Area RRM (bottom) during the Storm2008 event. From left to right: TKE, vertical velocity variance, third-moment vertical velocity, moisture variance, and moisture flux. Each panel consists of a time–evolution shading plot on the left and a vertical profile averaged over the simulation period on the right. For clarity, the colorbars differ between the 3.25 km and 100 m simulations, with the former being much larger.

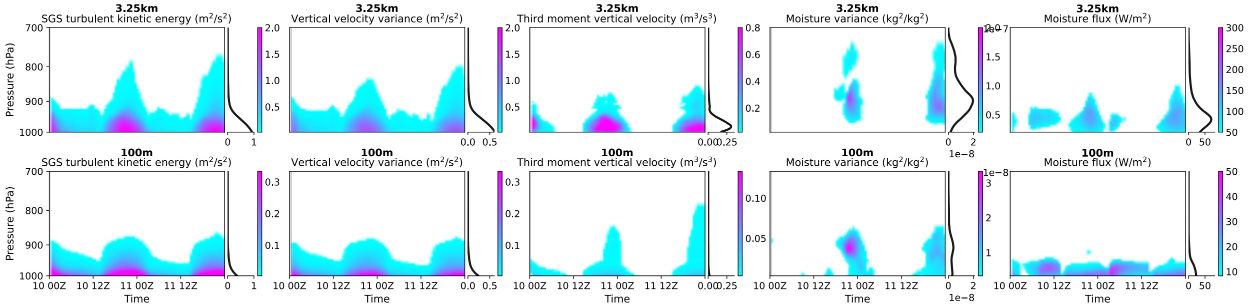


Figure 19. Same as Fig. 18 but for the Stratocumulus2023 event.

At 100 m resolution, simulations are close to, but largely below, the turbulence gray zone, where the grid spacing becomes comparable to the dominant eddy scale. The gray zone typically spans the transition from mesoscale models, which rely on ensemble-averaged vertical fluxes, to LES, where most turbulent motions are explicitly resolved and subgrid closures play only a minor dissipative role \citep{Wyngaard2004}. In this transitional regime, subgrid transport is best treated with 3D turbulence schemes that represent the full stress tensor \citep[e.g.,][]{Wyngaard2004,Chow2019,Honnert2020}, whereas SHOC currently parameterizes only vertical mixing. Thus, at coarser resolutions such as 800 m, a 3D implementation of SHOC would likely be beneficial. At 100 m, near the lower edge of the gray zone, the need for 3D turbulence remains uncertain and case-dependent, pending the implementation and testing of such a scheme in SCREAM."

"\subsection{Energy spectra} → in Methods

For global models, spherical harmonics are a natural method for spectral decomposition, but they are not suitable for limited-area regional models and RRMs. For regional outputs, Discrete Fourier Transforms (DFT) and Discrete Cosine Transforms (DCT) are commonly used. DFT requires detrending or windowing, while detrending can artificially remove large-scale gradients, and windowing can distort spectra for already periodic fields \citep{Errico1985, Denis2002}. DCT mirrors the field to ensure symmetry before applying a Fourier transform, and is reliable for fields with spectral slopes between –4 and 1. DCT was originally developed for digital image, audio, and video compression (e.g., JPEG), but it has also been used to diagnose energy spectra in numerical simulations \citep[e.g.,][]{Denis2002, Selz2019, Prein2022}.

We used the scipy.fft package (\url{https://docs.scipy.org/doc/scipy/tutorial/fft.html}, last access: 17 September 2025) for Discrete Cosine Transforms. Since we only output high-frequency relative vorticity, divergence, and \emph{w} profiles, we computed rotational and divergent KE spectra as well as \emph{w} spectra at every level using 10 min instantaneous outputs. The raw outputs were on the dynamical GLL grid; they were horizontally interpolated using the NCO-native first-order conservative algorithm to 0.03° (3.25 km California RRM) and 0.001° (100 m Bay Area RRM) over the small domain of the 100 m mesh (237.5E–238.5E, 37.3N–38.3N), and vertically interpolated to SCREAM's 128 reference pressure levels using NCO's default method. The two-dimensional spectra were projected onto the zonal and meridional directions, then averaged in time (from the 6th simulation hour to the end of the simulation) and in the vertical (within 100 hPa around 200, 500, and 850 hPa, respectively).

\subsection{Energy spectra} → in Results

Energy spectra provide a benchmark for evaluating the transition from large-scale quasi-2D motions to small-scale 3D turbulence. The canonical k^{-3} slope at synoptic scales and $k^{-5/3}$ slope at mesoscale wavelengths \citep{Nastrom_Gage1985} are widely used to assess effective resolution and numerical diffusion in atmospheric models \citep[e.g.,][]{Skamarock2004, Jablonowski_Williamson2011, Caldwell2021}. Numerous studies have examined KE spectra in global and regional models \citep[e.g.,][]{Bierdel2012, Skamarock2014, Durran2017, Menchaca_Durran2019, Prein2022, ZhangY2022, Khairoutdinov2022, Silvestri2024}, with some studies have emphasized the rotational and divergent components \citep[e.g.,][]{Hamilton2008, BlažicaN2013, Selz2019}. Spectra of vertical velocity (\emph{w}) are also informative, as they emphasize divergent motions and typically peak at mesoscale wavelengths

\citep{Bryan2003, Schumann2019}. Figures~\ref{spectraStorm2008}–\ref{spectraStratocumulus2023} show KE and \emph{w} spectra for the Storm2008 and Stratocumulus2023 cases.

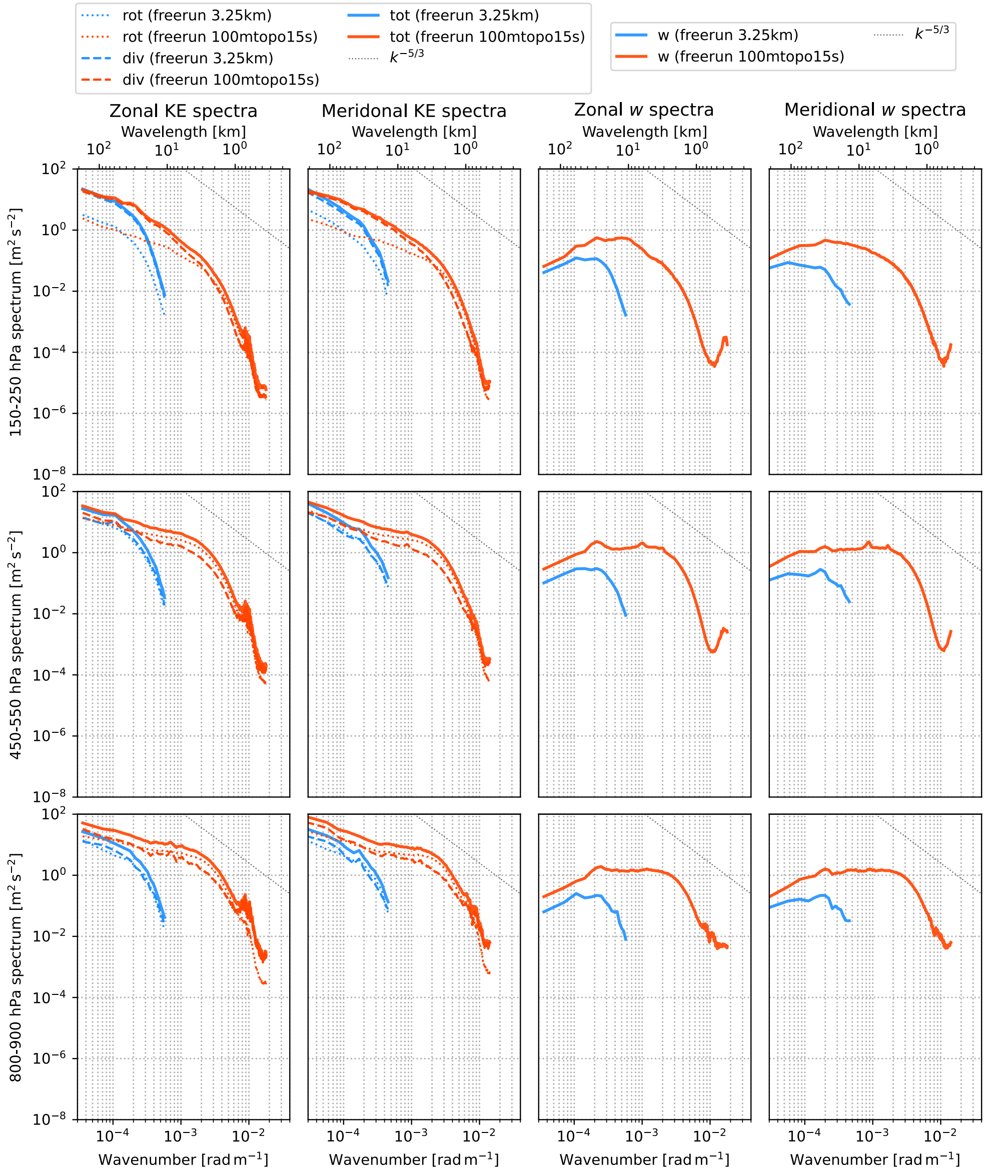


Figure 20. Energy spectra simulated by the 3.25 km California RRM (blue) and the 100 m Bay Area RRM (orangered) for the Storm2008 case. From left to right: zonal kinetic energy (KE), meridional KE, zonal vertical velocity (w), and meridional w spectra. From top to bottom: averages centered at 200 hPa, 500 hPa, and 850 hPa, each using a 100 hPa vertical window. In the KE spectra, the total, rotational, and divergent components are shown as thick solid, thin dotted, and thin dashed lines, respectively. A reference $k^{-5/3}$ slope line is shown in the top-right corner.

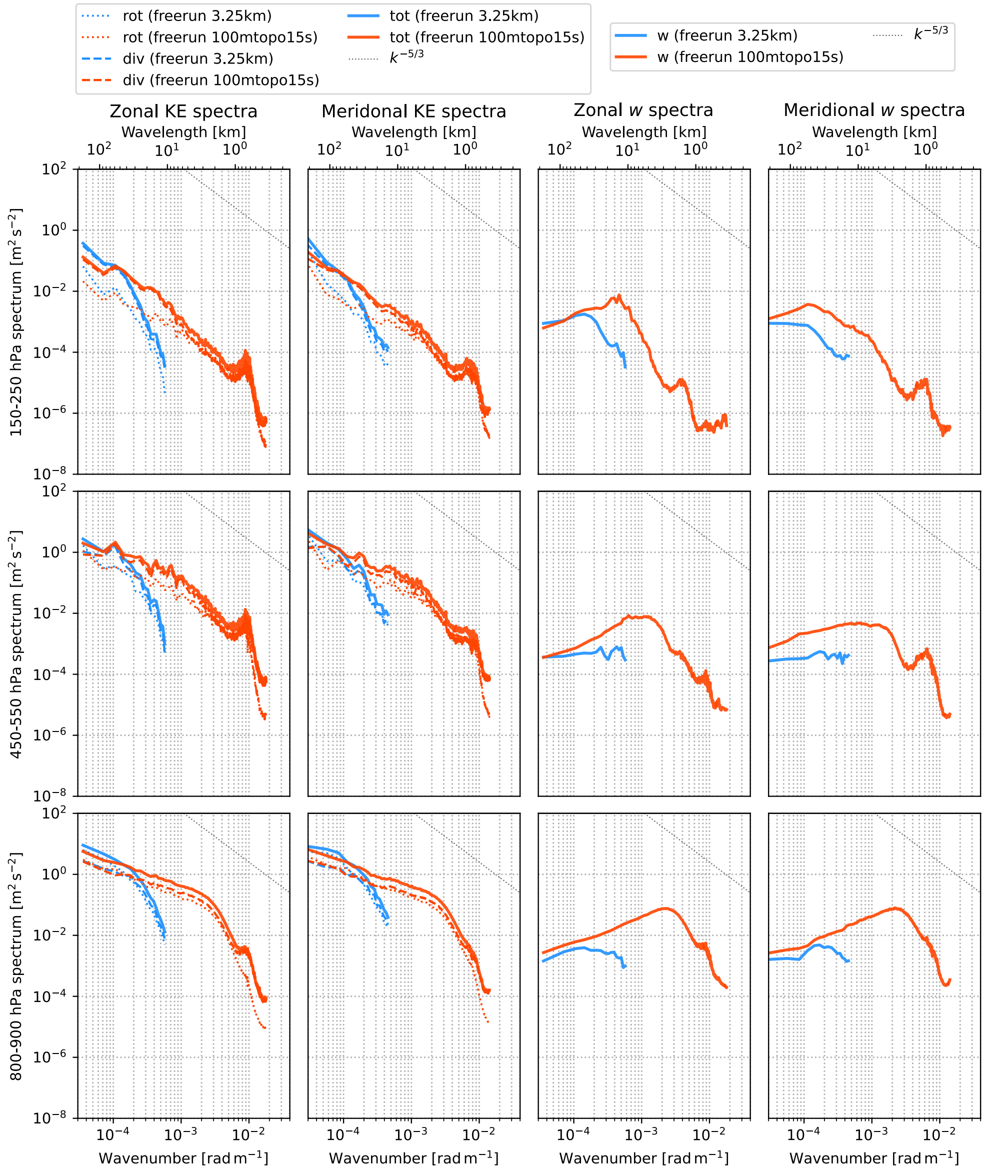


Figure 21. Same as Fig. 20 but for the Stratocumulus2023 event.

First, the CA-3km simulations roll off sooner than the global spectra in \citet{Caldwell2021}; however, we caution that the two differ in important ways (DCT over a region vs. spherical harmonics globally; two two-day events vs. 40-day statistics). For BA-100m, the effective resolution in Stratocumulus2023 is shorter than in Storm2008 if the roll-off standard is applied. However, within the mesoscale regime (10 km–100 m), the Storm2008 KE spectra flatten relative to $k^{-5/3}$ before steepening again, which corresponds to the earlier roll-off. The mesoscale flattening of the KE spectra in this case may reflect a genuine accumulation of mesoscale energy, given that this was a record-breaking extreme event. In both cases, it is robust that BA-100m contains much more small-scale energy than CA-3km.

A notable feature in both events is a sharp KE increase between 1 km and 500 m (approximately) in BA-100m. We suspect this results from the blocky mountain effect, caused by the mismatch between the 800 m cubed-sphere topography (the highest-resolution global dataset available) and the model's 100 m grid spacing. Toolchain memory limits prevented higher-resolution topography, so the effective topographic resolution lags behind the model Δx, producing unnaturally flat peaks and steep slopes. These slopes can generate excess high-wavenumber energy, consistent with the abrupt KE rise below 1 km. An alternative is contamination by inflow of coarser-resolution energy from the surrounding 800 m mesh via lateral boundaries. However, this is inconsistent with: 1) the effect being stronger in Stratocumulus2023 than in Storm2008, whereas boundary advection should amplify it in the latter; and 2) the 800 m mesh itself having an effective resolution of \textasciitilde4.8 km, with KE decaying rapidly beyond that, making a rise near 500 m hard to explain. To confirm the blocky mountain hypothesis, sensitivity tests with 100 m cubed-sphere topography are needed. To rule out lateral boundary effects, larger RRM domains need be tested \citep[cf.][]{Bogenschutz2024}. In either case, toolchain memory upgrades are essential.

The \emph{w} spectra exhibit a mesoscale peak, consistent with \citet{Bryan2003, Schumann2019}, with a cutoff near 1–3 km in BA-100m. Below 1 km, the ratio of \emph{w} to KE spectra approaches unity in Storm2008 but remains below one in Stratocumulus2023, except in the lower troposphere. The temporal evolution further shows rapid development of small- and mesoscale \emph{w} spectra, which are well developed within the first simulation hour (not shown).

Overall, KE spectra vary substantially across these two events, reflecting the influence of different forcing mechanisms and dynamical regimes. This aligns with \citet{Menchaca_Durran2019} and \citet{Selz2019}, and does not support a universal mesoscale KE spectrum."

10. As noted in the manuscript, despite improvements, the model consistently underestimates the surface pressure at the majority of stations. Where does this bias come from? Is it the model dynamics or the initial conditions? Authors should include some discussion on this.

Thanks for asking this! Since all simulations were initialized using ERA5 data (with both atmospheric and land surface conditions consistently initialized), our study is unable to isolate or quantify the contribution of initial conditions to the surface pressure bias. However, one potential factor related to initialization is the surface adjustment procedure. Because ERA5 has much coarser topography than SCREAM, failing to adjust surface pressure according to the high-resolution model topography would introduce spurious low or high pressure and hydrostatic imbalance at the model surface, which in turn triggers strong artificial gravity wave oscillations during spin-up. To mitigate this, the surface pressure adjustment was applied based on the difference in elevation between the source (ERA5) and target (SCREAM) topographies. This procedure uses the hydrostatic equation by assuming a dry adiabatic lapse rate (Trenberth, 1993). However, this assumption may introduce bias relative to the true surface pressure.

To our knowledge, some groups have started using O(10)-km reanalysis data blended with km-scale assimilated forecasts (e.g., 3 km High-Resolution Rapid Refresh forecast) to initialize km-scale simulations. These groups might have examined whether initializing with ERA5 vs. ERA5 + HRRR leads to meaningful differences. Replacing ERA5 with an alternative reanalysis product is also a possibility, although it would require careful coordination of atmospheric and land surface spin-up under consistent forcing. An alternative view is that, although the model spectrum tends to lack sufficient mesoscale kinetic energy during initialization, the mesoscale spectrum can rapidly develop (Skamarock, 2004). In this case, providing only large-scale kinetic energy in the initial conditions may not lead to significant loss. Whether model dynamics alone could lead to surface pressure errors of 1–2 hPa remains unclear.

We have added a brief discussion in the Results section:

> “Despite improvements, the model consistently underestimates surface pressure at the majority of stations. It remains unclear whether this bias is related to model dynamics. One possible factor tied to the initial conditions is the surface adjustment procedure, which assumes a dry adiabatic lapse rate. Some groups have begun blending reanalysis data at O(10)-km resolution with \emph{k}-scale analysis products to initialize \emph{k}-scale models. An alternative view suggests that, while mesoscale kinetic energy is initially lacking, it can rapidly spin up \citep{Skamarock2004}, so supplying only large-scale energy may not result in significant degradation.”