Author Comment (AC) — Response to Referees

egusphere-2026-1697: Multi-model high-resolution analysis of Tropical-Like Cyclone Daniel with WRF and ICON

Legend: *RC1 in blue*, *RC2 in green*, AC in black.

P. Serafini, on behalf of all co-authors

We thank both Referees for their careful reading of the manuscript and for comments that led to a substantial improvement of the analysis. Each point below follows the journal format: the Referee comment is quoted first, our response follows, and the resulting change in the manuscript is reported. Line numbers refer to the revised version.

The main structural changes in the revised manuscript are: (i) all track-error statistics have been recomputed after the discovery of a processing error (RC2, G1/G2b); (ii) a statistical significance assessment of the inter-configuration track differences has been added, accounting for serial correlation (RC2, G3); (iii) the non-harmonisable physics differences between WRF and ICON are now described explicitly and connected to the cross-model results (RC1 S1; RC2 G4); (iv) the resolution terminology has been redefined and the "gray zone" framing corrected throughout (RC2, G5); (v) the conclusions have been reorganised into case-specific findings, literature-consistent findings and open hypotheses (RC2, G6); (vi) new diagnostics have been added: precipitation difference maps and intensity distributions (RC1, S2), a CPS sensitivity analysis (RC2, SC1) and TASM asymmetry and displacement tests (RC2, SC2).

Response RC1

General comment

The manuscript presents a high-resolution, multi-model analysis of Medicane Daniel, comparing WRF and ICON simulations at convection-permitting resolution and assessing the sensitivity of the simulated track, intensity, precipitation, and tropical-like structure to different treatments of convection. The main findings show that both models are able to reproduce the overall life cycle and track of Daniel, but with substantial differences in cyclone intensity, precipitation distribution, and internal storm structure depending on both the model and the convection scheme. In particular, the study highlights the relevance of shallow-convection parameterization at approximately 2 km resolution, showing that even in convection-permitting simulations not all convective processes are explicitly resolved, and that a parameterization can still be beneficial for representing convective processes.

I appreciated the work because it provides a robust and systematic comparison between two widely used numerical weather prediction models, WRF and ICON, and between different convection schemes at convection-permitting resolution. This is a very timely topic and is fundamental for improving the simulation of high-impact precipitation events, such as those associated with "medicanes", where convective processes play a crucial role in storm intensification, precipitation and warm-core development.

I believe the manuscript fits well within the scope of the journal and deserves publication. However, a few specific comments should be addressed before it can be accepted for final publication.

Response: We thank the Referee for the careful reading and for the encouraging assessment of the

study. All specific and minor comments are addressed point by point below.

Changes in the manuscript: None for this point.

Specific comment 1

The objective of the present study is highly relevant for the community. In particular, understanding the role of convective parameterization at 2 km gray-zone resolution is important for high-impact Mediterranean systems such as medicanes. The comparison between WRF and ICON, and among the different convective configurations, is well structured, detailed, and clearly presented. However, I think the manuscript would benefit from a more explicit discussion of the physical mechanisms behind the differences among the simulations. This is particularly relevant in Section 4.2, where substantial differences are found in central sea-level pressure and 10 m wind speed (Fig. 5), as well as in precipitation and FSS scores (Figs. 6-7).
For example, the shallow-convection configurations appear to strongly affect the simulated medicane characteristics, but the analysis could better explain why this occurs physically. A clearer discussion of how the shallow-convection schemes in WRF and ICON influence boundary-layer processes, latent heat release, precipitation and cyclone characteristics would strengthen the interpretation of the results.
In addition, some differences between WRF and ICON may also arise from other model-dependent physical parameterizations, such as microphysics, turbulence/PBL schemes, surface fluxes, radiation, or gravity-wave drag. Although the effort to make the two configurations comparable is appreciated, these remaining differences in the physical setup could contribute to the different model responses. A short additional discussion, or where possible some supporting diagnostics, would help distinguish the role of convection schemes from the broader influence of the model physics and dynamical cores.
Overall, this would not require a major restructuring of the manuscript, but rather a more explicit physical interpretation of the differences already shown.

Response: We agree that the physical interpretation deserved more space, and we address the two parts of the comment separately.

(i) Mechanism of the shallow-convection sensitivity. At 2 km grid spacing, non-precipitating shallow cumuli remain unresolved. The two shallow schemes tested here (GRIMS in WRF–SH; the shallow-only Tiedtke–Bechtold mode in ICON–SH) vent boundary-layer moisture into the lower free troposphere. The sub-cloud layer dries, grid-scale moisture convergence weakens, and the onset of resolved deep convection is delayed; latent heating is released over a wider area instead of being concentrated in a few grid-point updrafts. This chain is consistent with the weaker but better organised cyclone in WRF–SH (Fig. 5), with the improved grid-scale localisation of extreme rainfall in that configuration (Fig. 7b), and with the smoother eyewall structure in the TASM diagnostics (Fig. 10b). Conversely, in the CU experiments the parameterized deep heating acts on top of the resolved convection and feeds a positive feedback between heating and central pressure fall, which explains the over-intensification of ICON–CU. We present this as a diagnostic interpretation: a quantitative attribution would require moisture-budget and heating-profile analyses that lie beyond the scope of this study, and we say so explicitly in the revised text. The new difference maps and precipitation distributions added in response to your comment 2 (below) support the same reading: the SH configurations reduce the amplitude of the localised over- and under-estimation dipoles produced by the purely explicit runs.

(ii) Unmatched physics. We agree that the residual differences in microphysics, PBL/turbulence, surface, radiation and gravity-wave drag contribute to the cross-model contrast. A dedicated paragraph has been added to Sect. 3.1 describing each non-harmonisable difference and its likely relevance, and the Conclusions now carry the corresponding caveat. We have also added two

sentences to the Discussion connecting these differences to the headline results (see our reply to Referee 2, comment G4, where the same issue is raised).

Changes in the manuscript: A new paragraph is added in Discussion and Conclusion to answer part (i) `(lines 514-522)`

> " The sensitivity to the shallow-convection treatment admits a consistent physical reading. At 2 km grid spacing, non-precipitating shallow cumuli are not resolved. The shallow schemes (GRIMS in WRF, the shallow-only Tiedtke Bechtold mode in ICON)515 vent boundary-layer moisture into the lower free troposphere; the sub-cloud layer dries, grid-scale moisture convergence weak- ens and the onset of resolved deep convection is delayed, so that latent heating is released over a wider area rather than in a few grid-point updrafts. This mechanism is consistent with the weaker but better organised cyclone in WRF–SH, with the improved localisation of extreme rainfall (Fig. 7(b)) and with the smoother eyewall structure in the TASM diagnostics (Fig. 10). In the CU experiments, instead, the parameterized deep heating acts on top of the resolved convection and feeds a positive feedback520 between heating and central pressure fall, consistent with the over-intensification of ICON–CU. This interpretation is diagnos- tic; a quantitative attribution would require moisture-budget and heating-profile analyses beyond the scope of this study "

In addition, Sect. 3.1 now contains the paragraph describing the non-harmonisable physics differences `(lines 158-177)`

> " Despite the effort to balance the configurations for the two simulations as much as possible, the physical parameterization suites of WRF and ICON could not be fully harmonized, reflecting the independent development lineages of the two modelling systems (Table 1). For each physics category we retained each model's standard, operationally validated configuration rather than forcing an artificial correspondence that neither code is designed to support. The most consequential divergence concerns cloud microphysics: the WDM6 scheme (Lim and Hong, 2010) predicts number concentration only for the warm-rain species, retaining a single-moment treatment for the ice-phase hydrometeors, whereas the two-moment scheme (Seifert and Beheng, 2006) used in ICON prognoses both mass and number concentration across all hydrometeor categories, including the ice phase. Radiative transfer is likewise computed by conceptually distinct codes, with RRTMG (Iacono et al., 2008) following a correlated-k approach, in contrast to the more modular, multi-solver architecture of ecRad (Hogan and Bozzo, 2018). Boundary-layer turbulence differs in closure order: the non-local, first-order YSU scheme (Hong et al., 2006), paired with a first-order Smagorinsky closure for residual 3D mixing in WRF, compared to ICON's prognostic, 1.5-order TKE-based scheme (Raschendorfer, 2001). Surface exchanges are computed by the single-tile Unified Noah land surface model in WRF (Niu et al., 2011), as opposed to the explicitly tiled TERRA scheme in ICON (Schrodin and Heise, 2001), which resolves sub-grid land-cover heterogeneity relevant to the complex coastlines of the Mediterranean basin. Gravity-wave drag is included in both configurations, though following distinct orographic formulations (Choi and Hong, 2015; Orr et al., 2010). Finally, convection is parameterized by mass-flux schemes of different heritage: the KIM Simplified Arakawa Schubert scheme in WRF (Han and Pan, 2011; Kwon and Hong, 2017), versus the Tiedtke–Bechtold scheme in ICON (Tiedtke, 1989; Bechtold et al., 2008), the latter treating shallow and deep convection as explicitly separate regimes. These structural differences, intrinsic to the two model architectures, represent an additional, model dependent source of uncertainty that should be considered alongside the dynamical and resolution-related aspects of the comparison discussed above. "

and the Conclusions carry the corresponding caveat `(lines 496-499)`.

> " Despite the effort to balance the configurations for the two simulations as much as possible, the physical parameterization suites of WRF and ICON could not be fully harmonized, reflecting the independent development lineages of the two modelling system. These structural differences, intrinsic to the two model archi- tectures, represent an additional, model-dependent source of uncertainty. "

Specific comment 2

Regarding the precipitation analysis, I suggest adding a few additional diagnostics to better assess the differences between the simulations and the observations, both in terms of spatial distribution and precipitation intensity.
First, in Fig. 6, the accumulated precipitation fields are useful to compare the general patterns among the models and IMERG. However, the spatial differences would be easier to interpret if model-minus-observation difference maps were also provided. This would help identify more clearly where each configuration overestimates or underestimates precipitation, and whether the errors are mainly related to displacement, intensity, or spatial extent of the rainfall structures.

Response: We thank the Referee for this suggestion. We computed model-minus-IMERG difference maps of the event-total accumulated precipitation for all seven configurations (Fig. 1). Both model fields and IMERG were first conservatively regridded to the common 0.1° verification grid, the same used for the FSS, so that the differences are not contaminated by resolution mismatch. The maps show that the dominant error structure along the cyclone track is a dipole pattern centred on the observed rainfall maxima, indicating that displacement, rather than a systematic intensity bias, drives most of the local disagreement during the marine phases. Over Greece, the WRF configurations overestimate accumulation along the windward slopes of the Pindus range, while the ICON configurations underestimate precipitation on the southern flank of the cyclone during the transition phase. The dipole amplitudes are smallest in the SH configurations of both models, consistent with sub-grid transport distributing convective precipitation more evenly and limiting grid-scale moisture convergence.

Changes in the manuscript: A new appendix subsection "Precipitation insight" with the corresponding figure (model-minus-IMERG event totals on the common grid, seven panels) has been added, and the a summary of the results is inserted in Sect. 4.2 `(lines 265-267)`

> " Model-minus-IMERG difference maps of the event-total precipitation, computed on the common verification grid (Fig. A4), show that the leading error structure along the track is a displacement dipole rather than a uniform intensity bias. WRF overestimates accumulation over the Pindus slopes, ICON underestimates it on the southern flank of the cyclone, and the dipole amplitudes are smallest in the SH configurations of both models. "

Second, the authors could consider computing the probability density function, or a similar distribution-based diagnostic, of precipitation during the event for each simulation and for the observations. This would provide a clearer comparison of the precipitation intensity distribution, especially for the highest percentiles and extreme values. Such an analysis would complement the FSS results and help clarify whether the models differ mainly in the localization of heavy precipitation or also in their ability to reproduce the intensity of extreme rainfall.

Response: We computed the frequency distributions of event-total precipitation for all simulations and for IMERG (Figs. 2 and 3). To avoid areal-averaging artefacts, all distributions are evaluated on the common 0.1° verification grid. All configurations reproduce the observed distribution up to approximately the 50th percentile, with a consistent behaviour up to 85th percentile. Above it, the model tails are heavier than the IMERG tail; the excess is largest for ICON–EXP and smallest for the ICON–CU configuration, in agreement with the FSS results at the 99th and 99.9th percentile thresholds. Part of the tail disagreement lies in the reference itself: IMERG Final applies a monthly climatological gauge adjustment and underestimated the observed maxima over Thessaly during this event (Flaounas et al., 2025), so the model tail excess over land should be read as an upper bound. Overall these results give no more informations than what already presented in the previous analysis.

Changes in the manuscript: The distribution figure has been added to the appendix subsection "Precipitation insight" because of its support to the previous analysis and this is mentioned in Sect. 4.2 `(lines 446-447)`

" The frequency distributions of event-total precipitation, evaluated on the common verification grid (Fig. A5-A6), show agreement with FSS scores. "

Minor comment 1

Lines 73–75: "As a result, widespread thunderstorms developed, with cloud tops exceeding 13 km, producing extreme precipitation and severe flooding throughout central and eastern Greece, as well as parts of Bulgaria and Turkey. Surface stations recorded more than 750 mm of daily rainfall and up to 1235 mm in 4 days in the eastern parts of the Thessaly region." Please provide appropriate references for the description of the flooding impacts in Greece, Bulgaria, and Turkey, as well as for the reported precipitation records from surface stations.

Response: We agree. The rainfall totals originate from the NOANN network of the National Observatory of Athens; we now cite Lagouvardos et al. (2017) for the network and Flaounas et al. (2025) for the event totals, and Hewson et al. (2024) for the description of the flooding impacts across Greece, Bulgaria and Turkey.

Changes in the manuscript: Added citation in `(lines 80-81)`

" Surface stations recorded more than 750 mm of daily rainfall and up to 1235 mm in 4 days in the eastern parts of the Thessaly region (Lagouvardos et al., 2017). "

Minor comment 2

Line 316: The statement "Fully parameterized configurations increase both the intensity and strength of the cyclone" should be clarified. Based on Fig. 5, this seems to be clearly valid only for ICON, where ICON–CU produces a stronger cyclone than the other ICON configurations. However, for WRF, WRF–EXP appears to produce slightly stronger 10 m winds and lower central sea-level pressure than WRF–CU, while only WRF–SH simulates a weaker cyclone. I suggest revising this sentence to better distinguish the different responses of WRF and ICON to the fully parameterized convection configuration.

Response: The Referee is right: the statement holds unambiguously only for ICON. We have revised the sentence to separate the two model responses.

Changes in the manuscript: Added sentence to clarify at `(lines 392-394)`

" Fully parameterized configurations increase both the intensity and strength of the cyclone, likely by creating deep convection cells that rotate in phase around an axis. This is particularly true for ICON, where ICON–CU reaches the lowest CSLP and the strongest MW10; while in WRF the WRF–CU experiment reaches intensities similar to those of WRF–EXP. "

Minor comment 3

Line 322: Please clarify the meaning of "mean accumulated precipitation per grid cell" and briefly explain how this quantity was computed. For example, please specify whether this quantity represents the spatial average of the event-total accumulated precipitation over the full analysis domain, or whether it is computed over a specific area around the cyclone.

Response: The quantity is the spatial average of the event-total accumulated precipitation over the full analysis domain. It is computed on the common 0.1° verification grid shared by both models and IMERG, so that the differing cell-area distributions of the native WRF (Lambert) and ICON (icosahedral) grids do not affect the comparison. We have clarified this in the text.

Changes in the manuscript: Added sentence in Sect. 4.2 `(lines 400-402)`

“ The mean accumulated precipitation per grid cell is an indicator of the total rainfall in the domain. This quantity has been computed as the spatial average of the event-total accumulated precipitation over the full analysis domain, on the common verification grid shared by both models and IMERG. ”

Response RC2

Summary

This manuscript presents a side-by-side, high-resolution (2 km) intercomparison of WRF and ICON for Medicane Daniel (September 2023), with a focus on sensitivity to the convection scheme (explicit, shallow-only, grayzone, and fully parameterized). The authors take care to match the two models as closely as possible (domain, vertical discretization, initial/boundary conditions, resolution, timestep) and introduce two appealing diagnostics: a Python algorithm to produce comparable vertical level distributions across the two coordinate systems, and the "Temporal Annular Symmetric Mean" (TASM) for characterizing the time-mean, axisymmetric warm-core structure.
The topic is timely and relevant, Daniel is an important and well-documented case, and the matched-configuration philosophy is a genuine strength. The TASM and the level-matching algorithm are worthwhile methodological contributions. However, before the paper can be accepted, several issues regarding the verification reference, internal numerical consistency, the fairness of the "convection-scheme" attribution, and the framing of the resolution regime need to be addressed. My main concerns are detailed below.

Response: We thank the Referee for the thorough review and for the positive assessment of the study design and of the methodological contributions. The comments led to substantive changes: the track statistics have been recomputed after the discovery of a processing error (G1/G2b), a significance assessment of the inter-configuration differences has been added (G3), the unmatched-physics limitation is now explicit and connected to the results (G4), the resolution terminology has been corrected (G5), and the conclusions have been restructured into case-specific findings, literature-consistent findings and open hypotheses (G6). Each point is addressed in detail below.

Changes in the manuscript: None for this point; see the individual replies.

General comment 1 – Reconcile track errors with Figure 4(c)

Section 4.1 reports track errors that are substantially smaller than those shown in Figure 4(c). For example, the text states that WRF–EXP has a mean error of 33 km and an RMSE of 39 km, and that ICON–CU has a mean error of 55 km and an RMSE of 65 km. In contrast, Figure 4(c) appears to show much larger values for the same simulations, including WRF–EXP with $\mu \approx 76$ km and $\sigma \approx 48$ km, and ICON–CU with $\mu \approx 113$ km. These values are not mutually consistent as currently presented. Since the ranking of the seven simulations by track skill is central to the paper's conclusions, the authors should clarify whether the text and Figure 4(c) refer to different metrics, time windows, smoothing procedures, or samples. If not, the relevant numbers should be corrected, and a single internally consistent set of track-error statistics should be used throughout the manuscript.

Response: We thank the Referee for spotting this inconsistency. In tracing its origin, prompted also by comment G2b, we found a genuine processing error: the displacement statistics reported in the text had been computed against a 6-hourly smoothed version of the observed track. The smoothing artificially reduced the reported errors. All track statistics have been recomputed against the 3-hourly HRV track, which is the correct reference; Fig. 4 has been regenerated, Table 3 has been re-verified, and a single consistent set of numbers is now used throughout the text, figures, tables and conclusions. The relative ranking of the configurations is unchanged: the WRF family attains the lowest errors, and ICON–SH remains the best ICON configuration. Sect. 3.3.1 now states correctly the observed HRV track.

Changes in the manuscript: In Sect. 4.1, all mean errors and RMSE values have been replaced with the recomputed ones `(lines 336-347)`

" The WRF–EXP simulation shows the best performance, with a mean error of ~~33~~63 km and a RMSE of ~~39~~76 km Fig.4(c). The cyclone is accurately tracked throughout its life cycle, with a slight eastward bias during the landfall phase. The WRF–SH simulation has a mean error of ~~40~~74 km and a RMSE of ~~50~~96 km. The cyclone is correctly tracked in the early phase, but it deviates significantly during the landfall phase, with a westward bias. The WRF–CU simulation has a mean error of ~~45~~76 km and a RMSE of ~~55~~86 km. The cyclone is poorly tracked in the early phase, with a northward bias, but the track improves during the landfall phase, with a slight eastward bias.
The ICON–EXP simulation has a mean error of ~~38~~98 km and a RMSE of ~~45~~110 km. The cyclone is well tracked in the early phase, but it deviates significantly from the observed track during the landfall phase, with a westward bias. The ICON–SH simulation has a mean error of ~~42~~83 km and a RMSE of ~~48~~97 km. The cyclone is poorly tracked in the early phase, with a northward bias, but improves during the landfall phase, with a slight eastward bias. The ICON–GZ simulation has a mean error of ~~50~~105 km and a RMSE of ~~60~~129 km. The cyclone is poorly tracked throughout its life cycle, with a significant northward bias. The ICON–CU simulation has a mean error of ~~55~~92 km and a RMSE of ~~65~~102 km. Again, the cyclone is poorly tracked in the early phase, with a northward bias, but it improves during the landfall phase, with a slight eastward bias. "

General comment 2 – Strengthen the observational reference for track and intensity

The observed track and the observed CSLP/wind time series rely heavily on Hérincs (2023), which appears to be a non-peer-reviewed report rather than an operational best-track product or a peer-reviewed observational dataset. Since the model track RMSEs, CSLP evolution and maximum-wind conclusions are evaluated against this reference, the observational basis needs to be strengthened. I recommend that the authors: (a) cross-validate the observed track and intensity against independent sources, such as ERA5, ECMWF operational analyses, ASCAT center/wind estimates, available station/ship/buoy data, and the diagnostics discussed in Flaounas et al. (2025); and (b) explicitly justify the adequacy of the manually selected SEVIRI-HRV cyclone center as the primary reference track. The manuscript states that this track has an uncertainty of about 15 km at 3-hourly resolution. Given that some reported model differences and landfall-position errors are only a few tens of km, the observational uncertainty may overlap with the inter-model differences used to rank the simulations. The authors should therefore include an observational uncertainty estimate, or at least discuss how this uncertainty affects the robustness of the model ranking.

Response: *(a) Cross-validation of the Hérincs (2023) estimates.* Hérincs (2023) reconstructs the pressure evolution from buoy, ship and land-station records (shown in our Fig. 3a), cross-checked against the UK Met Office analysis, and derives the wind estimates from station and satellite data. Following the Referee's suggestion, we cross-validated these estimates against ERA5 and a higher-resolution regional reanalysis CERRA. The comparison (Figs. 4 and 5) shows a consistent temporal evolution across datasets, with two systematic differences. First, the Hérincs pressures are slightly lower than the reanalysis values, while the absolute minimum is similar. Second, in Hérincs the minimum occurs offshore before landfall, whereas the reanalyses place it several hours later, over land. We consider the offshore timing more consistent with the expected behaviour of a marine warm-core cyclone: over the sea the vortex continues to extract latent and sensible heat from the surface, while the low surface roughness limits frictional spin-down; both supports are cut off at landfall. The land-based minimum in the reanalyses likely reflects their coarse effective resolution and the scarcity of in-situ MSLP data over the southern Ionian, which limit the analysed depth of a small-core vortex until it reaches the better-observed coast. For the wind, the reanalyses show a positive bias of up to $10\,\mathrm{m\,s^{-1}}$ with respect to Hérincs, but the temporal trend is consistent. On this basis we retain the Hérincs dataset as a qualitative reference for the temporal evolution, and we have added a brief version of this discussion to the revised manuscript.

(b) Adequacy of the SEVIRI-HRV track and impact of its uncertainty. This comment prompted us

to identify the processing error described under G1: a 6-hourly smoothed track had been used for the displacement statistics. After correcting to the 3-hourly HRV track, we validated it against the tracks extracted from ERA5 and CERRA, applying a 15 km buffer to represent the stated observational uncertainty (Fig. 6). The HRV track is consistent with the reanalysis tracks within this buffer for most of the cyclone lifetime, which supports its use as the primary reference. Regarding the impact of the uncertainty on the ranking: we report raw displacement errors and display the 15 km uncertainty band in Fig. 4; because the uncertainty applies equally to all configurations, it cannot reorder them, and we state this explicitly in Sect. 4.1. During the cyclogenesis phase the errors of several configurations fall within the band, and in that window the configurations are treated as observationally indistinguishable. The formal significance assessment of the inter-configuration differences requested in comment G3 complements this analysis.

Changes in the manuscript: The 15 km uncertainty is showed in Fig.4 and relative sentence is added in Sect. 3.3.2 `(lines 249-251)`

> " Since the uncertainty on the observed track applies equally to all simulations and through the track, it does not affect the error evaluation of the configurations, but it is still taken into account while evaluating the statistical significance of the results. "

General comment 3 – Assess statistical significance of inter-configuration differences

The manuscript ranks the seven simulations partly on the basis of track-error statistics, but the reported spread appears large relative to the differences between configurations. For example, the text reports mean errors of roughly 33–55 km, while Figure 4(c) appears to show larger μ values and substantial σ values. In either case, the separation between several configurations is small compared with the variability of the track error along the cyclone life cycle. It is therefore not clear whether the apparent differences between individual configurations, or between WRF and ICON as model families, are statistically or practically distinguishable.
Please add a robustness assessment of the track-skill ranking (statistical significance test?). At minimum, the authors should explicitly discuss whether the differences among configurations are robust enough to support statements such as "WRF performs slightly better than ICON" and "explicit and shallow cumulus schemes generally yield better results than fully parameterized schemes."

Response: To assess the true robustness of the track-skill ranking, we have added a dedicated serial-correlation-aware significance assessment.
The method deserves a short description, as a naive test assuming independent hourly pairs would severely overstate significance. We compare configurations through paired differences of the hourly displacement errors at common valid times. Consecutive hourly errors are strongly serially correlated because position errors persist over many hours; both the standard paired t-test and the Wilcoxon signed-rank test assume independent pairs, so raw p-values computed on the full hourly sample would be spuriously small.
We account for this serial correlation in two complementary ways:

- (i) Paired t-test with effective sample size: The test is applied using an effective sample size $n_{\text{eff}} = n(1-r_1)/(1+r_1)$, where r_1 is the lag-1 autocorrelation of the difference series s (Zwiers and Von Storch, 2004; Wilks, 2011).
- (ii) Subsampled Wilcoxon signed-rank test: As a distribution-free check, the Wilcoxon test is applied to differences subsampled at 12-hourly spacing to ensure temporal independence relative to the cyclone's evolution.

A difference is reported as significant only when both tests reject the null hypothesis at the 5% level ($\alpha = 0.05$).
The outcome of this serial-correlation-aware analysis is as follows:

- Sample Size & Autocorrelation: With $n = 60$ common hourly samples and lag-1 autocorrelations (r_1) of the difference series ranging between 0.466 and 0.872, the effective sample sizes range between $n_{\text{eff}} = 4.1$ and $n_{\text{eff}} = 21.8$.
- Pooled WRF vs. ICON Families: The difference between the pooled WRF-family and pooled ICON-family series ($-23.34\,\text{km}$) is not significant ($t = -1.224$, $p = 0.2847$ for the n_{eff}-corrected t-test; $p = 1.0000$ for the 12-hourly Wilcoxon test).
- Pairwise Comparisons: Among the pairwise comparisons, no pair was found to be statistically significant once serial correlation was accounted for; the remaining pairs are not distinguishable from baseline variability.
- Observational Uncertainty Benchmark: As a secondary, phase-resolved result, the mean displacement error of every configuration exceeds the 15 km observational uncertainty outside the cyclogenesis phase (one-sample t-test with the same n_{eff} correction), which quantifies where in the life cycle the models lose the track.

Changes in the manuscript: Added Section "Robustness Assessment of tracks error" in Supplementary materials with the workflow above
Modified statement "WRF performs slightly better than ICON" `(lines 350-352)`

> " While WRF exhibited a lower mean track error compared to ICON, rigorous significance testing accounting for serial correlation demonstrates that this difference is not statistically significant ($p = 0.28$), being largely overshadowed by the substantial variability along the cyclone life cycle "

Modified statement "explicit and shallow cumulus schemes generally yield better results than fully parameterized schemes." `(lines 355-358)`

> " Although explicit and shallow cumulus schemes yielded lower absolute mean errors in our sample compared to fully parameterized schemes, these improvements are statistically indistinguishable once serial correlation is accounted for. Therefore, a strict ranking of individual configurations based solely on track error cannot be robustly supported for this specific event. "

General comment 4 – The "convection-scheme sensitivity" attribution is confounded by unmatched physics

The manuscript aims to quantify sensitivity to convective treatment and to compare WRF and ICON under configurations that are "as comparable as possible". This is a valuable objective. However, the cross-model interpretation should be more cautious. While convection treatment is varied within each model, other physical parameterizations are not matched between WRF and ICON. In particular, Table 1 shows different microphysics schemes — WRF WDM6 versus ICON Seifert–Beheng two-moment — and different boundary-layer/turbulence treatments — WRF YSU + Smagorinsky versus ICON prognostic TKE. These differences are not secondary for a medicane-like system: microphysics–convection interactions, boundary-layer moisture transport, turbulent mixing and surface fluxes can directly affect intensity, precipitation distribution, latent heating, vertical warm-core structure and track evolution.
Therefore, systematic results such as WRF producing larger accumulated precipitation, deeper CSLP/stronger 10 m winds, or a deeper/more vertically coherent warm core than ICON cannot be attributed uniquely to convection-scheme treatment. They may partly reflect the combined model-physics package. I am not asking for a full additional matrix of microphysics/PBL experiments, but the limitation should be made explicit. The authors should separate conclusions about within-model convection sensitivity from conclusions about WRF–ICON differences, discuss the likely influence of microphysics and PBL/turbulence choices on the headline results, and temper statements that imply that the model-family differences are primarily controlled by convection parameterization alone. (I am not asking for a full additional experiment matrix, but the limitation must be made explicit and

its likely influence on the results discussed. I have seen myself huge differences in precipitation fields and tracks in medicanes, among different MP physics.)

Response: We fully agree, and we have restructured the presentation accordingly. Microphysics alone can displace medicane tracks and reshape rainfall fields substantially, as the Referee notes from direct experience and as documented in the sensitivity literature (Miglietta et al., 2015; Ricchi et al., 2017); nothing in our experimental design excludes such contributions to the cross-model contrast. Three changes implement the request. First, a dedicated paragraph in Sect. 3.1 now describes each non-harmonisable physics difference (WDM6 vs Seifert–Beheng two-moment microphysics, RRTMG vs ecRad radiation, YSU+Smagorinsky vs prognostic-TKE turbulence, Noah vs tiled TERRA surface, the two gravity-wave-drag formulations, KSAS vs Tiedtke–Bechtold convection) and identifies them as an additional, model-dependent source of uncertainty. Second, the Discussion now connects these differences to the headline results, so that the reader is warned where the confounding is most likely to act. Third, all cross-model statements in the revised text are phrased as differences between full model configurations; causal statements about the convection scheme are restricted to the within-model sensitivity experiments, where the rest of the physics is held fixed.

Changes in the manuscript: Added paragraph in Sect. 3.1 `(lines 158-177)`

> " Despite the effort to balance the configurations for the two simulations as much as possible, the physical parameterization suites of WRF and ICON could not be fully harmonized, reflecting the independent development lineages of the two modelling systems (Table 1). For each physics category we retained each model's standard, operationally validated configuration rather than forcing an artificial correspondence that neither code is designed to support. The most consequential divergence concerns cloud microphysics: the WDM6 scheme (Lim and Hong, 2010) predicts number concentration only for the warm-rain species, retaining a single-moment treatment for the ice-phase hydrometeors, whereas the two moment scheme (Seifert and Beheng, 2006) used in ICON prognoses both mass and number concentration across all hydrometeor categories, including the ice phase. Radiative transfer is likewise computed by conceptually distinct codes, with RRTMG (Iacono et al., 2008) following a correlated-k approach, in contrast to the more modular, multi-solver architecture of ecRad (Hogan and Bozzo, 2018). Boundary-layer turbulence differs in closure order: the non-local, first-order YSU scheme (Hong et al., 2006), paired with a first-order Smagorinsky closure for residual 3D mixing in WRF, compared to ICON's prognostic, 1.5-order TKE-based scheme (Raschendorfer, 2001). Surface exchanges are computed by the single-tile Unified Noah land surface model in WRF (Niu et al., 2011), as opposed to the explicitly tiled TERRA scheme in ICON (Schrodin and Heise, 2001), which resolves sub-grid land-cover heterogeneity relevant to the complex coastlines of the Mediterranean basin. Gravity wave drag is included in both configurations, though following distinct orographic formulations (Choi and Hong, 2015; Orr et al., 2010). Finally, convection is parameterized by mass-flux schemes of different heritage: the KIM Simplified Arakawa Schubert scheme in WRF (Han and Pan, 2011; Kwon and Hong, 2017), versus the Tiedtke–Bechtold scheme in ICON (Tiedtke, 1989; Bechtold et al., 2008), the latter treating shallow and deep convection as explicitly separate regimes. These structural differences, intrinsic to the two model architectures, represent an additional, model-dependent source of uncertainty that should be considered alongside the dynamical and resolution-related aspects of the comparison discussed above. "

Added sentences in Discussion and Conlusions `(lines 570-574)`

> " Part of the systematic WRF–ICON contrast may stem from the unmatched physics rather than from the convective treatment. The single-moment ice phase in WDM6, as opposed to the fully two-moment Seifert–Beheng scheme, plausibly alters the vertical distribution of latent heating and hence the warm-core depth diagnosed in the TASM; the non-local YSU mixing, compared with ICON's prognostic-TKE closure, modulates the boundary-layer moisture supply and the near-surface wind maxima. Cross-model statements in this study should hence be read as differences between full model configurations, whereas causal statements about the convection scheme are restricted to the within-model experiments. "

General comment 5 – Clarify and justify the "grayzone" framing for ~ 2 km

The manuscript uses ~ 2 km grid spacing and includes an ICON–GZ experiment described as a "grayzone" configuration. However, the terminology needs to be clarified. In much of the current convection-permitting modelling literature, horizontal grid spacings around 2 km are usually treated as convection-permitting, while the deep-convection gray zone is more commonly associated with somewhat coarser grid spacings, where deep convective updrafts are only partly resolved. At the same time, 2 km is still not convection-resolving in a strict sense, especially for shallow convection, turbulent transport and individual convective updraft dynamics. Therefore, the manuscript should define explicitly what is meant by "gray zone" in this study.
This clarification is particularly important because the selected WRF cumulus scheme, KSAS, is itself a scale-aware mass-flux scheme developed for gray-zone applications. Thus, the contrast between "explicit", "parameterized" and "grayzone" configurations is not as clean as the current wording suggests. Please distinguish clearly between: (i) the grid-spacing regime of the simulations; (ii) the physical gray zone of partially resolved convection; and (iii) the specific ICON grayzone tuning used in ICON–GZ. The operational recommendations should then be revised so that they follow consistently from this definition, without implying that all 2 km simulations fall into the same gray-zone category or that the explicit/parameterized distinction is unambiguous at this resolution.

Response: We adopted the three-way distinction requested by the Referee, which indeed removes an ambiguity that ran through the original text. In the revised manuscript: (i) the 2 km grid spacing is classified as convection-permitting, not as a deep-convection gray zone, which is associated with coarser spacings (typically 3–10 km); (ii) we state that 2 km remains inside the physical gray zone for shallow convection, boundary-layer turbulence and individual updraft dynamics, which are only partially resolved; (iii) ICON–GZ is described strictly as a specific NWP tuning of the ICON deep-convection scheme, not as a statement about the resolution regime. A sentence has also been added noting that KSAS is itself a scale-aware mass-flux scheme, so the "fully parameterized" label in WRF–CU refers to the activation of both deep and shallow components rather than to a scale-unaware treatment. The operational recommendations in Sect. 5 have been reworded to follow from this definition, and the residual "grayzone" wording in the Conclusions has been replaced with "convection-permitting". Concerning the domain and resolution choice: with ≈9 km lateral boundary conditions, the single 2 km domain keeps the grid-ratio jump at ≈1:4.5, avoiding both the boundary artefacts of a direct downscaling to 1 km and the interpolation cascade of intermediate nests (Miglietta et al., 2023); a matched multi-nest setup is in any case precluded by the differing nesting constraints of the two models (odd integer ratios in WRF, powers of 2 in ICON). Kilometre-scale explicit-only convection lies outside the scope of this work and exceeds the computational budget of most operational implementations.

Changes in the manuscript: Added paragraph in Sect. 3.1 `(lines 127-135)`

> " The choice of a single domain with a grid spacing of ≈ 2 km was carefully evaluated. Given the ≈ 9 km LBCs, a direct downscaling to 1 km would introduce severe boundary reflections, while intermediate nests introduce cascading interpolation artifacts and complex dynamical noise (Miglietta et al., 2023). Furthermore, differing structural nesting ratio constraints between the models (e.g., odd integers in WRF versus powers of 2 in ICON) would produce unmatched parent domains, compromising a clean model intercomparison. Physically, ≈ 2 km represents an optimal intermediate "sweet spot" between coarse parameterizations and fully resolved ones (Brumer et al., 2026). The grid-spacing regime, ≈ 2 km is classified as convection-permitting rather than a classic deep-convection gray zone (typically 3–10 km); nevertheless, it sits squarely in the physical gray zone for shallow convection, boundary-layer turbulence, and individual updraft dynamics, which remain only partially resolved. "

General comment 6 — single-case scope vs. generalised conclusions

The manuscript acknowledges that the analysis is based on a single event, but several conclusions

are phrased more generally than the experimental design can support. Statements such as "the inclusion of a shallow convection parameterization proved to be important for both models" or that WRF may be "better tuned for the diabatic processes dominating TLC maintenance" should either be softened to apply specifically to Daniel under the present configuration, or more explicitly contextualized within the broader medicane sensitivity literature.
This is particularly important because the Introduction itself notes that there is no consensus on which model configuration performs best for TLCs, and that previous studies have shown strong sensitivity to microphysics, PBL schemes, initialization time, forcing dataset and cumulus parameterization. I therefore recommend that the authors revise the conclusions to distinguish clearly between: (i) findings that are robust within this Daniel case study; (ii) findings that are consistent with previous medicane studies; and (iii) hypotheses that remain to be tested across additional cases or initialization times. In particular, the authors should discuss where Daniel agrees with, or departs from, prior results in Ricchi et al. (2017, 2019), Pytharoulis et al. (2018), Miglietta et al. (2015), and Saraceni et al. (2023). Without this contextualization, operational recommendations about shallow convection or model-family performance should be phrased as case-specific rather than general guidance.

Response: We agree that the experimental design supports case-specific statements only, and we revised the Conclusions along the three-tier structure suggested by the Referee. The general statements flagged in the comment have been restricted to the present case: "The inclusion of a shallow convection (SH) parameterization proved to be important for both models" now reads "[. . .] in this experiment proved to be important for both models [. . .] at least for the Daniel case", and "better tuned for the diabatic processes dominating TLC maintenance" now reads "dominating TLC Daniel maintenance in these simulations". In addition, a new Discussion paragraph places the results in the context of the studies listed by the Referee, stating where Daniel agrees with and departs from them. The interaction with comment G3 is also relevant here: where the significance tests indicate overlapping configurations, the corresponding ranking statements are presented as indicative for this case.

Changes in the manuscript: The softened sentences above are in the revised Conclusions `(lines 509-513)`
Added paragraph in Discussion `(lines 555-562)`

> " The present findings sit consistently within the medicane sensitivity literature. The first-order control of the convective treatment on track and intensity confirms the single-case results of (Miglietta et al., 2015), and the benefit of a dedicated shallow convection treatment at kilometre scale agrees with (Saraceni et al., 2023). The large spread among configurations, with no single setup performing best across all phases and metrics, mirrors the multi-physics results of (Ricchi et al., 2017, 2019; Pytharoulis et al., 2018). Daniel departs from most previously studied events in its prolonged, multi-stage evolution, which stresses model physics across heterogeneous regimes within a single integration. Accordingly, the operational indications given below are to be read as case-specific: findings established for Daniel under the present configurations, findings consistent with prior studies, and hypotheses that remain to be tested across additional cases and initialization times. "

Specific comment 1

Section 3.5.1 applies Hart's original Cyclone Phase Space methodology, using a circular radius of $R = 300$ km around the cyclone centre. Please justify why the original Hart CPS framework was selected for Daniel rather than a medicane-specific or medicane-adapted approach, such as that discussed in Miglietta et al. (2025), which is already cited in the manuscript. [. . .] The authors should clarify whether the chosen CPS setup is appropriate for identifying the warm-core and symmetry characteristics of a compact medicane, and explain how the interpretation of the CPS diagrams is affected by using the original Hart framework rather than a medicane-specific adaptation.

Response: We acknowledge that medicane-adapted CPS variants exist: smaller radii are frequently employed (Miglietta et al., 2025), and Picornell et al. (2014) propose a 100 km radius with the 925–700–400 hPa layers in place of the standard 900–600–300 hPa. Miglietta et al. (2025) also note that the application of differing criteria and thresholds has produced inconsistent medicane lists, which points to a real sensitivity of the diagnosed structure to the chosen configuration. To quantify this sensitivity for our case, we ran a dedicated test on one simulation (ICON–SH), computing the CPS parameters for radii of 100, 150, 200, 250, 300, 400 and 500 km and for both layer configurations (Figs. 7 and 8). The exact numerical values of B, $-V_T^L$ and $-V_T^U$ fluctuate considerably between setups: the Mean Absolute Percentage Differences reach 277–312 % across layer sets and 117–164 % across radii, with the percentages inflated where the parameter values approach zero. The structure of the distributions, however, changes little: effect sizes remain small throughout (mean Cohen's d between -0.28 and 0.35). We note that the CPS parameter time series are themselves serially correlated, so the nominal significance counts of the pairwise comparisons overstate the distinguishability of the setups, which reinforces the effect-size reading. Against the full ensemble of tested configurations, the adopted setup (300 km, standard layers) is close to the ensemble centre: mean z-scores of -0.12, -0.26 and -0.33 for B, $-V_T^L$ and $-V_T^U$, an average absolute z-score of 0.24 and an average Cohen's d of 0.19. We retain the original Hart framework because it captures the phase evolution of Daniel without anomalous or skewed results while preserving direct comparability with the synoptic climatology and with most of the existing TLC literature. Finally, we clarify the coloured thresholds in the CPS diagrams: they are indicative contours, qualitatively inspired by the SSHWS, used only as a visual aid for comparing simulations; they carry no quantitative classification value, since a rigorous medicane-specific calibration would require a large sample of high-resolution simulations that does not currently exist.

Changes in the manuscript: A summary is added to Sect. 3.5.1 `(lines 305-311)`

> " Medicane-adapted CPS variants employ smaller radii and shifted layers (e.g., 100 km and 925–700–400 hPa (Picornell et al., 2014; Miglietta et al., 2025)).A sensitivity test on ICON–SH, varying the radius between 100 and 500 km and testing both layer sets, shows that the numerical values of B, $-V_T^L$ and $-V_T^U$ fluctuate between setups while the effect sizes remain small (mean Cohen's d between -0.28 and 0.35), and that the adopted configuration (300 km, standard layers) lies near the centre of the tested ensemble (mean absolute z-score 0.24). The original Hart framework is retained for comparability with the existing literature on tropical and tropical-like cyclones. The coloured thresholds in Figs. 8-9 are indicative, qualitatively inspired by the SSHWS, and serve only as a visual aid for comparing simulations. "

Specific comment 2

TASM axisymmetry assumption (Sect. 3.5.2). The TASM is a useful idea, but the manuscript itself notes that TLC symmetry is poorly defined. Please (a) quantify the degree of asymmetry being averaged over (e.g., azimuthal variance), and (b) discuss how the 15 km center-position uncertainty propagates into the annular means.

Response: *(a)* We quantified the asymmetry through the azimuthal standard deviation $\sigma_\phi(r, z)$ of θ_e for each radius and vertical level (Eq. A1 in the revised Appendix). In the inner core, σ_ϕ generally remains below 2 K; it increases outward, with a localised maximum near 180 km at 500 hPa, and exceeds 3 K beyond a 250 km radius, where the symmetric structure degrades rapidly (Fig. 9). The axisymmetry assumption underlying the TASM is hence well supported within the inner core (up to 200–250 km), which is the region of interest for the warm-core diagnosis; outside it, the averaging must be read as a mean over an increasingly asymmetric field.

(b) We tested the propagation of the centre-position uncertainty directly: the tracked centre of the reference simulation (WRF–SH) was displaced by 0.14° (≈15 km) in the four cardinal directions and the TASM recomputed for each displacement (Fig. 10). The 2D spatial correlation between the displaced and the reference TASM fields lies between 0.9990 and 0.9992. The azimuthal averaging

damps the localised noise introduced by small misplacements, so a 15 km centre error does not propagate appreciably into the annular means. We note that the 15 km figure refers to the observed track, while the TASM is computed on model fields with the model-derived centre; the test shows that the diagnostic is insensitive to centre errors of this magnitude regardless of their origin.

Changes in the manuscript: A new Appendix subsection "TASM axisymmetry and displacement" has been added, containing the definition of σ_ϕ (Eq. A1), the asymmetry map (Fig. A7) and the displacement test (Fig. A8), with the quantitative statements above.

> " To assess the degree of asymmetry averaged over the annular domains, we calculated the azimuthal standard deviation of the equivalent potential temperature (θ_e) for each radius and vertical level.
> The target control (CTRL) simulation is the WRF–SH one, as it provides the best overall result. The azimuthal mean is defined as:
>
> $$\overline{\theta_e}(r,z) = \frac{1}{N}\sum_{i=1}^{N}\theta_e(r,\phi_i,z)$$
>
> Consequently, the azimuthal dispersion is computed as:
>
> $$\sigma_\phi(r,z) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\theta_e(r,\phi_i,z) - \overline{\theta_e}(r,z)\right]^2} \quad (1)$$
>
> where r is the radial distance from the cyclone center, z is the vertical level, ϕ_i represents the i-th azimuthal point within the annulus, and N is the total number of grid points within the annular ring. The parameter $\sigma_\phi(r,z)$ provides a quantitative, localized estimate of the field's asymmetry.
> The axisymmetry assumption holds remarkably well in the inner core of the cyclone. In this central region, σ_ϕ generally remains below 2 K Fig. A7. As expected, the asymmetry naturally increases moving radially outward. Specifically, we identify a localized maximum of asymmetry at approximately 180 km from the center at the 500 hPa level. Beyond a radius of 250 km, the σ_ϕ values exceed 3, indicating a rapid degradation of the symmetric structure. Therefore, we can confidently state that the axisymmetry assumption underlying the TASM is physically robust and well-justified within the inner core of the medicane (up to a radius of 200–250 km).
> To address the propagation of the 15 km positional uncertainty into the annular means, we performed a dedicated spatial sensitivity test on the CTRL simulation dataset. The tracked center of the cyclone was artificially displaced by 0.14° (approximately 15 km) in the four cardinal directions (North, South, East, and West). For each of these displaced configurations, we recalculated the TASM and compared it against the original, centered CTRL output Fig A8.
> The statistical comparison reveals that the 2D spatial correlation between the displaced TASM fields and the original reference field is exceptionally high, ranging strictly between 0.9990 and 0.9992.
> Therefore, we can state that a center-position uncertainty of 15 km does not significantly propagate into the annular means. The azimuthal averaging effectively dampens the localized spatial noise caused by minor misplacements. Consequently, the TASM calculation proves to be highly resilient to typical tracking algorithm uncertainties, guaranteeing that the resulting thermal anomaly profiles are physically reliable and not artifacts of tracking precision. "

Specific comment 3

FSS description and interpretation (Sect. 3.4 / 4.2). The verbal description of FSS behavior around lines 330–333 is imprecise and at one point conflates "more cells than the observations" with low FSS in a way that does not match the symmetric nature of the score. Please restate the FSS interpretation precisely and add a "useful scale" criterion (e.g., the FSS $\approx 0.5 + f_0/2$ target of Roberts and Lean, 2008), which would make the skill claims more rigorous and easier to compare across thresholds.

Response: The Referee is right: the original wording conflated fractional-coverage mismatch with over-forecasting, which the symmetric score cannot separate. We revised the FSS description in Sect. 3.4, clarifying the symmetric nature of the score, and introduced the useful-scale criterion of Roberts and Lean (2008), with the target $FSS_{useful} = 0.5 + f_0/2$; since f_0 is very small for the upper percentile thresholds, the target reduces to $FSS_{useful} \approx 0.5$ there, and we report the corresponding f_0 values. Sect. 4.2 has been rewritten to evaluate the simulations through the minimum window size at which each configuration crosses the useful-scale target, in place of the previous cell-count description.

Changes in the manuscript: Sect. 3.4 now defines the criterion `(lines 282-293)`

> " ~~Hence, a perfect forecast is represented by the FSS approaching 1; low values represent a poor forecast. To be pointed out, the FSS strongly depends on the size of the window, if high values for FSS are obtained for small window, then the model performs well. However, high FSS values associated with large windows suggest a poor localization of the precipitation but a good accumulation value.~~ The FSS evaluates the spatial overlap of fractional coverage between model and observations, ranging from 0 (complete mismatch) to 1 (perfect forecast). Because the score is symmetric, a low FSS value indicates a general mismatch in fractional coverage, which can stem equally from spatial displacement, over-forecasting, or under-forecasting. To rigorously assess the spatial scale at which a model becomes skilful, we employ the "useful scale" criterion introduced by (Roberts and Lean, 2008) A forecast is considered skilful when the FSS exceeds a target value defined as $FSS_{useful} = 0.5 + f_0/2$, where f_0 is the domain-average fractional coverage of the observed event. Since f_0 is inherently very small for localized extreme precipitation (such as the upper percentiles analysed here), the target threshold is effectively $FSS_{useful} \approx 0.5$. The minimum window size required to reach this value represents the spatial scale at which the model provides a reliable precipitation forecast. "

Sect. 4.2 has been rewritten accordingly `(lines 424-448)`

> " ~~The FSS is now used for a quantitative analysis of the simulations. The FSS highlights the peculiarities of each simulation. If the size of the window is small and the FSS is high, the model agree well with the observations. As the window size increases, so does the number of grid cells that exceed the accumulation threshold; if FSS is close to one, the number of cells between simulation and observation is similar; if FSS is lower, the simulation has more cells than the observations. At low thresholds, all models have good scores (Fig. 7 (a)(b)(c)(d)(e)(f)(g) black and blue lines). The first differences are noticeable from the 85th percentile, where the best results are achieved by the fully parameterized ICON–CU scheme (Fig. 7 (g) green line). On the contrary, rising to the 95th percentile threshold, the best result for WRF is in the SH configuration, but ICON–EXP performs better at small window (Fig. 7 (b) and (c) yellow line respectively). The 99th and 99.9th thresholds show the most dramatic differences among the experiments: WRF–SH is the only one achieving FSS $>$ 0 on the 1x1 grid at both a threshold of 99% and 99.9% (Fig. 7 b red and violet lines), indicating a good localization of extreme precipitation accumulations. Conversely, ICON–CU yields the lowest FSS at the same threshold (Fig. 7 (g) violet line), indicating that there are much more cells exceeding the threshold. ICON–SH and ICON–EXP show a sharp increases and the highest FSS values at larger window size (Fig. 7 (c) and (f), red line), resulting in low displacement error and good amount of precipitation for extremes. On the overall, ICON–EXP attains the highest FSS values.~~
> The FSS is now used for a quantitative analysis of the simulated precipitation fields. Assessing the FSS as a function of the neighbourhood window size highlights the specific spatial capabilities of each simulation. High FSS values at small window sizes indicate an accurate spatial localization of the precipitation. Conversely, when the FSS only rises at larger window sizes, it implies that the model captures the overall precipitation frequency but suffers from significant displacement errors at the grid scale. At low thresholds, all models exhibit high skill scores, rapidly exceeding the $FSS_{useful} \approx 0.5$ criterion even at small spatial scales (Fig. 7a-d, black and blue lines). Noticeable differences emerge starting from the 85th percentile, where the fully parameterized ICON–CU scheme achieves the best performance at finer scales (Fig. 7g, green line). However, as the threshold rises to the 95th percentile, the best grid-scale result for WRF is obtained with the SH configuration, while ICON–EXP demonstrates the highest

skill among the ICON simulations at small window sizes (Fig. 7b and 7e, yellow line). The 99th and 99.9th percentiles highlight the most dramatic discrepancies among the experiments. WRF–SH is the only configuration achieving an FSS > 0 at the native grid scale (1×1) for both extreme thresholds (Fig. 7b, red and violet lines), indicating some exact spatial overlap for extreme local accumulations. Conversely, ICON–CU yields the lowest overall FSS at these extremes (Fig. 7g, violet line), reflecting a severe mismatch in fractional coverage. ICON–SH and ICON–EXP, while exhibiting displacement errors at the native grid scale, show a sharp increase in skill at larger window sizes, ultimately attaining the highest overall FSS values (Fig. 7e and 7f, red lines). This indicates that although these setups struggle to pinpoint the exact location of extreme cells, they successfully reproduce the correct fractional coverage of extreme precipitation over broader areas, reliably crossing the FSS_{useful} threshold at synoptic scales. Overall, taking both peak accumulations and spatial scales into account, ICON–EXP attains the most robust FSS performance.
Examining the minimum skillful ($FSS \geq 0.5$) window sizes across configurations and thresholds reveals that nearly all simulations achieve skill at the grid scale (1 grid-cell) up to the 85th percentile, whereas none reach skillfulness at the 99.9th percentile threshold. The main differences emerge at the 99th percentile threshold, where ICON–EXP achieves skillfulness at the smallest window size (19 grid-cells), followed by WRF–CU (41 grid-cells), WRF–EXP and WRF–SH (49 grid-cells), ICON–SH (67 grid-cells), and ICON–GZ (85 grid-cells). ICON–CU fails to reach an FSS threshold of 0.5. ”

Specific comment 4

Re-initialization experiment (lines 114–118). The three 72-h re-initialized runs and their degradation (attributed to hydrometeor imbalance and energy-budget) are interesting but presented without any figure or quantification. Please either show supporting evidence or present this as a brief methodological note without the strong claim.

Response: We followed the second option. The mechanistic attribution (hydrometeor imbalance and energy-budget arguments) exceeded what our diagnostics can support, and it has been removed; the revised text presents the outcome as a brief methodological note. For the Referee's reference, supporting figures are attached to this reply: they show the track (Fig. 11) and CSLP evolution (Fig. 12) of the three 72-h re-initialised runs against the continuous run, with a visible degradation of both after each re-initialisation, consistent with spin-up effects. Two snapshots of the 10 m wind speed field at 08-September 17:00 UTC and 10-September 00:00 UTC (Figs. 13 and 14) illustrate the spatial degradation of the vortex structure in the re-initialised runs.

Changes in the manuscript: The sentence in Sect. 3.1 has been updated `(lines 123-126)`

> “ ~~highlighting how the re-initialization of variables, such as hydrometeors, causes imbalances in the energy budget, with a consequent reduction in the models' ability to develop a TLC~~
> probably due to the spin-up (cold start) of each re-initialised run; a dedicated analysis is left for future work. ”

Specific comment 5

Observed CSLP reliability vs. CSLP rankings (lines 294–300). The discussion of the 996 hPa observed minimum possibly being an overestimate is welcome and honest. However, the CSLP skill rankings are then made against this admittedly unreliable value. Please add a corresponding caveat to the CSLP comparisons, or use a reanalysis-based intensity estimate as a secondary reference.

Response: We would like to clarify our approach on this point. Because of the limited quantitative reliability of the observed values, the considerations in Sect. 4.2 are qualitative, concerning the temporal evolution of the pressure and wind fields rather than absolute values. The skill ranking discussed in the manuscript is not based on the difference between simulated values and the observed 996 hPa minimum: it is a direct inter-comparison between the WRF and ICON configurations.

Lines 313–317, for instance, make no reference to the observed values. That said, we agree that the reader should not have to infer this, and we added an explicit caveat to Sect. 4.2. The reanalysis-based cross-validation added in response to comment G2a now also serves as the secondary intensity reference the Referee suggests.

Changes in the manuscript: Added at the end of the CSLP discussion in Sect. 4.2 `(lines 396-397)`

> " The observed CSLP minimum is used only as a qualitative reference for the temporal evolution; the quantitative statements in this section only compare the simulations with one another. "

Specific comment 6

Please specify which IMERG run (Early/Late/Final) and version were used. This affects the high-percentile (99th, 99.9th) FSS results that several key precipitation conclusions depend on.

Response: We used the IMERG "Final" run, version V07, satellite-gauge merged, half-hourly, on the $0.1° \times 0.1°$ grid. This is now stated in Sect. 3.2, and the full dataset citation with DOI is in the Data Availability section.

Changes in the manuscript: Added sentence in Sect. 3.2 `(lines 220-224)`

> " The dataset refers to the "final" run v07 satellite-gauge product merged into half-hourly $0.1° \times 0.1°$ (roughly 10x10 km nominal spatial resolution) fields.
> IMERG Final incorporates gauge information only through a monthly climatological adjustment and underestimated the observed maxima over Thessaly during this event (Flaounas et al., 2025); model overestimates diagnosed over Greece against IMERG should hence be read as upper bounds. "

Specific comment 7

Radius inconsistency for wind extraction. Section 3.4 states the maximum wind is extracted within 150 km of the center, whereas the Figure 5 caption states 100 km. Please reconcile.

Response: The extraction radius used in the processing code is 150 km, consistent with Sect. 3.4 and with the radius used by the tracking constraint; the Fig. 5 caption erroneously stated 100 km and has been corrected.

Changes in the manuscript: The Fig. 5 caption now reads "within a radius of 150 km from cyclone center".

Specific comment 8

Title of Paragraph 4.3 is "Physics", which does not accurately describe its content (an analysis of the cyclone's tropical characteristics and warm-core structure via CPS and TASM, rather than the model physics/parameterizations covered in Sect. 3.1). Consider retitling, e.g., "Tropical-like structure" or "Physical structure of the cyclone", for consistency with Sect. 3.5.

Response: We adopted the Referee's suggestion.

Changes in the manuscript: Sect. 4.3 is retitled "Physical structure of the cyclone".

Specific comment 9 — "dramatic lack of observations" vs. the NOANN network

The manuscript describes a "dramatic lack of ground-based observations" (Sect. 3.2). While accurate for the offshore Ionian and Libyan phases, this overstates the situation over Greece during the cyclogenesis/extratropical phase, where the dense National Observatory of Athens automatic network (NOANN/METEO) (Lagouvardos et al., 2017) provides high-resolution surface and rainfall observations — data used by several Daniel studies the authors already cite. Please consider incorporating NOANN gauge data (at least for the Greek extratropical phase) to provide independent ground truth for the precipitation verification, which currently relies on satellite-derived IMERG even though the cited literature documents IMERG biases over Thessaly for this event. At minimum, the "lack of observations" statement should be qualified by phase and region. In Line 74, the precipitation sums mentioned, come from the NOANN network. Please use either the Lagouvardos et al. 2017 citation or Flaounas et al., 2025 one.

Response: The Referee is right that the original statement overstated the data scarcity over Greece, and the revised Sect. 3.2 now qualifies it by phase and region, crediting the NOANN/METEO network explicitly. On the use of the gauges as an independent verification reference, we considered the option and retained IMERG as the single reference for the precipitation evaluation, for three reasons: the verification domain is mostly sea, where no gauge network exists, and a homogeneous reference is needed for a fair comparison across the whole event; IMERG Final incorporates gauge information, though only through a monthly climatological adjustment; and a direct gauge-to-grid comparison at kilometre scale introduces point-to-area representativeness problems that the gridded product avoids. At the same time, we acknowledge the point the Referee raises about the quality of the reference over Thessaly: IMERG underestimated the observed maxima there during this event (Flaounas et al., 2025). We added this limitation to Sect. 3.2, so that the model overestimates diagnosed over Greece against IMERG are explicitly presented as upper bounds. A dedicated gauge-based verification of the Greek extratropical phase is a natural follow-up, but it requires quality control and representativeness analysis of the station data that exceed the scope of this revision. The Line-74 citation has been added as requested (Lagouvardos et al., 2017 for the network, Flaounas et al., 2025 for the totals; see also our reply to RC1, Minor comment 1).

Changes in the manuscript: Modifications to Sect. 3.2 `(lines 211-224)`

> " ~~Because of the dramatic lack of ground-based observations, additional insight into the storm's structure and intensity was obtained through satellite-derived precipitation estimates.~~
> During the genesis and extra-tropical phases, the dense National Observatory of Athens automatic network (NOANN/METEO) (Lagouvardos et al., 2017) provides high-resolution surface and rainfall observations. However, offshore over the open Ionian Sea and in Libya, there is a dramatic lack of ground-based observations. For the sake of consistency, we have decided to use satellite-derived precipitation estimates for the evaluation over the whole domain. "

Specific comment 10

Figure 4: panels (a) and (b) rely on an eight-way color code that is defined only in the caption, with no in-figure legend. Several of the assigned colors are difficult to distinguish, making the overlaid tracks (a) and stacked-bar timeseries (b) hard to interpret. Please add an explicit legend to the figure and, if possible, adopt a more clearly separable color palette. Ensure color assignments are consistent across panels (a)–(c) and the corresponding text.

Response: Figure 4 has been regenerated with an in-figure legend and a more separable, colour-blind-safe palette (ICON–SH in green, ICON–GZ in magenta, ICON–CU in light blue; WRF colours unchanged). The assignments are consistent across panels (a)–(c), the caption and the text, and the same palette has been propagated to all figures that use the simulation colour code (Figs. 5–10). Panel (a) additionally shows the 15 km uncertainty buffer around the observed track (see G2b).

Changes in the manuscript: Fig. 4 and its caption have been replaced; the colour references in Sect. 4.1 have been updated accordingly.

> " WRF-EXP in red, WRF-SH in orange, WRF-CU in yellow; ICON-EXP in blue, ~~ICON-SH in petrol-green, ICON-GZ in cyan, ICON-CU in blue-grey, observed track is the black line.~~ ICON-SH in green, ICON-GZ in magenta, ICON-CU in light-blue, observed track is the black line with a shadowed buffer indicating the estimated 15 km error. "

Specific comment 11

Line 248: "The cyclone structure is typically analysed using vertical cross-sections of θe." The word "typically" asserts a methodological convention without support. Since this claim motivates the introduction of the TASM, please provide citations establishing that vertical θe cross-sections are the conventional diagnostic for (tropical-like) cyclone warm-core structure, or rephrase to avoid the unsupported generalization.

Response: We rephrased the sentence and added supporting citations.

Changes in the manuscript: Updated sentence `(lines 317-319)`

> " ~~The cyclone structure is typically analysed using vertical cross-sections of θ_e.~~ Vertical cross-sections of θ_e are often employed to inspect the moist thermodynamic and warm-core structure of tropical and tropical-like cyclones (Emanuel, 2018; Miglietta and Rotunno, 2019) "

Specific comment 12

The manuscript repeatedly invokes operational use to justify model and resolution choices "widespread use in... operational forecasting" (line 45) and "operationally this is approximately the resolution used by both models" (line 108) but never specifies which centers run WRF and ICON operationally, at what resolution, domain, and convection configuration. Please name the relevant operational forecast suites (e.g., the DWD ICON convection-permitting systems and the specific WRF deployments intended) with citations, and confirm that 2 km is genuinely representative of operational practice for both models rather than one. This is important because the 2 km choice and the operational recommendations in Section 5 rest on this assertion. The grayzone scheme's description as "NWP-specific tuning" (line 152) should likewise be attributed.

Response: We added the requested list of operational suites to the Introduction. For ICON: ICON-D2 at ∼2.2 km (DWD), ICON-2I at ∼2.2 km (ItaliaMeteo/ARPAE), ICON-CH2 at ∼2.1 km (MeteoSwiss). For WRF: the operational deployments at MeteoGalicia (∼4–1 km), CETEMPS (∼3–1 km) and CMCC (∼2 km). All the listed systems run convection-permitting configurations with shallow-only or scale-adapted convection treatments, so a grid spacing of ≈2 km is representative of current operational practice for both model families. The description of the ICON grayzone option as an NWP-specific tuning of the deep-convection scheme is now attributed to the ICON model documentation.

Changes in the manuscript: The Introduction now lists the operational suites with citations `(lines 47-51)`

> " Specifically, a horizontal grid spacing of ≈ 2 km is representative of the current convection-permitting operational standard adopted by major national meteorological services and regional research institutes like ICON-D2 at ∼2.2 km by (DWD, 2026), ICON-2I at ∼2.2 km by (ItaliaMeteo, 2026), ICON-CH2 at ∼2.2 km by (MeteoSwiss, 2026), WRF at ∼4 km – 1 km by (MeteoGalicia, 2026), WRF at ∼3 km – 1 km by (CETEMPS, 2026), WRF at ∼2 km by (CMCC, 2026). "

and the grayzone description in Sect. 3.1 carries a citation to the ICON documentation (`lines 188-191`).

> " The grayzone parametrization used in ICON–GZ is a NWP-specific tuning of the deep convection scheme in order to reduce the activity of the convection scheme to just a bit more than pure shallow convection. It includes an increased entrainment rate for parcel ascent calculations and a modified CAPE closure in order to suppress widespread convective drizzle (Deutscher Wetterdienst, 2025).
> "

Deutscher Wetterdienst: Working with the ICON model: ICON Tutorial 2025, page 107, Tech. rep., Deutscher Wetterdienst (DWD), https://doi.org/10.5676/DWD_pub/nwv/icon_tutorial2025, 2025.

Specific comment 13

Line 28: Missing a reference for socio-economic impacts. The sentence on the importance of correctly forecasting these phenomena "for early warning, prevention, and adaptation" would be well supported by the recent Reviews of Geophysics synthesis on the socio-economic impacts of Mediterranean cyclones. Please consider to add the reference Khodayar et al., 2025.

Response: The reference has been added at the indicated sentence.

Changes in the manuscript: Added reference to (Khodayar et al., 2025) in `(line 30)`

Minor comments

The Abstract characterizes Daniel as featuring a "rapid transition from a baroclinic disturbance to a compact tropical-like vortex." This appears to contradict the manuscript's own Case Study (Sect. 2), which describes a five-phase evolution over 6 days (4–10 September) with a gradual transition from 7–9 September, and especially the Discussion (lines 433–437), which explicitly states that "unlike systems that rapidly attain a mature structure, Daniel evolved through a prolonged sequence of physically distinct stages." Please reconcile these: either rephrase the Abstract (e.g., "multi-stage" transition, or restrict "rapid" to the final barotropic alignment on 9 September if that is what is meant) or justify the "rapid" characterization against the stated timeline.

Response: The Referee is right; the intended meaning was the final barotropic alignment, and the Abstract now says so explicitly.

Changes in the manuscript: The Abstract has been updated `(lines 1-4)`

> " ~~Medicane Daniel (September 2023) featured a rapid transition from a baroclinic disturbance to a compact tropical-like vortex, challenging short-range prediction.~~
> Medicane Daniel (September 2023) featured a multi-stage transition from a baroclinic disturbance to a compact tropical-like vortex, culminating in a rapid barotropic alignment on 9 September, challenging short-range prediction. "

Line 84: "70–80 km/[t]" — units appear to be km/h; correct the symbol.

Response: Both artefacts originated from a bulk search-and-replace in the LaTeX source. We corrected the two instances flagged and scanned the full source for residual "[t]" markers, finding none.

Changes in the manuscript: Substituted [t] with h

Line 90: "western Egyptian desert" → "Western Desert of Egypt".
Line 61: "future researche" → "future research."
Line 263: "Aegean Sea" → "Ionian Sea".

Line 359: "keeping a cold core and chaotic structure" — "chaotic" carries a specific dynamical-systems meaning that is not what is intended here. Replace with a more precise term such as "disorganized" or "incoherent."

Response: All four corrections have been applied as suggested; for line 359 we adopted "disorganized".

Changes in the manuscript: The four passages are corrected in the revised manuscript.

Line 378: "a horizontal extension of a few tents km" → "a few tens of km" (also line 390 region: check similar instances).

Response: Corrected, and the check requested by the Referee revealed one further instance of the same error class in the ICON–SH TASM paragraph ("in the first tenths of kilometers"), which has also been corrected to "in the first tens of kilometres". The full source has been scanned for similar misuses.

Changes in the manuscript: Line 378 now reads "a horizontal extension of a few tens of km"; the ICON–SH sentence now reads "the wind contours are packed in the first tens of kilometres from the center".

Line 459: SST is mentioned, but never defined. Define it in Line 77 when it is first mentioned.

Response: SST is now defined at its first mention.

Changes in the manuscript: Line 77 now reads "The Sea Surface Temperature (SST) in the south of Greece [. . .]".

Figure 5 vs. Section 3.4 radius mismatch (see Specific Comment 7).

Response: Addressed under Specific comment 7 above: the code uses 150 km; the caption has been corrected.

Changes in the manuscript: See Specific comment 7.

Figures 8, 9, and 10 are first introduced and discussed in Section 4.3 (Results) but are typeset within Section 5 (Discussion and Conclusion) [. . .] Please relocate them closer to their first citation to aid readability.

Response: The three figures have been relocated within Sect. 4.3, and their placement has been verified in the compiled revised manuscript. Float positions will be re-checked after the final edit pass, since the Copernicus class may shift them at typesetting.

Changes in the manuscript: Figures 8–10 are now typeset within Sect. 4.3, before the Conclusions.

References cited in this response

- Flaounas, E., Dafis, S., Davolio, S., Faranda, D., Ferrarin, C., Hartmuth, K., Hochman, A., Koutroulis, A., Khodayar, S., Miglietta, M. M., et al.: Dynamics, predictability, impacts and climate change considerations of the catastrophic Mediterranean Storm Daniel (2023), Weather and Climate Dynamics, 6, 1515–1538, https://doi.org/10.5194/wcd-6-1515-2025, 2025

- Hewson, T., Ashoor, A., Boussetta, S., Emanuel, K., Lagouvardos, K., Lavers, D., Magnusson, L., Pillosu, F., and Zsoter, E.: Medicane Daniel: An extraordinary cyclone with devastating impacts, ECMWF newsletters, 179, 33–47, https://doi.org/10.21957/th3wxk861d, 2024.
- Khodayar, S., Kushta, J., Catto, J. L., Dafis, S., Davolio, S., Ferrarin, C., Flaounas, E., Groenemeijer, P., Hatzaki, M., Hochman, A., Kotroni, V., Landa, J., Láng-Ritter, I., Lazoglou, G., Liberato, M. L. R., Miglietta, M. M., Papagiannaki, K., Patlakas, P., Stojanov, R., and Zittis, G.: Mediterranean Cyclones in a Changing Climate: A Review on Their Socio-Economic Impacts, Reviews of Geophysics, 63, https://doi.org/10.1029/2024RG000853, 2025.
- Lagouvardos, K., Kotroni, V., Bezes, A., Koletsis, I., Kopania, T., Lykoudis, S., Mazarakis, N., Papagiannaki, K., and Vougioukas, S.: The automatic weather stations NOANN network of the National Observatory of Athens: operation and database, Geoscience Data Journal, 4, 4–16, https://doi.org/10.1002/gdj3.44, 2017
- Miglietta, M. M., Mastrangelo, D., and Conte, D.: Influence of physics parameterization schemes on the simulation of a tropical-like cyclone in the Mediterranean Sea, Atmospheric Research, 153, 360–375, https://doi.org/10.1016/j.atmosres.2014.09.008, 2015.
- Miglietta, M. M., Buscemi, F., Dafis, S., Papa, A., Tiesi, A., Conte, D., Davolio, S., Flaounas, E., Levizzani, V., and Rotunno, R.: A high-impact meso-beta vortex in the Adriatic Sea, Quarterly Journal of the Royal Meteorological Society, 149, 637–656, https://doi.org/10.1002/qj.4432, 2023.
- Miglietta, M. M., Flaounas, E., González-Alemán, J. J., Panegrossi, G., Gaertner, M. A., Pantillon, F., Pasquero, C., Schultz, D. M., D'adderio, L. P., Dafis, S., et al.: Defining medicanes: Bridging the knowledge gap between tropical and extratropical cyclones in the Mediterranean, Bulletin of the American Meteorological Society, 106, E1955–E1971, https://doi.org/10.1175/BAMS-D-24-0289.1, 2025
- Picornell, M. A., Campins, J., and Jansà, A.: Detection and thermal description of medicanes from numerical simulation, Natural Hazards and Earth System Sciences, 14, 1059–1070, https://doi.org/10.5194/nhess-14-1059-2014, 2014
- Pytharoulis, I., Kartsios, S., Tegoulias, I., Feidas, H., Miglietta, M. M., Matsangouras, I., and Karacostas, T.: Sensitivity of a mediterranean710 tropical-like cyclone to physical parameterizations, Atmosphere, 9, 436, https://doi.org/10.3390/atmos9110436, 2018
- Ricchi, A., Miglietta, M., Barbariol, F., Benetazzo, A., Bergamasco, A., Bonaldo, D., Cassardo, C., Falcieri, F., Modugno, G., Russo, A., Sclavo, M., and Carniel, S.: Sensitivity of a Mediterranean Tropical-Like Cyclone to Different Model Configurations and Coupling Strategies, Atmosphere, 8, 92, https://doi.org/10.3390/atmos8050092, 2017.
- Ricchi, A., Miglietta, M. M., Bonaldo, D., Cioni, G., Rizza, U., and Carniel, S.: Multi-Physics Ensemble versus Atmosphere–Ocean Coupled Model Simulations for a Tropical-Like Cyclone in the Mediterranean Sea, Atmosphere, 10, 202, https://doi.org/10.3390/atmos10040202, 2019
- Roberts, N. M. and Lean, H. W.: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events, Monthly Weather Review, 136, 78–97, https://doi.org/10.1175/2007MWR2123.1, 2008
- Saraceni, M., Silvestri, L., Bechtold, P., and Bongioannini Cerlini, P.: Mediterranean tropical-like cyclone forecasts and analysis using the ECMWF ensemble forecasting system with physical parameterization perturbations, Atmospheric Chemistry and Physics, 23, 13 883– 13 909, https://doi.org/10.5194/acp-23-13883-2023, 2023.
- Wilks, D. S.: Statistical methods in the atmospheric sciences, vol. 100, Academic press, 2011

- Zwiers, F. W. and Von Storch, H.: On the role of statistics in climate research, International Journal of Climatology: A Journal of the Royal Meteorological Society, 24, 665–680, https://doi.org/10.1002/joc.1027, 2004

Figures

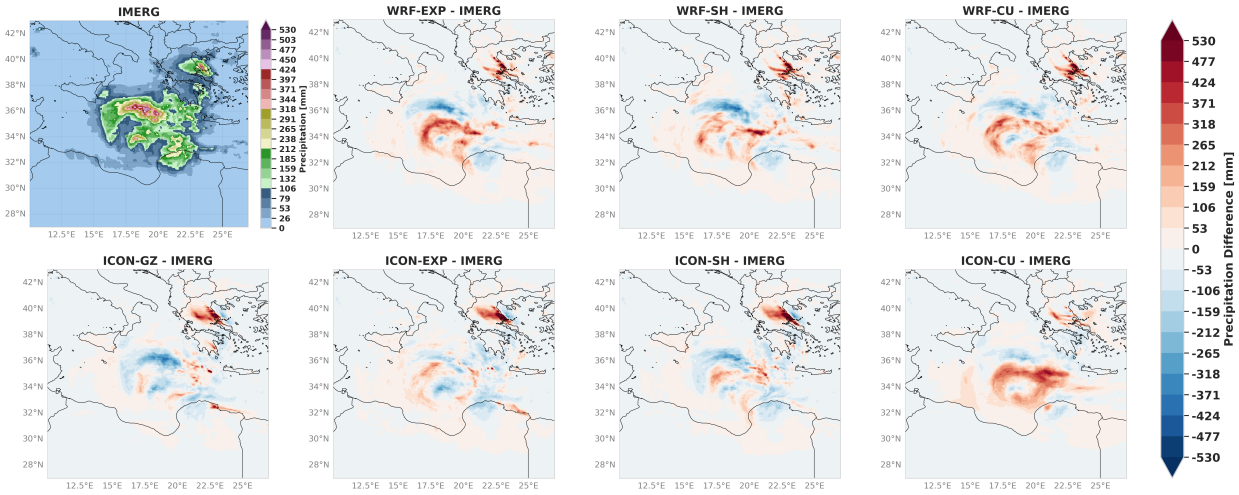

Figure 1: Total Precipitations Regridded difference Simulation–IMERG

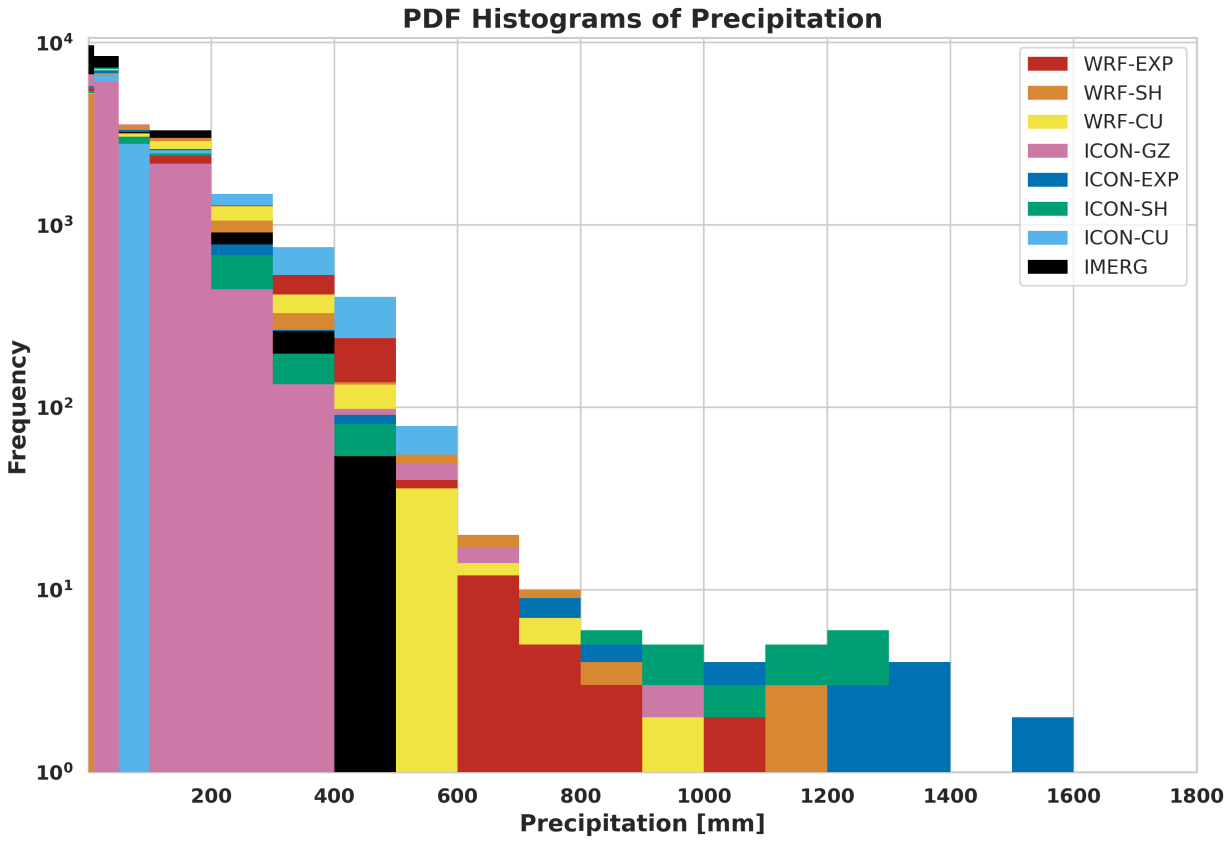

Figure 2: Probability Density Function of regridded precipitations

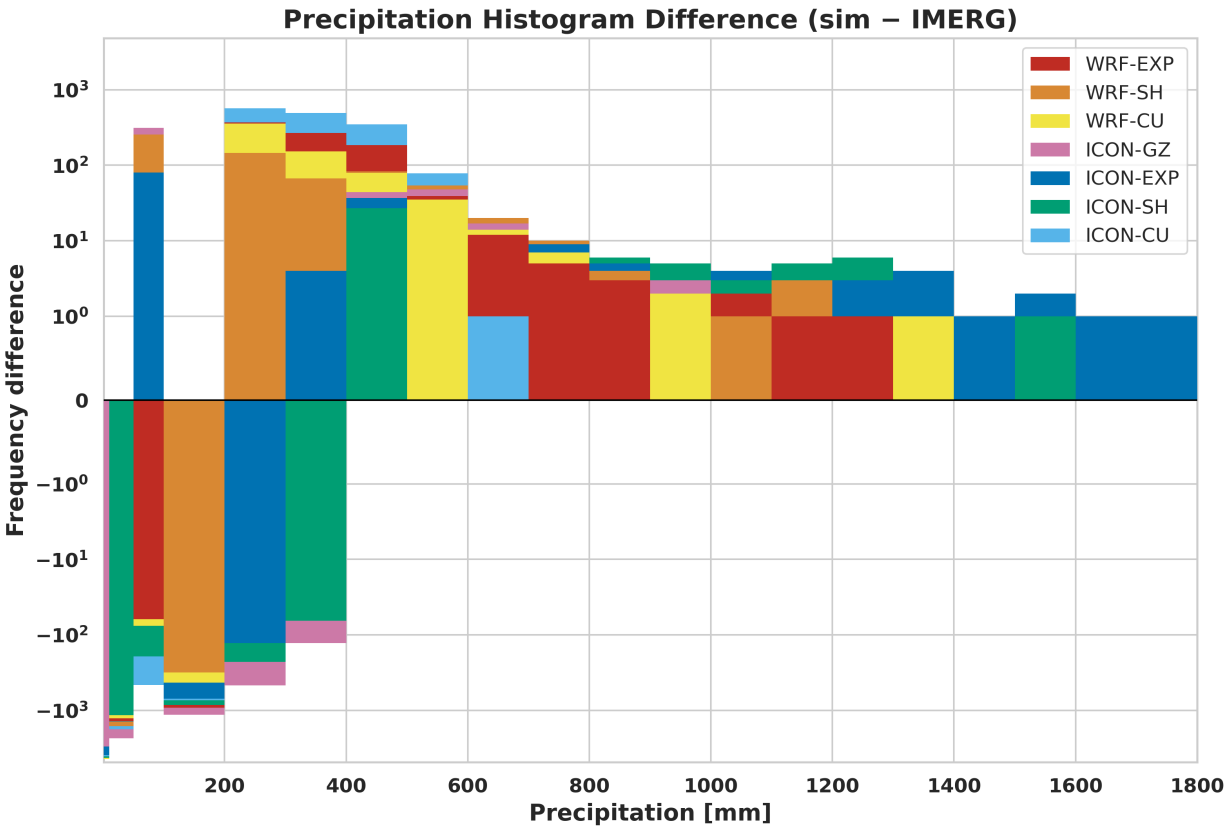

Figure 3: Probability Density Function of regridded precipitations Simulation–IMERG

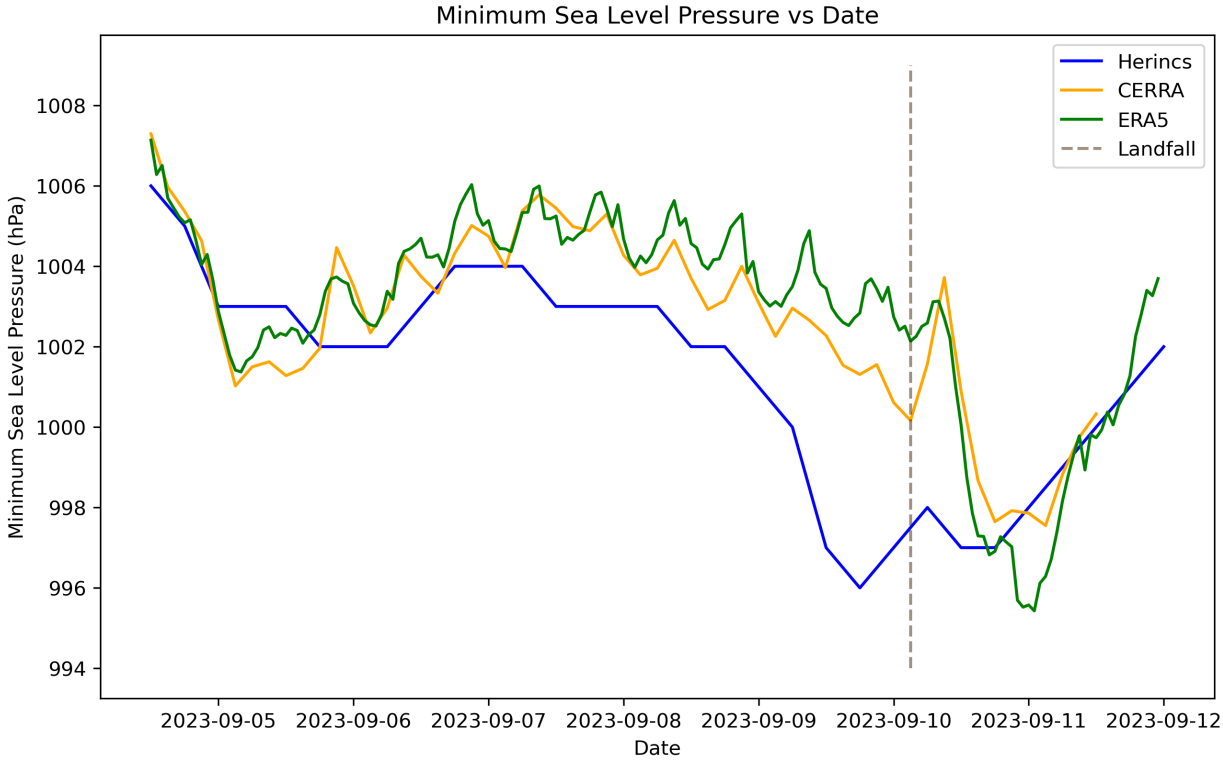

Figure 4: Timeseries CSLP for Hérincs, ERA5 and CERRA

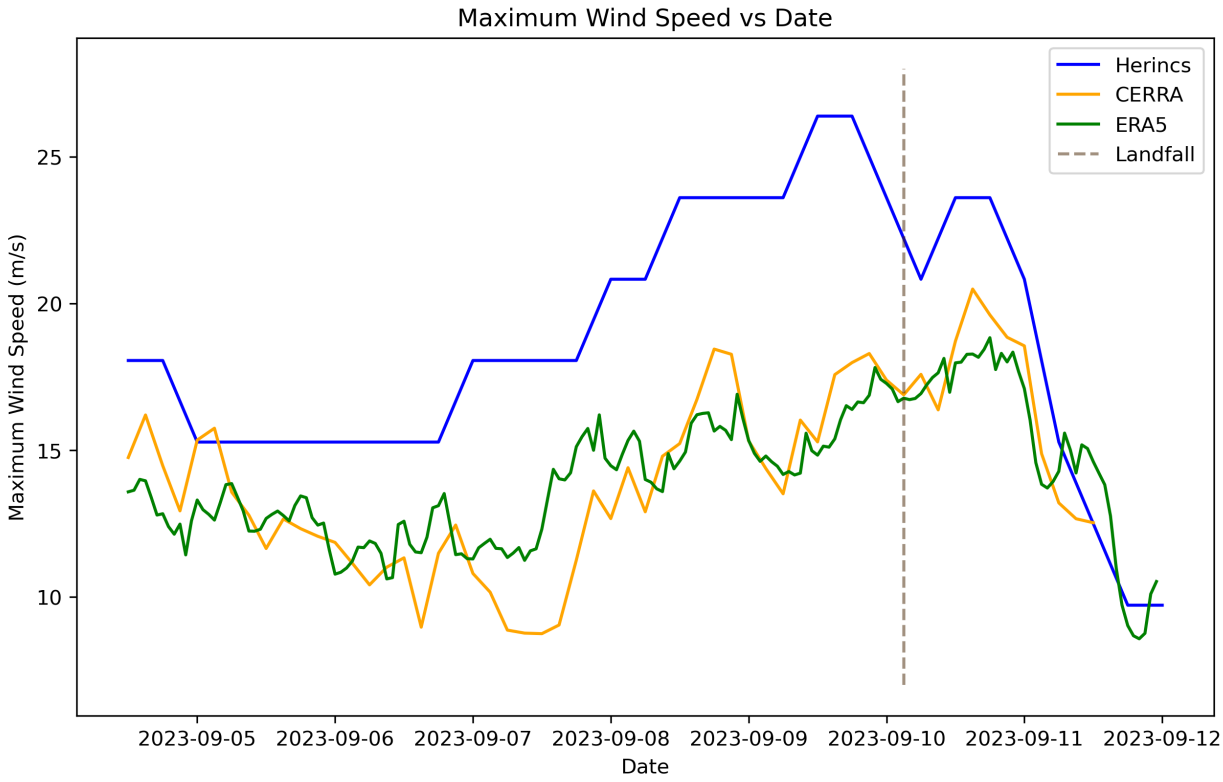

Figure 5: Timeseries Max Wind for Hérincs, ERA5 and CERRA

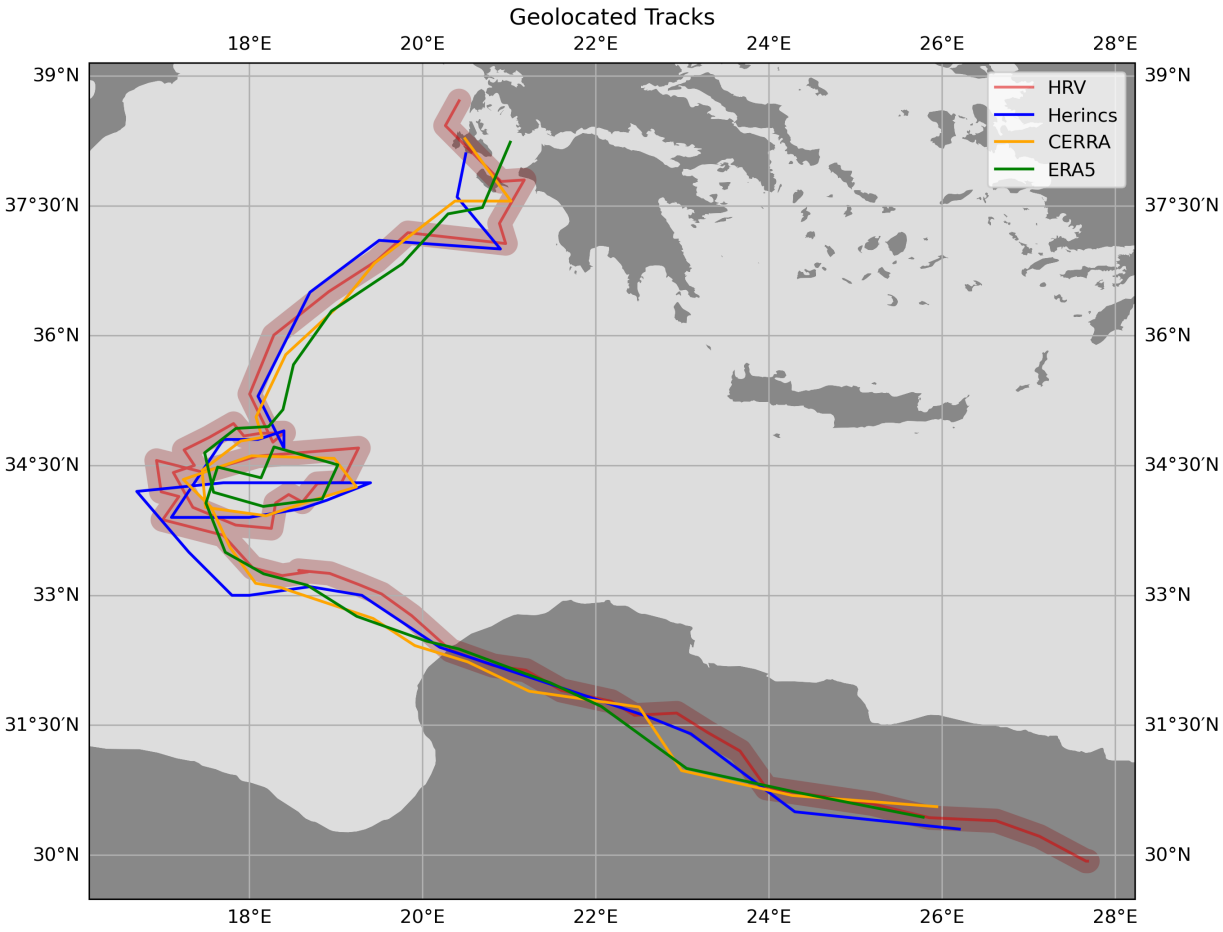

Figure 6: Tracks comparison Observed HRV, Hérincs, ERA5 and CERRA

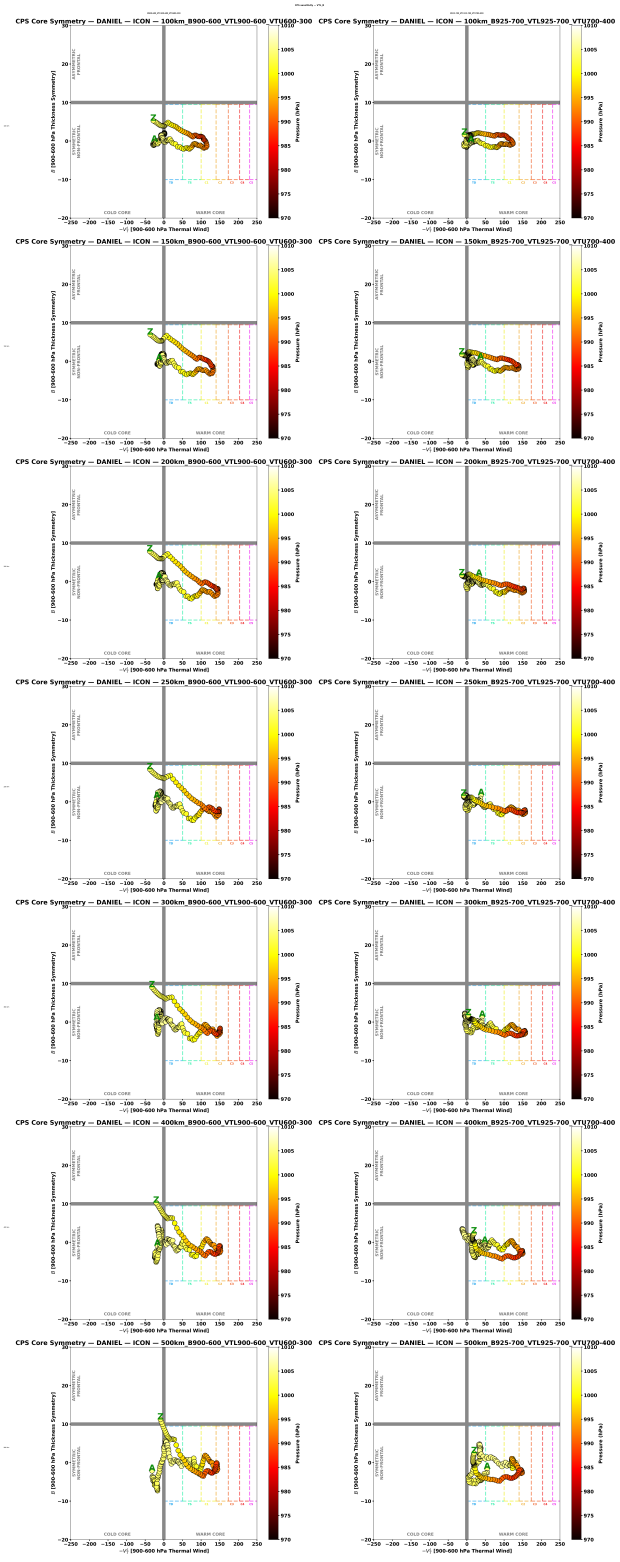

Figure 7: Hart Cyclone Phase Space sensitivity test B vs VTL

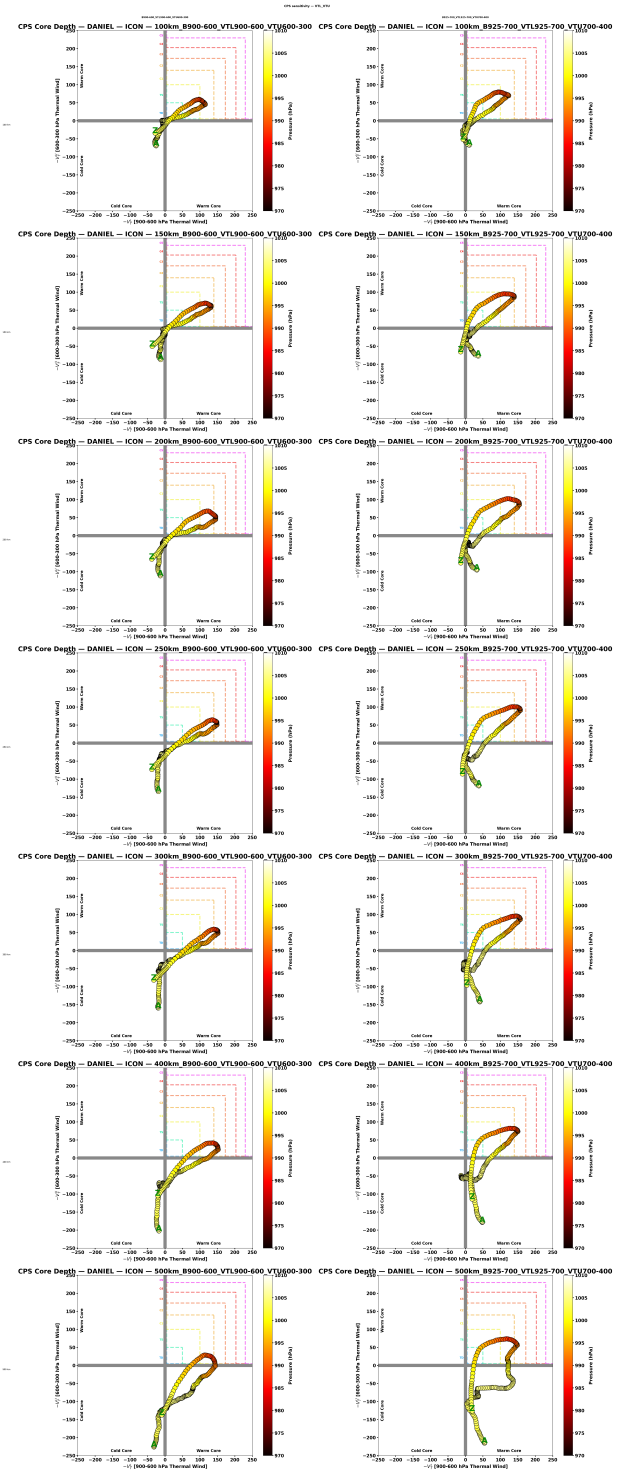

Figure 8: Hart Cyclone Phase Space sensitivity test VTU vs VTL

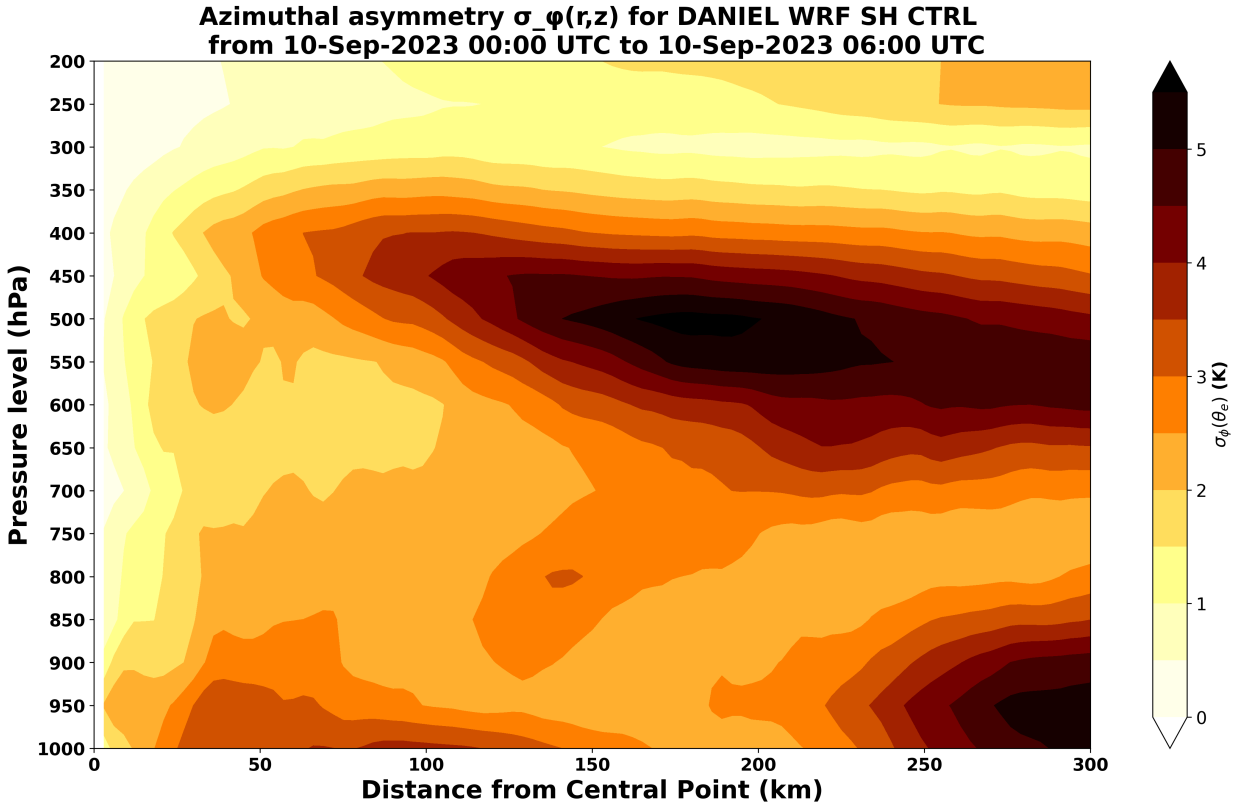

Figure 9: TASM axisymmetry evaluation on WRF–SH simulation

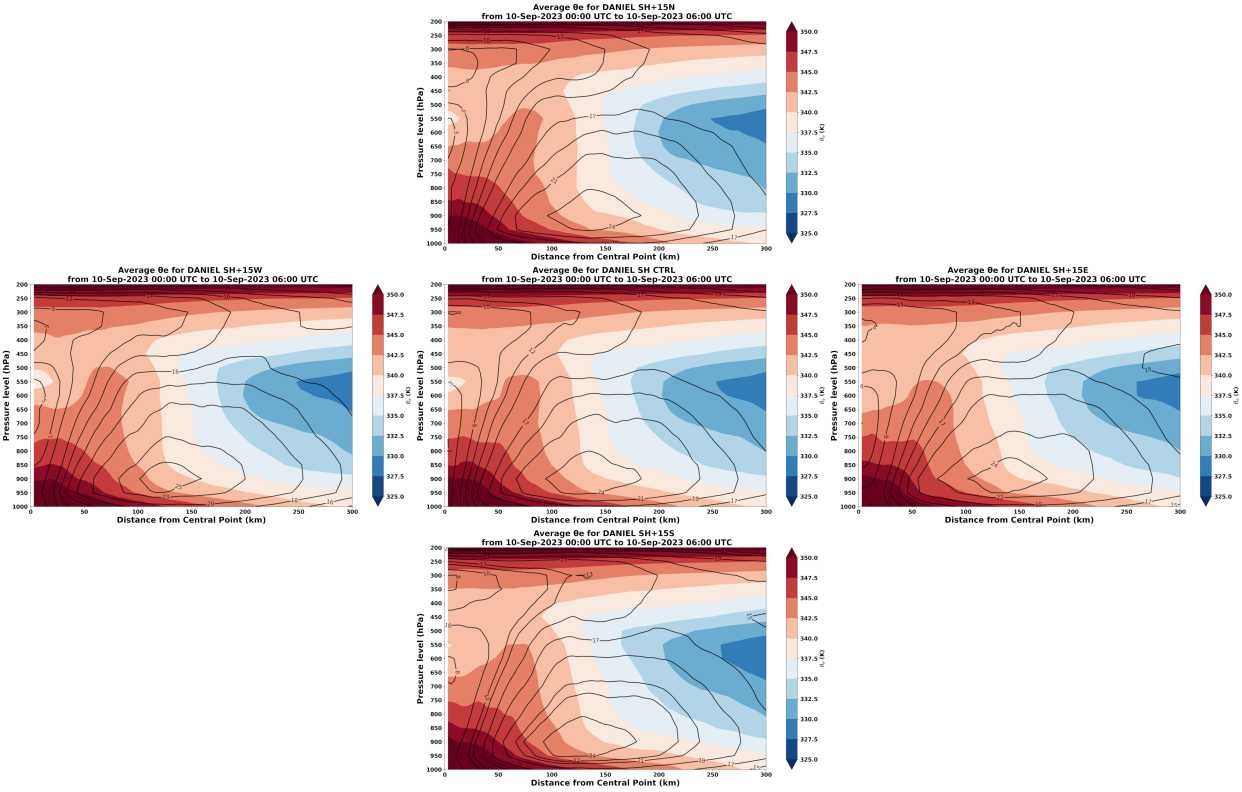

Figure 10: TASM displacement effect on WRF–SH simulation

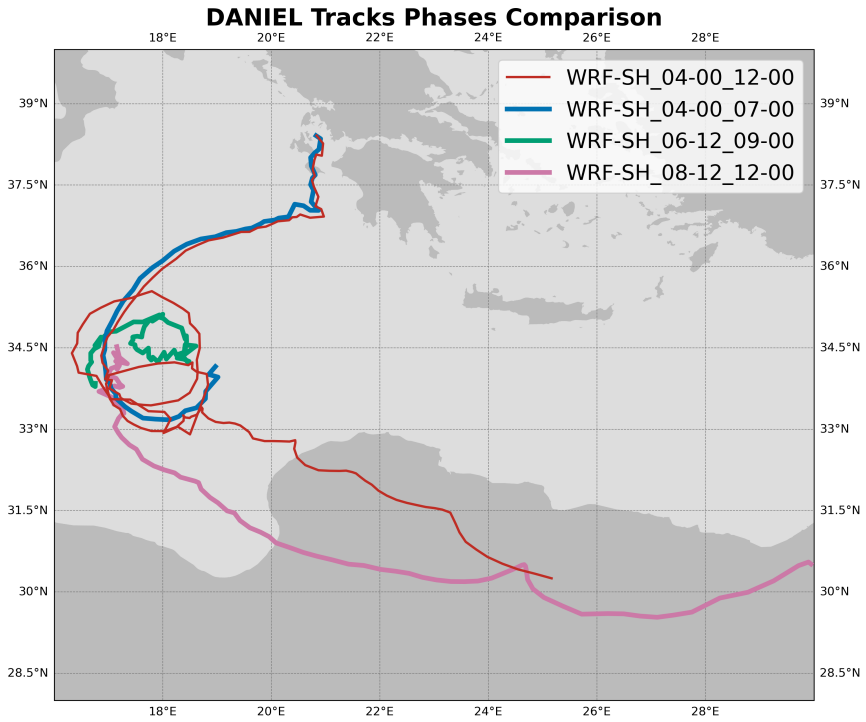

Figure 11: Tracks comparison between different initialization phases

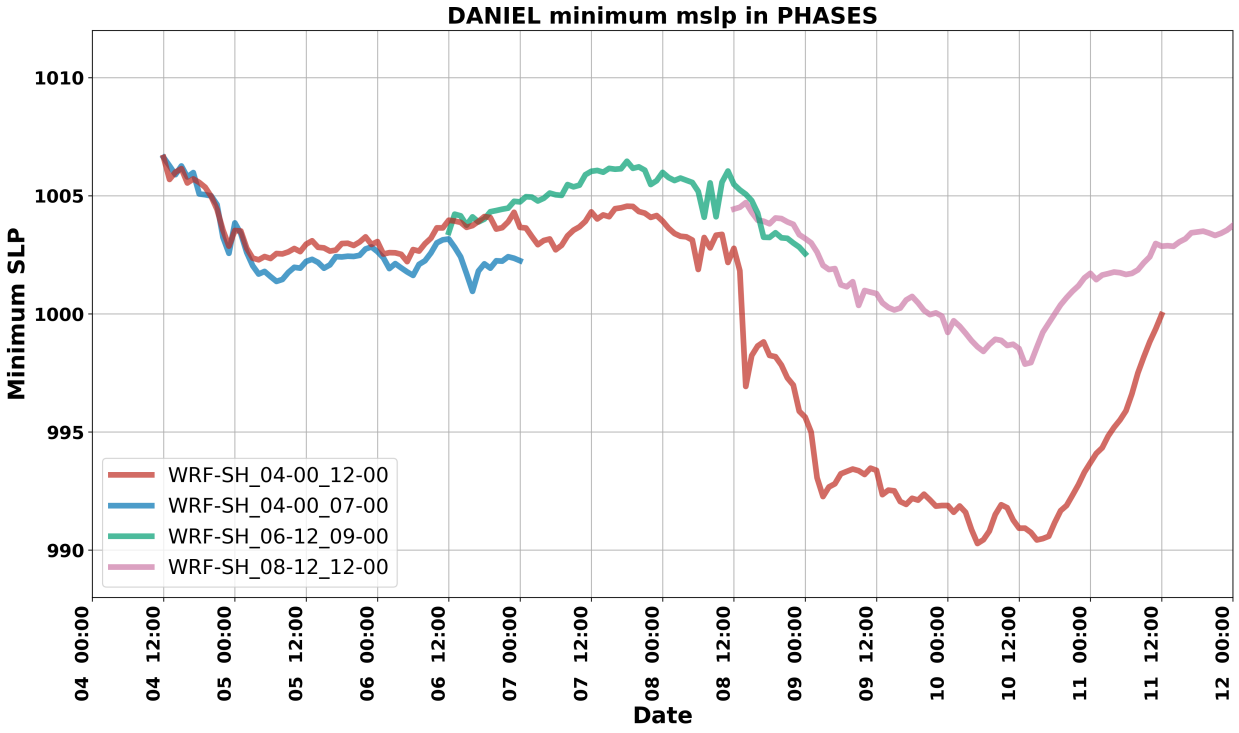

Figure 12: CSLP comparison between different initialization phases

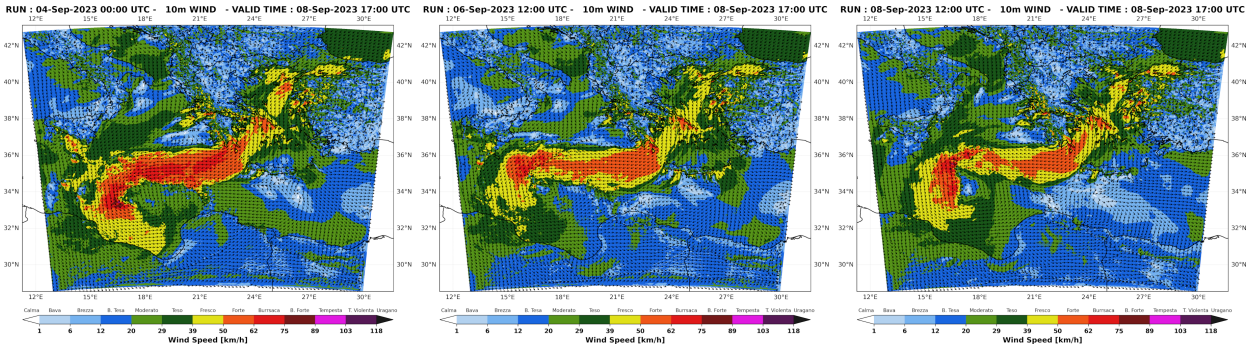

Figure 13: 10 m Wind Speed field comparison between different initialization phases at 08-September-2023 17:00 UTC

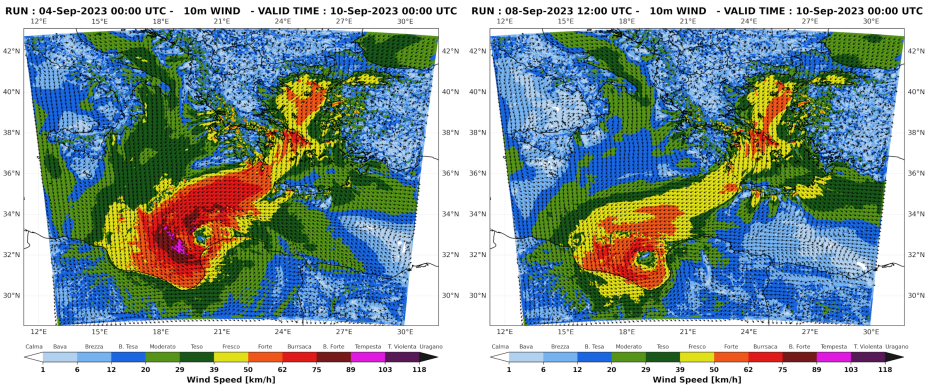

Figure 14: 10 m Wind Speed field comparison between different initialization phases at 10-September-2023 00:00 UTC