

Residual Corrective Diffusion Modeling for Km-scale Atmospheric Downscaling

Corresponding Author: Dr Morteza Mardani

This file contains all editorial decision letters in order by version, followed by all author rebuttals in order by version.

Version 0:

Decision Letter:

**** Please ensure you delete the link to your author home page in this e-mail if you wish to forward it to your coauthors ****

Dear Dr Mardani,

Your manuscript titled "Residual Diffusion Modeling for Km-scale Atmospheric Downscaling" has now been seen by 2 reviewers, and we include their comments at the end of this message. They find your work of interest, but some important points are raised. We are interested in the possibility of publishing your study in Communications Earth & Environment, but would like to consider your responses to these concerns and assess a revised manuscript before we make a final decision on publication.

We therefore invite you to revise and resubmit your manuscript, along with a point-by-point response that takes into account the points raised. Please highlight all changes in the manuscript text file.

In particular, please ensure that your revised manuscript meets the following editorial thresholds:

* Demonstrate that your downscaling approach is robust (including thorough validation) and more efficient than existing approaches

* Fully explain your methodology to the extent that it is reproducible and ensure that your code is openly available

We are committed to providing a fair and constructive peer-review process. Please don't hesitate to contact us if you wish to discuss the revision in more detail.

Please use the following link to submit your revised manuscript, point-by-point response to the referees' comments (which should be in a separate document to any cover letter), a tracked-changes version of the manuscript (as a PDF file) and the completed checklist:

Link Redacted

**** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first ****

We hope to receive your revised paper within six weeks; please let us know if you aren't able to submit it within this time so that we can discuss how best to proceed. If we don't hear from you, and the revision process takes significantly longer, we may close your file. In this event, we will still be happy to reconsider your paper at a later date, as long as nothing similar has been accepted for publication at Communications Earth & Environment or published elsewhere in the meantime.

Please do not hesitate to contact us if you have any questions or would like to discuss these revisions further. We look forward to seeing the revised manuscript and thank you for the opportunity to review your work.

Best regards,

Joe Aslin

Deputy Editor,

EDITORIAL POLICIES AND FORMATTING

We ask that you ensure your manuscript complies with our editorial policies. Please ensure that the following formatting requirements are met, and any checklist relevant to your research is completed and uploaded as a Related Manuscript file type with the revised article.

Editorial Policy: [Policy requirements](https://www.nature.com/documents/nr-editorial-policy-checklist.pdf)
(Download the link to your computer as a PDF.)

Furthermore, please align your manuscript with our format requirements, which are summarized on the following checklist: [Communications Earth & Environment formatting checklist](https://www.nature.com/documents/commsj-phys-style-formatting-checklist-article.pdf)

and also in our style and formatting guide [Communications Earth & Environment formatting guide](https://www.nature.com/documents/commsj-phys-style-formatting-guide-accept.pdf).

*** DATA: Communications Earth & Environment endorses the principles of the Enabling FAIR data project (<http://www.copdess.org/enabling-fair-data-project/>). We ask authors to make the data that support their conclusions available in permanent, publically accessible data repositories. (Please contact the editor if you are unable to make your data available).

All Communications Earth & Environment manuscripts must include a section titled "Data Availability" at the end of the Methods section or main text (if no Methods). More information on this policy, is available at <http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf>.

In particular, the Data availability statement should include:

- Unique identifiers (such as DOIs and hyperlinks for datasets in public repositories)
- Accession codes where appropriate
- If applicable, a statement regarding data available with restrictions
- If a dataset has a Digital Object Identifier (DOI) as its unique identifier, we strongly encourage including this in the Reference list and citing the dataset in the Data Availability Statement.

DATA SOURCES: All new data associated with the paper should be placed in a persistent repository where they can be freely and enduringly accessed. We recommend submitting the data to discipline-specific, community-recognized repositories, where possible and a list of recommended repositories is provided at <http://www.nature.com/sdata/policies/repositories>.

If a community resource is unavailable, data can be submitted to generalist repositories such as [figshare](https://figshare.com/) or [Dryad Digital Repository](http://datadryad.org/). Please provide a unique identifier for the data (for example a DOI or a permanent URL) in the data availability statement, if possible. If the repository does not provide identifiers, we encourage authors to supply the search terms that will return the data. For data that have been obtained from publically available sources, please provide a URL and the specific data product name in the data availability statement. Data with a DOI should be further cited in the methods reference section.

Please refer to our data policies at <http://www.nature.com/authors/policies/availability.html>.

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

The paper puts forward a method for downscaling weather variables over Taiwan. The main novelty appears to be

- Downscaling of multiple variables at the same time, particularly for typhoon downscaling.
- The combined use of two machine learning models, one a deterministic UNet and the other a diffusion model.
- The application of downscaling to Taiwan.

Overall, I think there are novel, interesting results here that deserve publishing in a journal in some form. It addresses interesting questions, but in my opinion does not do so in a particularly thorough manner, and does not seem to me to satisfy this journal's publication criteria for the following reasons:

1. "The paper provides strong evidence for the main conclusions": I don't think the evidence provided is particularly strong, as only a small number of samples are used in evaluation, and the analysis of the probabilistic qualities of the model are

limited to the CRPS, which does not give much insight into how well calibrated the model is. Additionally, a lot of emphasis seems to be put onto case studies, from which it is hard to draw robust conclusions.

2. "The underlying data are generated and processed at a state-of-the-art level, including tracking of uncertainties": Related to the above, there isn't any accounting for sampling variability from using relatively few samples. The methodology in places is not clear enough to answer this as well, e.g. it isn't clear from the methodology if/how data is used to perform model validation prior to testing, or how the testing samples are distributed in the year.

3. "A paper should [...] represent an advance in understanding that is expected to influence thinking in the field": In my opinion the results that are most likely to influence thinking in the field are the combination of deterministic model and generative model to perform downscaling, and downscaling multiple variables at once. But I did not find the analysis to be presented in a way that gives robust take-home messages about either of these.

For these reasons, I am not recommending that it is accepted into this journal in its current form.

In addition to this, the paper comes across as not having been checked through thoroughly, and I found it structured in a very unintuitive way that makes it difficult to read the results. The exception is the background sections on Diffusion models which seem well written, although perhaps unnecessarily detailed relative to the other methodology sections. More detailed comments about these are given below, there are a few spelling/grammar/typos, I've tried to flag a few of them but it could do with a proof read.

Page 1:

"since training costs are superlinear with respect to the resolution of training data." Can you provide a reference for this?

"ML downscaling enters as an advanced (non linear) form of statistical downscaling with potential to emulate the fidelity of dynamical downscaling." To me this reads as though ML downscaling is only useful for emulation, which is true for this study but not true in general (and in subsequent paragraphs you discuss works that downscale towards observations).

Page 2

"Convolutional Neural Networks, which reduce the input dimensions" not strictly true for all CNNs, e.g. those using residual blocks keep the same dimensions, and depending on the number of channels the dimensions could increase, so I'm not sure of the validity of this statement?

I think It would be helpful to briefly mention other non-ML downscaling techniques, and describe why they are not suitable competitors to ML methods.

"and difficulties in capturing long tails of distributions". Is this a problem that diffusion models overcome though? If so, please give refs. If not, this is misleading, as it sounds like you are suggesting it can.

"diffusion models offer sample diversity and training stability [25, 15] alongside demonstrable skill in probabilistically generating km-scales". But other works using GANs have also demonstrated sample diversity and skill in probabilistically generating fine scales, this makes it sound like they don't. It seems to me like the main benefit of diffusion models is training stability?

"Moreover, diffusion models have been used directly for probabilistic weather forecasting and nowcasting [30, 33, 36]." There's also Price et. al. 2023 (GenCast) to include here.

"The key contributions of this paper are: ..." a few comments on this list:

- Why is the simultaneous downscaling of multiple variables not part of the key contributions?
- "CorrDiff provides realistic measures of stochasticity and uncertainty," this is quite a strong statement given the only probabilistic measure employed is the CRPS. Typically you would expect at least a spread-error plot to see how under or over dispersed the ensemble is, which CRPS doesn't show you.
- "CorrDiff reproduces the physics of coherent weather phenomena remarkably well" If this conclusion is purely based on the case studies given, then this statement is too strong. I would say that, for the case studies considered it does this, and to say anything more than that requires more evidence.

"were produced using a radar-assimilating Weather Research and Forecasting (WRF) physical simulator". To help the reader, I think it best to explicitly say that this is a dynamical downscaling model.

Page 3:

"The motivation for CorrDiff is that naively applying conditional diffusion models to directly learn $p(x|y)$ was unsuccessful. The large uncertainty of variables such as radar reflectivity at these scales requires substantial noise levels during the forward process demands a large number of steps in backward process.

This impedes learning and leads to poor sample fidelity". This seems like a very interesting point, and seems quite central to

one of the key points of this paper, are the authors able to elaborate on why other studies may not have come across this problem? Intuitively you might expect that downscaling multiple variables simultaneously requires more training time, might this also be a factor?

“based on law of total variance” missing ‘the’

Page 4

Figure 2 (a) uses superscripts in the y label, but looks like something has gone wrong with those for Figs 2 (b) and (c) I found the Results section confusing to read, think it would be clearer if the baselines (RF and Reg) are defined in sec 2.

“The target data (WRF) span four years (2018-2021) and additional first four months of 2022, at one hour time resolution. We use the first 3 years for training and the rest for testing.” Two points here:

1. I didn’t find it clear how you have used the data. Was a portion of the training data used for validation ? If so, how much / what dates? And it would be good to state how the test data is used (i.e. 2021 used for skill scores, etc.). It seems like the 2022 data is only used for analysis of the cold front, but it’s not clear why, and I can’t see this written anywhere.

2. Are all the training and test years fairly homogeneous? Particularly with respect to events like typhoons, since this has bearing on your case study in Sec 3.3.2.

Page 5

“Compared with several baselines, CorrDiff shows the highest deterministic skill (MAE) followed by the UNet regression (Reg), the random forest (RF)” I think it would make it much clearer to introduce these baselines in Section 2. Also, why not just call the UNet model UNet? it would be easier to read the results.

“Skill scores evaluated from 204 date and time combinations taken randomly from the out-of-sample year (2021).” Two points:

- This seems like a small number of samples, given the total available of around 8000. How can we be confident with this few samples that the results are significant, and truly represent the behaviour of the model over the whole year, and for all times of day? I suggest using many more samples, or at least using a method like bootstrapping to estimate sampling variability.
- Are all metrics and models evaluated on the same 204 date/time combinations? Or do they all use different random samples?

“A separate RF is fit with scikit-learn for each of the 4 output channels with 100 trees and the default hyperparameters.” Isn’t this then an unfair representation of the suitability of Random Forests to this task?

In section 3.2 you switch from using 204 samples to 192 samples. It’s not really clear why, and again this seems like a very small number of samples to draw robust conclusions from. See point above about bootstrapping.

One limitation in this section is that it is pixel-wise comparison, and uses only a univariate measure (CRPS). To make these results more compelling, it would be interesting to see results of a neighbourhood score (e.g. pooled CRPS as used in Ravuri et al 2021, Fractions Skill Score, etc).

“In contrast, for 2-meter temperature (Fig. 2 b) the performance of UNet-regression is comparable to CorrDiff;” also seems true for kinetic energy; particularly since with only 192 samples, variability in these results may be high.

“CorrDiff is able to match the target distribution, including the heavy-tailed structure, which is significantly improved from the the prediction based on the UNet alone. These rare extreme values reflect most of the risk.” It isn’t clear what percentile these values are, and so is hard to tell how extreme they are. Please elaborate what percentiles these are, or else a similar number like return period. Given the small number of samples, it also seems unlikely that your testing samples shed much light on really extreme regimes.

“For windspeed (Fig. 2 d), the UNet and the baseline models underestimate the probability of winds faster than 20 m/s.” Perhaps I’m misreading the graph, but in Fig 2d it looks like The Reg, CorrDiff and target are basically indistinguishable, so this statement doesn’t seem to be true?

“For 2-meter temperatures (Fig. 2 e), the baseline models overestimate “ not clear what you mean by baseline; I would interpret this as everything except CorrDiff. But here this seems to exclude UNet, or at least it should, as the UNet curve seems to align almost exactly with the target

“the match of slopes at high values attests that the power-law relationship regulating the heavy-tail structure of rain extremes is emulated skillfully.” Hard to tell what happens at the high end; maybe a log-log plot would be more informative, to see how the Diffusion model performs at the end of the tails?

“the variability of the generated radar reflectivity field at four distinct times.” How were these times chosen? At random?

"When analyzing 200 samples" My understanding is you mean an ensemble size of 200, but this isn't that clear from the text.

Page 6

Figure 3.

- This Figure would be a lot clearer by adding a title for each column (like you have done in Fig 4)
- Units required on colorbar axes

"Fig. 4 shows an example of CorrDiff downscaling a cold front." How is this chosen? Is this typical of the kind of frontal predictions seen?

Page 7

"Cold front on Feb 2, 2022 at 20 UTC." Inconsistent formatting of times throughout document.

"Both the location of the front, and the magnitude of the horizontal wind and temperature gradients (sharpening of the front) associated with it, are captured well by CorrDiff although some mispositioning of the exact front location is inevitable" This is a strange sentence, it sounds like you are saying that CorrDiff predicts the front location well, whilst also it is inevitable that it doesn't predict it well...

3.3.2 Tropical Cyclone case study: Not clear why you are not showing results for the UNet (Reg) here, since it seems to also perform well at improving the winds?

Page 8:

"Further analysis 5 expands beyond this" This is meant to link to supplementary info but shows up as a link to Methods

"Although no target data is available, when comparing the maximum windspeed and radius of maximum windspeed from CorrDiff predictions to a reference from the Japan Meteorological Agency best track dataset [4] CorrDiff is found to correct at least 75% of the error in intensity for windspeed below 50m/s but only 50% of the error for higher windspeeds S4." Again I'm not sure why UNet isn't also included in this comparison, as it would be insightful to see if the corrections here are due to the deterministic or probabilistic component.

"An attractive use case for CorrDiff is to directly downscale forecasts from global models." I'm not really sure what the value of this analysis is; it seems to say more about how close GFS forecasts are to ERA5 than anything else. Is there any use-case where we would apply this model trained on ERA5 to GFS data, instead of just training it on GFS in the first place? These results also make CorrDiff appear worse at first glance, until you realise that it is on different data.

Page 9

"the model is shown to skillfully correct kinetic energy spectra, generate realistic probabilities and downscale coherent structures accurately" Two points:

- Whilst this is true, it also misses that the UNet alone does almost as well on the kinetic energy spectra (and its hard to tell how significant this difference is given the small number of samples used). It seems like the main thing that differentiates CorrDiff from the UNet is the ability to represent complex spatial fields like radar reflectivity.
- There isn't much to go on to infer that the probabilities are realistic, since there is no target ensemble data to compare with. Again, I suggest some kind of analysis such as spread-error to back this up.

Methods: I'm not sure why Methods are after Discussion. Is this a specification of the journal? It seems like this would be a lot clearer if a lot of this came before the results and was combined with some of the other background.

Page 11

"5.1 Background on diffusion models" This seems like an overly detailed background on diffusion models; are there not other existing articles on diffusion models that the reader could be directed to for further details? Or else can this part moved to the Supplementary information?

"we resort to UNet-regression. A UNet network is supervised with training data $\{(x_n, y_n)\}_{n=1}^N$ to learn the regression. We adopt a UNet architecture that is commonly used in denoising diffusion models." Lots of repetition

Page 12

"Each sample in the input dataset consists of 20 channels of information see ??." Broken link to section.

"This includes four pressure levels (500 hPa, 700 hPa, 850 hPa, and 925 hPa) with four corresponding variables: temperature, eastward and northward components horizontal wind vector, and Geopotential Height. Additionally, the dataset includes single level variables such as 2 meter Temperature, 10 meter wind vector, total column water vapor." Any explanation of why these inputs / pressure levels are used ?

Page 13

"Subsequently, vertical interpolated from sigma levels to isobaric levels then save as a custom CWA DMS format." Check grammar

"Additionally, any missing or corrupted data points represented by "inf" or "nan" values are eliminated from the dataset." Can you clarify? you mean any sample which has one or more nan/inf value is removed?

"To avoid over-fitting we divide the data into training and testing sets. Three years of data 2018-2020 are used for training (2,4154 samples total). For testing we use the full fourth year, 2021, as well as the first four months (January to April) of 2022." What about a validation set? How did you tune your hyperparameters?

"To ensure compatibility and consistency, we employ the same UNet architecture used in EDM (Elucidated Diffusion Models) for both the regression and diffusion networks(see S1). This architecture is based on the UNet model proposed in [53]." Repetition

"For time embedding in the diffusion process, we utilize Fourier-based position embedding. However, in the regression network, time embedding is disabled." Why no time embedding for the U Net? How do we know what benefits in skill are down to diffusion model vs just knowing the time of day?

"Additionally, we introduce 4 channels for sinusoidal positional embedding." Please provide reference if not describing exactly what this is.

"During the training phase, we use the Adam optimizer with a learning rate of ..." Need to explain how you chose these numbers; are these all based on ref [27]? Or chosen during hyperparameter tuning? It would be good to specify this at this point.

"The regression network solely receives the input conditioning channels. On the other hand, the diffusion training incorporates the 20 input conditioning channels from the coarse-resolution ERA5 data, which are concatenated with 4 noise channels to generate the output for each denoiser." This isn't very clearly written. I believe you mean that both UNet and Diffusion model take 20 input conditioning channels, and the diffusion model also takes in an additional 4 noise channels. But the wording makes this hard to read

"A necessary condition for calibration is that the spread-error relationship be 1-to-1 when averaged over sufficient samples" It seems strange to discuss spread-error and rank histograms and then not use them. Additionally, something like a spread-error plot would provide useful information about the calibration of the CorrDiff ensemble.

"CRPS is defined as" Please define the step function symbol used

"For a deterministic forecast, $F(y) = 1\{y \geq x_0\}$ " needs $\geq$ symbol

"CRPS is equivalent to the mean absolute deviation." I think its clearer to refer to it as mean absolute error, since this is the standard way to say this.

Page 18

"To the best of our knowledge ML downscaling of tropical cyclones, which requires accounting for different physics, across many channels and channel synthesis was not shown before." If this is the case, then I think this needs to be emphasised in the main body of text

Page 19

"In order to examine the realism of individual samples from CorrDiff and their quality compared with the target data figure shows an animation" No figure is referenced in the pdf.

"Comparison of simulated Typhoons with historical records": I'm not sure why the results in this section aren't more prominent. It's hard to draw robust conclusions from the case study, but this at least gives some insight into the overall performance of CorrDiff.

"One imperfection of CorrDiff is that it shrinks all storms, including those with the correct size or those that were too small in the ERA5 input data (panel a)." This seems like a key finding that should be reported in the main text.

Page 21

Y labels on Figs. S2 c and e are m/s, inconsi

Page 22

Figure S3 seems to be missing?

"Panel (a): abscissa: mean (dot) and one standard deviation (bars) of predicted radius of maximum winds for the ERA5 (red)

and CorrDiff (orange) storms, grouped by their observed values; ordinate: corresponding observed values." Written in quite a confusing way. Referring to abscissa and ordinate seems redundant here. E.g. could be written "Panel (a) mean (dot) and one standard deviation (bars) of predicted radius of maximum winds for the ERA5 (red) and CorrDiff (orange) storms, grouped by their observed values".

Reviewer #2 (Remarks to the Author):

Summary

The authors propose a novel approach to downscale global reanalysis or forecast data to a reliable high-resolution ensemble with a horizontal resolution of 2 km. They have trained the combination of a U-Net and Corrector Diffusion-network on a dataset of 4 years of ERA5 data and the corresponding high-resolution WRF forecasts (with radar DA). The area under study was a region surrounding Taiwan, and multiple variables were used as input (20) and output (4) channels. Precipitation was not explicitly added as an input channel, but only as an output channel.

The authors highlight the fidelity of their method in reproducing WRF's spectra and distributions for rainfall and 10m wind. They show its enhanced performance for CRPS for the test dataset, the improvement for wind speed for historical typhoon data as well as the performance for a "live" forecast of typhoon Loinu. The method substantially reduces the required computational power (1.300 times more energy efficient than the traditional forecasts).

Overall assessment and recommendations:

The proposed method is an important, timely and original contribution to ML-driven weather prediction, potentially enabling the generation of fast and reliable high-resolution forecasts even with a modest training dataset. The authors position their approach accurately within the field. Their approach to separate the learning process into a U-Net for the mean and a CorrDiff-network for the more localized residuals appears to be highly effective. Moreover, the methodology has a lot of potential not just in the context of downscaling, but also e.g. for the fast generation of large ensembles from high-resolution NWP forecasts, without the need for an ensemble training data set.

I recommend that the authors make the architecture or code available to the scientific community. This would greatly increase the impact of this work and, most importantly, warrant the research results' reproducibility.

The manuscript is well-written apart from some minor typographical errors, some of which are listed below. The structure of the manuscript, in which the results and discussion precede the details of the methodology, is unconventional but reads well.

Specific remarks:

There are some details in the methodology which were not entirely clear to me, so I would appreciate if the authors could clarify these points:

P2: You mention the path towards an end-to-end ML downscaling system. What do you mean by that, exactly (and how is this more end-to-end than previous ML-based downscaling techniques)? Do you mean it generalizes to both reanalysis and different forecast models?

P2: What was the update cycle / frequency of the CWA's forecasts, and which lead times were used as target for the training? Please clarify this in the manuscript or S.I. The forecast skill is not constant in leadtime and may also suffer from spin-up effects, how was this dealt with?

P3 and P9: Is there any potential for a pretrained model to be used over a different domain? Does it learn the local statistics, or the actual physics of downscaling? Latitude, longitude, orography, land-sea mask, etc. are not provided as input channels, so I suppose not. Would such a more generic model be feasible and what would be required? This would be highly valuable, considering the cost of retraining such a model.

P3 and further: you refer to "generated rainfall reflectivity field". Do you use a radar simulator to generate these fields? Why not just consider the (log) total precipitation (intensity or accumulation)? Reflectivity depends not just on precipitation intensity but also precipitation type, distance from the radar, etc. so may not be the most useful variable to look at. Moreover, it suggests that you are comparing to observations in Fig. 3, which you are not, if I understand correctly; you compared to WRF output.

P11: time here is not related to time in the forecast, please make sure this distinction is clear.

P12: only here it becomes clear that the precipitation-related output variable is in fact 1hr maximum radar reflectivity; please mention this earlier. Why look only at hourly maxima instead of accumulations, which are more useful to hydrologists and to calculate longer accumulations (e.g. 24h rainfall)?

P12: Is any other data than radar data assimilated in the CWA-WRF model – such as station observations, useful for correcting T2m? P1 mentions "radar and other sensors".

No real mention is made of the fact that this downscaling approach bypasses, or even replaces, data assimilation. Do I understand correctly that this approach yields an improvement wrt coarse global forecasts comparable to WRF downscaling with DA, even though this approach does not assimilate the high-resolution data? One would assume that the most recent observations are required to fill in the missing information needed to improve model skill at higher resolutions.

P13: It is not clear to me from the architecture of the CorrDiff method how you ensure continuity, or realistic temporal autocorrelations, between the subsequent generated high-resolution fields (for a single member forecast). Realistic time correlations are important e.g. for hydrological applications. Could you clarify this?

To enhance readability:

Please explain the meaning of your acronyms (e.g. SDE, EDM...) and variables (ω in $d\omega(t)$...) upon first use.

Typographical and other minor remarks:

P1: weather hazard -> weather hazards

Good estimates ... exist-s-

P2: Building from -> Building upon/on

P3: Missing "and" in "The large uncertainty...."

P6: same calendar date: and time / hour?

P7 caption of figure 4:

"true wind vectors" do you mean "target"?

"arrows are along front wind component" is a bit confusing, please rephrase

P8: "it[is] most accurate"
"CorrDiff seem[s] to struggle"

P10: donwscaling

P12: Missing reference: "see ??"
hydrometers -> hydrometeors

P13: " Subsequently, vertical interpolated from sigma levels to isobaric levels then save as a custom CWA DMS format." - this sentence has grammatical errors and contains jargon without explanation.

Supplementary Information:

The supplementary information still has a lot of typographical and grammar errors and I suggest the authors carefully go through this document before resubmitting it.

Some specific remarks:

S3: I cannot see the animation, I tried multiple PDF readers. I am interested to see how realistic the time evolution is. Would it be possible to share this video another way? I've used TIB (<https://www.tib.eu/en/publishing-archiving/digital-preservation/>) in the past.

As for the caption, it is misleading to refer to WRF as "truth" - rather call it target (as you do elsewhere).

P22: You mention the improved efficiency of the inference, but this is only a part of the total picture. How many MWh did the training take, and how often do you estimate you will need to retrain?

** Visit Nature Research's author and referees' website at www.nature.com/authors for information about policies, services and author benefits**

Communications Earth & Environment is committed to improving transparency in authorship. As part of our efforts in this

direction, we are now requesting that all authors identified as 'corresponding author' create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the Manuscript Tracking System by clicking on 'Modify my Springer Nature account' and following the instructions in the link below. Please also inform all co-authors that they can add their ORCIDs to their accounts and that they must do so prior to acceptance.

<https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research>

For more information please visit <http://www.springernature.com/orcid>

If you experience problems in linking your ORCID, please contact the [Platform Support Helpdesk](http://platformsupport.nature.com/).

Version 1:

Decision Letter:

**** Please ensure you delete the link to your author home page in this e-mail if you wish to forward it to your coauthors ****

Dear Dr Mardani,

Your manuscript titled "Residual Corrective Diffusion Modeling for Km-scale Atmospheric Downscaling" has now been seen by our reviewers, whose comments appear below. In light of their advice we are delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Earth & Environment.

We therefore invite you to revise your paper one last time to address the remaining concerns of our reviewers. At the same time we ask that you edit your manuscript to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

EDITORIAL REQUESTS:

Please review our specific editorial comments and requests regarding your manuscript in the attached "Editorial Requests Table".

*****Please take care to match our formatting and policy requirements. We will check revised manuscript and return manuscripts that do not comply. Such requests will lead to delays. *****

Please outline your response to each request in the right hand column. Please upload the completed table with your manuscript files as a Related Manuscript file.

If you have any questions or concerns about any of our requests, please do not hesitate to contact me.

SUBMISSION INFORMATION:

In order to accept your paper, we require the files listed at the end of the Editorial Requests Table; the list of required files is also available at <https://www.nature.com/documents/commsj-file-checklist.pdf>.

OPEN ACCESS:

Communications Earth & Environment is a fully open access journal. Articles are made freely accessible on publication. For further information about article processing charges, open access funding, and advice and support from Nature Research, please visit <https://www.nature.com/commsenv/open-access>

At acceptance, you will be provided with instructions for completing the open access licence agreement on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

Please use the following link to submit the above items:

Link Redacted

**** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first ****

We hope to hear from you within two weeks; please let us know if you need more time.

Best regards,

Joe Aslin

Deputy Editor,
Communications Earth & Environment
<https://www.nature.com/commsenv/>
Twitter: @CommsEarth

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

The authors have clearly put in a lot of effort to address concerns from the previous round of reviewing. Whilst the authors note they are unable to incorporate many samples due to time and space restrictions, they do include estimates of sampling uncertainty which demonstrate their results are robust. I also appreciate the addition of rank histograms and spread-skill plots which provide a clearer picture of how the ensemble behaves. Overall then I would recommend this revised version is accepted for publication.

One comment I would make is in relation to my 3rd major comment in the previous review (which I can see probably wasn't that clear in retrospect). What I felt was lacking previously was a clear message about the relative utility of the UNet versus the Diffusion model for this application. Whilst there has been much improvement here, I would say there is still room to improve the strength of the message. For example, the power spectra seem to show that UNet and CorrDiff perform similarly on 2m temperature and 10m wind speed, so I think it would be illuminating to see the UNet represented in the Frontal System Case Studies so that readers can get an idea of the differences of these approaches.

The following are minor comments / typos.

Minor comments / suggestions:

Section 3.2: I'm not sure why the CorrDiff MAE is taken for the ensemble mean, why not just a single member ensemble? To my mind that seems the more natural quantity to report on, as the limit of the CRPS as the ensemble goes to one member. Or perhaps worth reporting on both? The same applies to Table S3.

Fig. 3: Just checking that these definitions are consistent with Leutbecher and Palmer, Ensemble Forecasting, 2008.

- "the RSME of mean prediction for the..." just checking you are referring to ensemble mean here? (also note typo RSME -> RMSE).

- "The standard deviation is adjusted with a factor $\sqrt{1 + 1/n}$ " : My understanding is that you are using the unbiased estimator of the standard deviation, hence the difference compared to the correction factor of $\sqrt{(N+1) / (N-1)}$ in Leutbecher and Palmer. Just checking this is the case.

In their response to my point on line 45 (previous manuscript), the authors say they have updated the text but looks like this part of the updated text hasn't come through:

... We use 2018 through 2020 for training. For testing we used 205 random samples from 2021, and the rest used for validation. ..."

Typos:

Section 2: "Our input y ... is a low-resolution meteorological data"

"with a forward and a backward processes working in tandem."

Reviewer #2 (Remarks to the Author):

In this revised version of the manuscript, the authors present their combined U-Net and Corrector Diffusion-based downscaling approach with a correction of a bug in their code, which originally led to an overestimation of the model's performance. I appreciate the authors' scientific integrity and effort in admitting and correcting this bug. Even with a slightly reduced skill, I believe the model is a valuable advancement upon the state of the art. I also appreciate the intention of the authors to make the architecture openly available.

The authors have implemented nearly all the requested changes and added a valuable new supplementary section containing a statistical analysis which discusses the significance of their improvements.

The authors answered my question regarding the lead time of the WRF-CWA forecasts: the reference corresponds to the analysis, not to a forecast. However, this is still not clear from the manuscript itself. Moreover, this makes me wonder how the 1-hour maximum radar reflectivity was computed. This is (based on) a prognostic variable, meaning that only the analysis is insufficient, and you need to run the WRF model forward in time to compute it. Could the authors please clarify this?

Minor remarks:

Please correct: "radar reflectivity - a much more complex observable that relates to the *sixth moment of the cloud water droplet distribution*" -> *sixth moment of the raindrop size distribution*

There are some missing capitals in the References section (e.g. taiwan, era5, ...)

The manuscript and SI still contain some typographical / editing errors (e.g. "metrics reported in 1 -> metrics reported in Section 1") - please review carefully.

** Visit Nature Research's author and referees' website at www.nature.com/authors for information about policies, services and author benefits**

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

Response to the Reviewers

August 6, 2024

Dr. Morteza Mardani
NVIDIA Corporation
Santa Clara, CA

Email: mmardani@nvidia.com

Dear Editor:

We extend our sincere gratitude to you and both reviewers for their constructive feedback and positive opinion regarding our proposed method and findings. Your insights have been invaluable in enhancing the quality of our work.

During the revision phase, as we were preparing to finalize the software for public release, and in the process of attending to new diagnostics requested by the reviewers, we identified a critical bug in our code. This bug inadvertently mixed the training and testing data, resulting in overly optimistic results presented in the original submission. We have corrected this bug and have meticulously reproduced all the results in the paper.

We are confident that the updated results affirm that CorrDiff remains a viable approach, and hope you and the reviewers agree, despite the fact that some results have deteriorated somewhat compared to the initial submission. To mitigate over-fitting, we selected an earlier Regression checkpoint trained on only 2 million images. While this decision compromised the UNet (regression model) performance, it provided more room for improvement during the diffusion step. The key changes in results are as follows:

1. CorrDiff’s skill, as measured by MAE and CRPS scores, has notably decreased but still outperforms the baselines.
2. CorrDiff’s spectra maintain a similar match to the target, with some degradation in radar reflectivity.
3. The distributions continue to closely match the target, showing some improvements in 2-meter temperature while experiencing some degradation in radar reflectivity.
4. The weather front case study demonstrates comparable skill to previous results.
5. The typhoon case study has been updated:
a. The original typhoon analyzed proved too challenging and has been moved to the Supplementary Information (SI).
b. A new typhoon case from 2023 has been added to the main text.
c. In this new case, as well as other cases detailed in the SI, CorrDiff recovers approximately 50% of the intensity error between ERA5 and the target.

To us this deterioration is reassuring in hindsight as it more accurately reflects what one ought to expect from the performance on out-of-distribution data typically encountered in these settings.

In response to the reviewers' comments, we have included several additional experiments and analyses. Notably, we have added a new section on statistical analysis that discusses the calibration and statistical significance of the samples generated by CorrDiff. This section provides a deeper understanding of the robustness and reliability of our method. Overall, we believe that addressing the reviewers' comprehensively exhaustive comments and fixing the bug have significantly improved the manuscript's quality.

We look forward to a positive review in the next round. Detailed point-by-point responses to the reviewers' comments are attached. Thanks again to these reviewers for their time spent offering such constructive feedback. We are very grateful for it and hope the revised manuscript is proof that we have taken it all to heart.

Sincerely,

Morteza Mardani, on behalf of the co-authors

Contents

1	Major updates	3
2	Reviewer 1	3
2.1	Major comments	4
2.2	Minor comments	5
3	Reviewer 2	23
3.1	Overall assessment and recommendations	23
3.2	Specific remarks	24

1 Major updates

Based on the constructive feedback provided by the reviewers, we have significantly revised the manuscript. During the revision phase, we identified a subtle bug in the train-test data split. This bug had inadvertently mixed test data with training data, leading to overly optimistic results in the original submission. We have retrained the model (both UNet regression and Diffusion) and invested considerable effort to correct this bug and to produce accurate evaluation of the new model. The underlying performance has unavoidably degraded as is listed above. The source code has been publicly updated to ensure reproducibility, and it is available at this link ¹. The major updates are summarized as follows:

- **Corrected Train-Test Split:** In the revised version we retrained both regression and diffusion components of CorrDiff while ensuring the separation of train and test data. The revised regression model was trained to 2 million steps (instead of 40 million steps before) to avoid over-fitting it on the training data. As a result, the performance of this UNet regression on the 2021 data as a baseline has significantly changed. With proper data split, and when trained for 2M images, the revised regression model has an unavoidably larger, yet more realistic, mismatch with the out-of-sample target. This leaves room for diffusion model to improve the mean prediction. This change is evident in many of the revised figures. Additionally, to enhance the diffusion model, we further conditioned the denoising network on the output of the regression model to increase the global context for the residual diffusion model, aligning with the proposed residual learning methodology.
- **Updated Figures and Tables:** All figures, except Figure 1 and Supplementary Figure S1 (which are architectural sketches), have been updated. Table 1 has also been revised to reflect the corrected results. Furthermore, we added 6 extra figures and 1 table in the supplementary information as part of the revision process.

These changes have improved the manuscript’s accuracy and reliability, aligning it more closely with the reviewers’ insightful comments. To aid the reviewers in assessing the revised manuscript we include a document with the changes and additions highlighted.

2 Reviewer 1

The paper puts forward a method for downscaling weather variables over Taiwan. The main novelty appears to be

¹<https://github.com/NVIDIA/modulus/tree/main/examples/generative/corrdiff>

- Downscaling of multiple variables at the same time, particularly for typhoon downscaling.
- The combined use of two machine learning models, one a deterministic UNet and the other a diffusion model.
- The application of downscaling to Taiwan.

Overall, I think there are novel, interesting results here that deserve publishing in a journal in some form.

We thank the reviewer for recognizing our contributions. Additionally, we highlight the novel synthesis of high-resolution outputs, specifically radar reflectivity, from low-resolution inputs, alongside multivariate downscaling.

2.1 Major comments

It addresses interesting questions, but in my opinion does not do so in a particularly thorough manner, and does not seem to me to satisfy this journal's publication criteria for the following reasons:

1. "The paper provides strong evidence for the main conclusions": I don't think the evidence provided is particularly strong, as only a small number of samples are used in evaluation, and the analysis of the probabilistic qualities of the model are limited to the CRPS, which does not give much insight into how well calibrated the model is.

We thank the reviewer for taking the time to help us improve and expand the statistical assessment associated with our key findings. While the reviewer is correct that there is substantial variability in time in the metrics, by evaluating all the models on a common set of times we obtain robust statistical power. For example, the CRPS of CorrDiff for Radar reflectivity is lower than any of the other baselines for each of the 205 time samples. In the language of statistics, this is binomial test with a p-value of $(0.5)^{205} \approx 10^{-60}$. At the same time, during the revision we explored the needed ensemble size in our validation run, and found that it can be safely reduced to 32 members without any change to the results. Since the original 205 calendar times, each with 192 ensemble members, tends to occupy 150GB and require several hours to score, we used an updated protocol with 205 times and 32 ensemble members in the revised study.

We have added a section (SI 5.2) explaining our statistical analysis, a new figure with error bars (Fig S6), as well the the follow text in the caption of Table 1. In the manuscript we added:

"The differences between CorrDiff, Reg, and RF are all highly significant (see SI 5.2). CorrDiff has lower CRPS than Reg in 205/205 time samples."

Moreover, we added in the paper a section for model calibration (see sec 3.3), showing the spread skill ratios and rank histograms for all channels. The model is indeed under-dispersive and we acknowledge this deficiency. In the appendix we also added the pooled MAE and CRPS counterpart table for Table 1 in the main text. Although pooling improves the values of CRPS, the relative value compared with the baselines remain similar.

2. “The underlying data are generated and processed at a state-of-the-art level, including tracking of uncertainties”: Related to the above, there isn’t any accounting for sampling variability from using relatively few samples...

Please see response to the first Major Comment.

3. “A paper should ... represent an advance in understanding that is expected to influence thinking in the field”: In my opinion the results that are most likely to influence thinking in the field are the combination of deterministic model and generative model to perform downscaling, and downscaling multiple variables at once. But I did not find the analysis to be presented in a way that gives robust take-home messages about either of these. For these reasons, I am not recommending that it is accepted into this journal in its current form.

We thank the reviewer for endorsing the value of these main messages that we intended to convey to the *Nature Communications Earth and Environment* readership – that modern generative AI methods can be remarkably effective at multivariate downscaling and new channel synthesis, especially when used downstream of a regression backbone. We hope the evidence provided above has proved reassuring about the reviewer’s main concern regarding the statistical significance of the results, along with the changes made throughout the manuscript to improve clarity and better communicate statistical robustness.

In addition to this, the paper comes across as not having been checked through thoroughly, and I found it structured in a very unintuitive way that makes it difficult to read the results. The exception is the background sections on Diffusion models which seem well written, although perhaps unnecessarily detailed relative to the other methodology sections. More detailed comments about these are given below, there are a few spelling/grammar/typos, I’ve tried to flag a few of them but it could do with a proof read.

Many thanks to the reviewer for their time spent offering these constructive suggestions that have helped improve clarity of the manuscript throughout. We are grateful.

2.2 Minor comments

1. “since training costs are superlinear with respect to the resolution of training data.” Can you provide a reference for this?

Vision transformers² use attention mechanism to learn the relationship between sequence of tokens (here individual image pixels) that incurs quadratic complexity with the input resolution^{3 4}.

2. “ML downscaling enters as an advanced (non linear) form of statistical downscaling with potential to emulate the fidelity of dynamical downscaling.” To me this reads as though ML downscaling is only useful

²Dosovitskiy, A., et al. "An image is worth 16x16 words: Transformers for image recognition at scale." arXiv:2010.11929 (2020).

³Guibas J, Mardani M, Li Z, Tao A, Anandkumar A, Catanzaro B. "Efficient token mixing for transformers via adaptive fourier neural operators." International Conference on Learning Representations, 2021.

⁴Pathak J, Subramanian S, Harrington P, Raja S, Chattopadhyay A, Mardani M, Kurth T, Hall D, Li Z, Azizzadenesheli K, Hassanzadeh P. "Fourcastnet: A global data-driven high-resolution weather model using adaptive fourier neural operators." arXiv:2202.11214, 2022

for emulation, which is true for this study but not true in general (and in subsequent paragraphs you discuss works that downscale towards observations).

We very much agree and have changed “emulate” to “exceed”.

3. “Convolutional Neural Networks, which reduce the input dimensions” not strictly true for all CNNs, e.g. those using residual blocks keep the same dimensions, and depending on the number of channels the dimensions could increase, so I’m not sure of the validity of this statement?

We agree and have removed the phrase “which reduce the input dimensions” from this sentence.

4. I think It would be helpful to briefly mention other non-ML downscaling techniques, and describe why they are not suitable competitors to ML methods.

Thank you for this suggestion and we agree it adds helpful context and motivation. While we could not find a non ML statistical downscaling model that can match our exact problem (resolution gap, channels, location), we did edit the sentence to describe the general approach and common limitations of statistical downscaling:

" A common inexpensive alternative is to learn a statistical downscaling from these dynamical downscaling simulations and observations Wilby et al. (1998). This is typically done by learning the values of several parameters of a statistical mapping (e.g. quantile mapping, generalized linear regression) that best match the the high resolution “target” data Bano-Medina et al., (2020). In this context, ML downscaling enters as an advanced (non linear) form of statistical downscaling Rampal et al., (2024). with potential to exceed the fidelity of dynamical downscaling. "

5. “and difficulties in capturing long tails of distributions”. Is this a problem that diffusion models overcome though? If so, please give refs. If not, this is misleading, as it sounds like you are suggesting it can.

Diffusion models address long tail distribution challenges by enhancing mode coverage compared to GANs. Their iterative denoising process mitigates mode collapse, ensuring thorough exploration of the data distribution, including rare and outlier points. This results in a comprehensive representation of complex data, supported by optimized likelihood estimates and robust sampling mechanisms; see e.g.,^{5 6}. We included some of these references in the main text.

6. “diffusion models offer sample diversity and training stability [25, 15] alongside demonstrable skill in probabilistically generating km-scale”. But other works using GANs have also demonstrated sample diversity and skill in probabilistically generating fine scales, this makes it sound like they don’t. It seems to me like the main benefit of diffusion models is training stability?

While GANs have demonstrated sample diversity and fine-scale generation capabilities, diffusion models offer distinct advantages beyond training stability. Diffusion models excel in mode coverage, effectively capturing long-tail distributions and mitigating mode collapse, as evidenced by Dhariwal and Nichol (2021). Their iterative denoising process ensures thorough exploration of the data distribution, enhancing both the

⁵Dhariwal, P., & Nichol, A. (2021). Diffusion Models Beat GANs on Image Synthesis. Advances in Neural Information Processing Systems, 34, 8780-8794. Available at <https://arxiv.org/abs/2105.05233>.

⁶Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-Based Generative Modeling through Stochastic Differential Equations. Advances in Neural Information Processing Systems, 33, 12438-12448. Available at <https://arxiv.org/abs/2011.13456>

quality and variety of generated samples. Additionally, diffusion models optimize likelihood directly, providing a principled approach to probabilistic modeling and consistently producing high-fidelity samples with detailed resolution (Ho et al., 2020; Song et al., 2020). These features collectively position diffusion models as a robust alternative to GANs for generative tasks.

7. “Moreover, diffusion models have been used directly for probabilistic weather forecasting and nowcasting [30, 33, 36].” There’s also Price et. al. 2023 (GenCast) to include here.

Indeed, the inspiring work of Price et al. (2023) was not citable at the time of our original submission and we have expanded the indicated sentence with the following bolded suffix:

“... for probabilistic weather forecasting and nowcasting [30, 33, 36] – **including global ensemble predictions that can outperform conventional weather prediction on a range of important stochastic metrics at 0.25-degree resolution...**”

8. “The key contributions of this paper are: ..” a few comments on this list: Why is the simultaneous down-scaling of multiple variables not part of the key contributions?...

We have added the following bold text to our top-line contribution statement as follows:

“A physics-inspired, two-step approach (CorrDiff) to **simultaneously** learn mappings between low- and high-resolution weather data **across multiple variables** with high fidelity **alongside new channel synthesis.**”

9. “CorrDiff provides realistic measures of stochasticity and uncertainty,” this is quite a strong statement given the only probabilistic measure employed is the CRPS. Typically you would expect at least a spread-error plot to see how under or over dispersed the ensemble is, which CRPS doesn’t show you.

We agree with the reviewer and regret this error. This text has been removed, in lieu of the new results section on calibration in sec 3.3. Achieving optimal calibration is now emphasized as an important area for future work.

10. “CorrDiff reproduces the physics of coherent weather phenomena remarkably well” If this conclusion is purely based on the case studies given, then this statement is too strong. I would say that, for the case studies considered it does this, and to say anything more than that requires more evidence.

Thank you for this prompt to improve our text precision. The corresponding statement has been revised to:

“For two case studies considered, CorrDiff adds physically realistic improvements to the representation of under-resolved coherent weather phenomena – frontal systems and typhoons”.

11. “were produced using a radar-assimilating Weather Research and Forecasting (WRF) physical simulator”. To help the reader, I think it best to explicitly say that this is a dynamical downscaling model.

We agree and added “..., i.e. a dynamical downscaling approach.” to the end of this sentence.

12. “The motivation for CorrDiff is that naively applying conditional diffusion models to directly learn $p(x|y)$ was unsuccessful. The large uncertainty of variables such as radar reflectivity at these scales requires substantial noise levels during the forward process demands a large number of steps in backward process. This impedes learning and leads to poor sample fidelity”. This seems like a very interesting point, and seems quite central to one of the key points of this paper, are the authors able to elaborate on why other studies may not have come across this problem? Intuitively you might expect that downscaling multiple variables simultaneously requires more training time, might this also be a factor?

Indeed, the reviewer’s intuition is correct. We have observed that if the input (coarse resolution) and output (high resolution) are misaligned, conditional diffusion models struggle to learn a generative map capable of producing coherent images. This often results in hallucinations. The challenge becomes more pronounced when handling multiple variables simultaneously due to the increased dimensionality of the data.

In light of the reviewer’s suggestion, we have revised the corresponding section in the manuscript to clarify this important message as follows:

“The development of CorrDiff was motivated by the limitations observed when using conditional diffusion models to directly learn $p(x|y)$. This approach showed slow convergence and resulted in poor-quality images with incoherent structures. This was surprising because conditional diffusion models have been successfully applied to super-resolution tasks in natural image restoration, as demonstrated in works like Rombach et al.⁷. We hypothesize that the significant distribution shift between the input variables and challenging target variables, particularly the 1-hour maximum derived radar reflectivity (referred to as radar reflectivity), necessitates high noise levels during the forward process and numerous steps in the backward process. Our experiments indicated that these requirements hindered learning and compromised sample fidelity, as noted by Song et al.⁸. This issue is particularly relevant for the downscaling task, which must account for large spatial shifts, correct biases in static features like topography, and synthesize entirely new channels like radar reflectivity. By comparison, the task of super-resolution in natural images is much simpler, as it typically involves local interpolation and does not face the same level of distributional challenges.”

13. “Based on law of total variance” missing ‘the’

Corrected.

14. Figure 2 (a) uses superscripts in the y label, but looks like something has gone wrong with those for Figs 2 (b) and (c)

Corrected, thank you.

15. I found the Results section confusing to read, think it would be clearer if the baselines (RF and Reg) are defined in Sec 2. Two points here:

⁷Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-Resolution Image Synthesis with Latent Diffusion Models. arXiv preprint arXiv:2112.10752.

⁸Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2020). Score-Based Generative Modeling through Stochastic Differential Equations. arXiv preprint arXiv:2011.13456.

1. I didn't find it clear how you have used the data. Was a portion of the training data used for validation? If so, how much / what dates? And it would be good to state how the test data is used (i.e. 2021 used for skill scores, etc.). It seems like the 2022 data is only used for analysis of the cold front, but it's not clear why, and I can't see this written anywhere.

We regret the confusion and are grateful for these suggestions and questions to improve other readers' experience. In response to this and the following points we have grouped details of training and baselines into a short paragraph at the end of Section 2 that now reads as follows:

“Our target (WRF) dataset spans 2018 through 2021 at hourly time resolution. We use 2018 through 2020 for training and the rest for testing. We additionally use several days of typhoon data from 2023 and some snapshots of a coherent frontal weather system from 2022 for testing case studies. The input (coarse resolution) data are taken from the ERA5 reanalysis for the corresponding times. A UNet regression and random forest are trained as baselines. See 5 and S2 for details.”

16. Are all the training and test years fairly homogeneous? Particularly with respect to events like typhoons, since this has bearing on your case study in Sec 3.3.2.

To address this question, we have generated Figure 1 below (now also in the revised paper's Appendix) which shows a time series of domain mean, and the mean $\pm$ 2 and 4 standard deviations of the normalized channels. There does not seem to be a notable inter-annual trend in the data. As for extreme events: the training data contains 22 time instances stemming from 5 separate named Typhoons of category 1 or higher in the domain. From a rough estimate, frontal systems tend to pass through the domain between 4-8 times per year and typically last 8h in the domain (order 100 snapshots with a front).

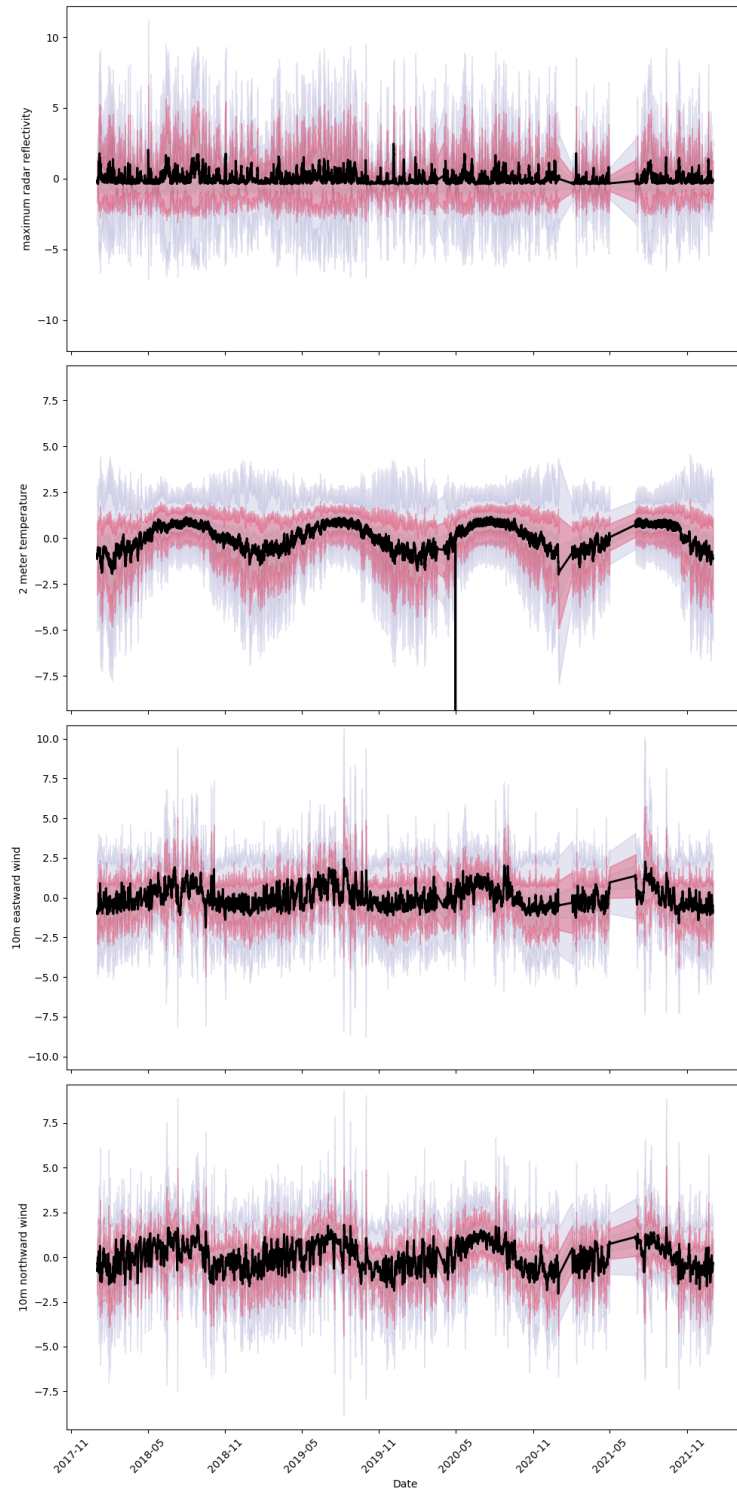


Figure 1: Time series showing the mean (black line), mean $\pm 2 \times$ standard deviation (red shading) and mean $\pm 4 \times$ standard deviation (grey shading) for the 4 channels. Statistics are computed latitude and longitude at each date and time in the dataset.

17. “Compared with several baselines, CorrDiff shows the highest deterministic skill (MAE) followed by the UNet regression (Reg), the random forest (RF)” I think it would make it much clearer to introduce these baselines in Section 2. Also, why not just call the UNet model UNet? it would be easier to read the results.

Thank you for this suggestion. We have implemented your suggestion by renaming “Reg” to “UNet” throughout to improve readability. We have also taken the reviewer’s suggestion to move the description of the baselines to a new subsection (3.1), as described in our reply to Minor Comment 15.

18. “Skill scores evaluated from 204 date and time combinations taken randomly from the out-of-sample year (2021).” Two points: (i) This seems like a small number of samples, given the total available of around 8000. How can we be confident with this few samples that the results are significant, and truly represent the behaviour of the model over the whole year, and for all times of day? I suggest using many more samples, or at least using a method like bootstrapping to estimate sampling variability. (ii) Are all metrics and models evaluated on the same 204 date/time combinations? Or do they all use different random samples?

(i) Please see reply to Major Comment 1 regarding statistical validity of the scores.

(ii) All metrics are spectra and distributions are evaluated on the same 205 (204 was a type-o) dates and times. To avoid such confusion we have consolidated these important details in the beginning of the revised results section as follows:

“For computing the metrics and spectra of CorrDiff and the baseline models we use the same random sample of 205 date and time combinations from 2021 as a validation set. For CorrDiff we generate a 32 ensemble member as a larger ensemble did not modify the results (not shown).”

19. “A separate RF is fit with scikit-learn for each of the 4 output channels with 100 trees and the default hyperparameters.” Isn’t this then an unfair representation of the suitability of Random Forests to this task?

Respectfully, it is unclear to us if the reviewer means that Random Forests are suitable or unsuitable for this task. Unfortunately, this is not a benchmark dataset so a commonly validated set of baselines is not available. We still wanted to present some comparisons, so we chose something simple. Because radar is not in the input set we could not use bilinear interpolation as this baseline. We chose RFs since they are a common off-the-shelf method deployed in atmospheric sciences—including for down-scaling (<https://doi.org/10.1002/2016WR019034>). We do not make any strong claims about the suitability of RFs, and did caution that “While crude, this RF provides a simple (and easily tuned) baseline for the performance of the UNet.”

We have moved the description of the RF from the figure caption to the main text (Section 3) and add some explanation:

“A separate RF with the 20 low-resolution input channels is fit with scikit-learn for each of the 4 output channels with 100 trees and the default hyperparameters. The RF is applied independently at each horizontal location similar to a 1x1 convolution. While crude, this RF provides a simple (and easily tuned) baseline for the performance of the UNet. We train separate RFs for each channel to ensure the best performance for each channel individually.”

20. In section 3.2 you switch from using 204 samples to 192 samples. It’s not really clear why, and again this seems like a very small number of samples to draw robust conclusions from. See point above about

	Radar	t2m	u10m	v10m
CorrDiff	1.64	0.5	0.73	0.82
UNet	1.98	0.53	0.89	0.99
RF	2.91	0.62	0.91	1.01
ERA5	-	0.83	0.97	1.05

Table 1: pooled CRPS counterpart table for Table 1, where the prediction and target from each model are pooled on 28 grid boxes.

bootstrapping.

Thank you for pointing this out, we apologies for the inconsistency.

“In section 3.2 you switch from using 204 samples to 192 samples. It’s not really clear why”:

This is a typo, we have unified the terminology, now with 32 samples:

“These results reflect reductions over space, time and for *CorrDiff* across 32 independently generated samples”

“and again this seems like a very small number of samples to draw robust conclusions from”

Please see replies to Major Comment 1 and Minor Comment 18 (ii).

21. One limitation in this section is that it is pixel-wise comparison, and uses only a univariate measure (CRPS). To make these results more compelling, it would be interesting to see results of a neighbourhood score (e.g. pooled CRPS as used in Ravuri et al 2021, Fractions Skill Score, etc).

Many thanks to the reviewer for this constructive suggestion that has added additional depth to our scoring results. We added a counterpart of Table 1 with pooled CRPS and pooled MAE in Table S3 in the SI section.(See also Table 1 below). The pooling changes the CRPS values but the relative score between models is not changed and therefore we relegated this result to the appendix.

22. “In contrast, for 2-meter temperature (Fig. 2 b) the performance of UNet-regression is comparable to CorrDiff;” also seems true for kinetic energy; particularly since with only 192 samples, variability in these results may be high.

We have already addressed the reviewer’s concern about sampling uncertainty with regard to our scalar score metrics in our reply Major Comment 1.

To address the reviewer’s related concern that sampling uncertainty might also obscure model intercomparison from the vantage point of spectra and probability distributions, we have computed a modified version of the associated figure from our initial submission (see Figure 2 below) in which error bounds from bootstrapping using a 10% sub-sample in time (20 times randomly selected from the available 205 times, with repetition) are included. The grey shading around the yellow line for the diffusion model now denotes one standard error computed from 1000 bootstrap subsamples, from the population of means. The fact that the grey line is barely visible across most of the analysis confirms that 205 random times across the year was a large enough sample to assess the performance of the model, even from the vantage point of intercomparing spectra and PDFs.

Regarding the kinetic energy spectrum, in our revised results we have found that the diffusion does add nontrivial variance across a particular length scale regime. The updated narrative around our power spectrum results now reads as follows:

“Relative to deterministic baselines, CorrDiff significantly improves the realism of power spectra for 10-meter kinetic energy (KE), 2-meter temperature and synthesized radar reflectivity. Variance missing from the UNet is restored by the corrective diffusion (blue-dashed vs blue-solid) – especially for the radar reflectivity channel at all length scales (Fig. 2c), but also for kinetic energy between 10–200 km length scales (Fig. 2a) and to a lesser extent for temperature on 10–50 km length scales (Fig. 2b). Temperature downscaling is an easier task that is expected to be mostly driven by sub-grid variations in topography that can be learned deterministically from the static grid embeddings.”

23. “CorrDiff is able to match the target distribution, including the heavy-tailed structure, which is significantly improved from the the prediction based on the UNet alone. These rare extreme values reflect most of the risk.” It isn’t clear what percentile these values are, and so is hard to tell how extreme they are. Please elaborate what percentiles these are, or else a similar number like return period. Given the small number of samples, it also seems unlikely that your testing samples shed much light on really extreme regimes.

We have revised discussion of our probability distribution analysis to avoid emphasis on tail structure in recognition of the reviewer’s concerns about sufficient sampling of low likelihood/high-impact extremes. The revised narrative now emphasizes how the corrective diffusion has outsize effects on the radar channel synthesis in comparison to other channels and includes the reviewer’s caveat about validating tail structures. Please note that with the revision of our ML model CorrDiff offers some level improvement in each of the channels relative to the UNet.

We remove the statement about extremes and wrote explicitly the values where CorrDiff improves upon the UNet:

“Evidently, synthesizing radar from only indirectly related inputs is the task that most benefits from the corrective diffusion component of CorrDiff.

This is corroborated by analysis of probability distributions (Fig. 2d-f) – for the radar reflectivity channel both the UNet and RF fail to produce realistic statistics, but CorrDiff is able to match the target distribution between 0 and 43 dbz while significantly improving on the UNet (Fig. 2f). In contrast to the radar channel, the hot and cold tails of the CorrDiff-generated surface temperature distribution are only incrementally improved relative to the UNet (Fig. 2e) and the overall windspeed PDF is virtually unchanged relative to the UNet, despite the scale-selective variance enhancements noted previously. Overall, CorrDiff produces encouraging probability distributions, with the caveat that apparent agreement of generated tail structures should be viewed as provisional given that our chosen validation sample of 205 independent calendar times imperfectly samples especially low likelihood/high-impact extremes.”

24. “For windspeed (Fig. 2 d), the UNet and the baseline models underestimate the probability of winds faster than 20 m/s.” Perhaps I’m misreading the graph, but in Fig 2d it looks like The Reg, CorrDiff and target are basically indistinguishable, so this statement doesn’t seem to be true?

Please see the revised manuscript text included in our replies to Minor Comments 22 and 23. We agree with the reviewer that - even after the updates to our ML model - the diffusion component of CorrDiff does not significantly impact the tail of the wind speed PDF, despite some scale selective enhancements to the

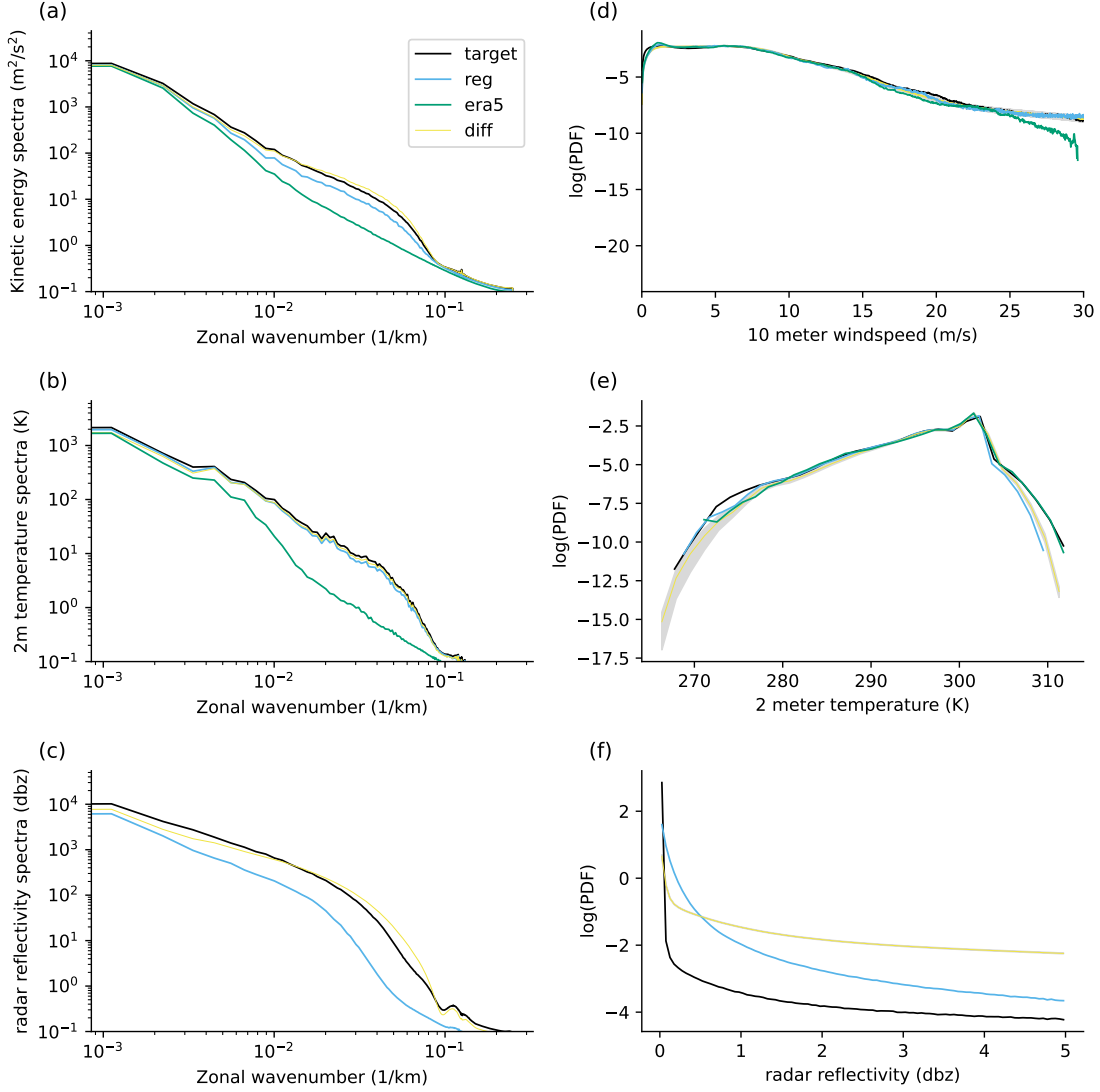


Figure 2: Assessment of sampling uncertainty on diagnostics of spectra and distributions, using bootstrapping of 1000 random subsamples with replacement, each using only 20 samples in time from the 205 times (10%). The solid yellow line labeled as *diff* denotes the diffusion model's average of all subsample means, and the gray shading around it denotes $\pm$ one standard error from the bootstrap ensemble. The standard error is barely visible in some plots and visible and small in others, confirming that sampling uncertainty is small for such diagnostics. Note: This analysis was done with the ML model that was originally submitted, as one of our first steps in the revision, and has not been repeated with our updated ML model as we don't expect the main finding would change.

kinetic energy power spectrum, and hope the reviewer agrees the revised narrative accurately reflects the revised results.

25. “For 2-meter temperatures (Fig. 2 e), the baseline models overestimate” not clear what you mean by baseline; I would interpret this as everything except CorrDiff. But here this seems to exclude UNet, or at least it should, as the UNet curve seems to align almost exactly with the target

Please see replies to Minor Comments 22–24.

26. “the match of slopes at high values attests that the power-law relationship regulating the heavy-tail structure of rain extremes is emulated skillfully.” Hard to tell what happens at the high end; maybe a log-log plot would be more informative, to see how the Diffusion model performs at the end of the tails?

Removed. Please see reply to Minor Comment 23.

27. “the variability of the generated radar field at four distinct times.” How were these times chosen? At random?

Good question. The associated text has been revised as follows:

“Fig. 4 illustrates the variability of the generated radar reflectivity field on four separate dates corresponding to distinct Taiwanese weather events. Two dates are chosen randomly (Fig. 4e-k). The other two correspond to a typhoon (Fig. 4a-d) and frontal event (Fig. 4m-o); more examples of both of these phenomena are provided in the Appendix for additional context.”

Please note especially the new suffix – to address the reviewer’s valid implicit concern that case study results can be unrepresentative, we have added considerably more sampling of frontal and typhoon events to the Appendix so the interested reader can take in more of the case-to-case variability.

28. “When analyzing 200 samples” My understanding is you mean an ensemble size of 200, but this isn’t that clear from the text.

Please see reply to Minor Comment 18 (ii). Here and elsewhere we have clarified our discussion of CorrDiff-generated ensembles. In the revised manuscript all ensembles have 32 members.

29. “Figure [3] would be a lot clearer by adding a title for each column (like you have done in Fig 4). Units required on colorbar axes

The suggested figure improvements have been implemented, with thanks.

30. Fig. 4 shows an example of CorrDiff downscaling a cold front.” How is this chosen? Is this typical of the kind of frontal predictions seen?

This event was chosen qualitatively, as a particularly coherent frontal case with weak background temperature gradients and winds, facilitating a clear illustrative analysis. To help readers better assess the representativeness of the frontal event case study, which we agree with the reviewer is important, we added

a new section in the appendix (see section 6.1) where we further analyzed this event as well as an event in 2023 to bolster the claim about the coherent synthesis of reflectivity at the frontal boundary and sharpening of wind and temperature fronts by CorrDiff.

31. “Cold front on Feb 2, 2022 at 20 UTC.” Inconsistent formatting of times throughout document.

Thanks you for this suggestion. We have unified time formatting across the document.

32. “Both the location of the front, and the magnitude of the horizontal wind and temperature gradients (sharpening of the front) associated with it, are captured well by CorrDiff although some mispositioning of the exact front location is inevitable” This is a strange sentence, it sounds like you are saying that CorrDiff predicts the front location well, whilst also it is inevitable that it doesn’t predict it well...

We have rewritten the narrative surrounding frontal super-resolution as follows and hope that it is clearer and also more precise:

“Fig. 5 shows an example of CorrDiff downscaling a cold front. Examining the target data (WRF in third column), the position of the front is clearly visible in the southeast portion of the domain, where a strong horizontal 2-meter temperature gradient (top) co-locates with both a strong divergence of the across-front wind (bottom) and a reversal in direction of the along-front wind on either side of the temperature front (middle). Compared to the target data the ERA5 representation of this front is smoother. CorrDiff partially restores sharpness to the front by increasing the horizontal gradients across all three field variables. Although the generated front has some differences in morphology compared to the ground truth, the consistency of its morphology across winds and temperature is reassuring. The intense rainfall associated with the convergence at the front can be seen in the radar reflectivity for the same date and time in bottom row of Fig. 4 above. The generated radar reflectivity is appropriately concentrated near the sharpened frontal boundary at the cold sector. We expand this analysis across more samples of frontal boundaries in the Appendix Section 6.1, which reveals that CorrDiff consistently adjusts the winds, temperature and high resolution radar reflectivity at the front, but that its skill in sharpening frontal gradients exhibits case-to-case variability.”

33. 3.3.2 Tropical Cyclone case study: Not clear why you are not showing results for the UNet (Reg) here, since it seems to also perform well at improving the winds?

We appreciate the reviewer pointing out here and elsewhere that the UNet performs a significant portion of our task, including for wind speed enhancement of typhoons. This finding has survived the corrections and updates to our revised ML model and we have implemented the reviewer’s suggestion to show the contribution of the UNet to all Typhoon wind super-resolution figures.

34. “Further analysis expands beyond this” This is meant to link to supplementary info but shows up as a link to Methods

Thank you for catching this broken and mislabeled section link. It has been resolved, alongside significant narrative updates in the associated section.

35. “Although no target data is available, when comparing the maximum windspeed and radius of maximum windspeed from CorrDiff predictions to a reference from the Japan Meteorological Agency best track dataset [4] CorrDiff is found to correct at least 75% of the error in intensity for windspeed below 50m/s but only 50% of the error for higher windspeeds S4.” Again I’m not sure why UNet isn’t also included in this comparison, as it would be insightful to see if the corrections here are due to the deterministic or probabilistic component.

We have expanded this analysis to include the UNet as requested. Please note that with the ML model used in the original submission the reviewer is probably correct that the UNet provided almost all of the improvement, but in the revised model the diffusion also contributes a secondary component.

36. “An attractive use case for CorrDiff is to directly downscale forecasts from global models.” I’m not really sure what the value of this analysis is; it seems to say more about how close GFS forecasts are to ERA5 than anything else. Is there any use-case where we would apply this model trained on ERA5 to GFS data, instead of just training it on GFS in the first place? These results also make CorrDiff appear worse at first glance, until you realise that it is on different data.

Thank you for pointing out this intrinsic tension of our fledgling attempt towards zero-shot forecast post-processing. On balance we decided to remove this section to avoid unnecessary confusion and distraction from our more fundamental points. In the revised manuscript we have opted instead to point to this as a logical direction for community work.

36. “the model is shown to skillfully correct kinetic energy spectra, generate realistic probabilities and downscale coherent structures accurately”

Two points: - Whilst this is true, it also misses that the UNet alone does almost as well on the kinetic energy spectra (and its hard to tell how significant this difference is given the small number of samples used). It seems like the main thing that differentiates CorrDiff from the UNet is the ability to represent complex spatial fields like radar reflectivity...

We agree that – even after our updates to the ML model – the diffusion subcomponent of CorrDiff has outsize benefits for the radar reflectivity synthesis task compared to super-resolution of other channels and hope the reviewer agrees the revised manuscript now emphasizes this at high levels; see e.g. revised text in response to Minor Comment 23. Please see response to Minor Comment 22 regarding our updated findings about the significance of the role of diffusion in improving kinetic energy spectra given the revised regression model.

37. ...There isn’t much to go on to infer that the probabilities are realistic, since there is no target ensemble data to compare with. Again, I suggest some kind of analysis such as spread-error to back this up.

We hope the reviewer is satisfied by the new spread-error and rank histogram analysis in Section 3.3 of the revision, which clarifies the limits of this model’s probabilistic calibration, and thank them again for encouraging us to add this. Associated claims on extremes and probabilistic performance have likewise been revised throughout. We are sure that the reviewer agrees a truly satisfying examination of probabilistic performance would include the use of CorrDiff in the context of an end-to-end forecasting pipeline. While beyond the scope of this study this has been more clearly suggested for future work in the revised conclusions.

38. Methods: I am not sure why Methods are after Discussion. Is this a specification of the journal? It seems like this would be a lot clearer if a lot of this came before the results and was combined with some of the

other background.

The reviewer is correct that these are journal specifications. In *Nature Communications Earth and Environment* the Article main text section does not include the Methods see guidelines: “Ideally, the main text (not including Abstract, Methods, References and Figure legends) should be limited to 5,000 words.”

We regret that this style, which is not our typical preference either, made the manuscript more cumbersome to review.

39. “5.1 Background on diffusion models” This seems like an overly detailed background on diffusion models; are there not other existing articles on diffusion models that the reader could be directed to for further details? Or else can this part moved to the Supplementary information?

We hope the reviewer agrees it is natural for reasonable people to have varying preferences on such matters of writing style. Given that our target audience is broad and the readership not generally ML experts, we decided to keep this section in the main paper for the sake of completeness. This succinct summary helps the non-expert readership better understand the mechanisms behind diffusion models for CorrDiff.

40. “we resort to UNet-regression. A UNet network is supervised with training data $(\mathbf{x}_n, \mathbf{y}_n)$ $N_n=1$ to learn the regression. We adopt a UNet architecture that is commonly used in denoising diffusion models.” Lots of repetition

Repetition has been reduced in the indicated paragraph and the text condensed. Thanks to the reviewer for helping improve the clarity of this prose:

“In order to predict the conditional mean $\mu = \mathbb{E}[\mathbf{x}|\mathbf{y}]$, we employ to a UNet-regression model trained on a dataset $\{(\mathbf{x}_n, \mathbf{y}_n)\}_{n=1}^N$. We utilize the specific architecture described in Song et al., (2019) that incorporates attention and residual layers, allowing it to effectively capture both short and long-range dependencies in the data (see section 2 and Fig. S1). The model is optimized using a Mean-Squared-Error (MSE) loss during training.”

41. “Each sample in the input dataset consists of 20 channels of information see ??.” Broken link to section.

Thank you for catching this latex compilation error, which has been resolved. Please note that the input actually contained only 12 channels, which has been corrected throughout (see next comment).

42. “This includes four pressure levels (500 hPa, 700 hPa, 850 hPa, and 925 hPa) with four corresponding variables: temperature, eastward and northward components horizontal wind vector, and Geopotential Height. Additionally, the dataset includes single level variables such as 2 meter Temperature, 10 meter wind vector, total column water vapor.” Any explanation of why these inputs / pressure levels are used?

There was a typo in the original submission. We used 12 input channels. The 20 channels listed are from earlier ablations but we found no meaningful difference.

To the reviewer’s question of how we chose these variables, please refer to the bolded excerpt in the revised version of this paragraph:

“Each sample in the input dataset consists of 12 channels of information; see Table S2. This includes two pressure levels (500 hPa and 850 hPa) with four corresponding variables: temperature, eastward and northward components of the horizontal wind vector, and geopotential height.

Additionally, the dataset includes single-level variables such as 2-meter Temperature, 10-meter wind vector and total column water vapor. **This input channel set is admittedly somewhat arbitrary, but serves the purposes of (i) allowing a reasonable sparse representation of the thermodynamic state of the macro-scale atmosphere, while (ii) intentionally including only limited information about atmospheric water via the total vertical integral of an invisible trace gas - water vapor; this ensures the task of synthesizing radar reflectivity – a much more complex observable that relates to the sixth moment of the cloud water droplet distribution – is appropriately ambitious, as a strong test of CorrDiff’s generative component.”**

43. “Subsequently, vertical interpolated from sigma levels to isobaric levels then save as a custom CWA DMS format.” Check grammar

Resolved, with thanks. The indicated section now reads as follows:

“A vertical interpolation from hybrid coordinates (i.e., sigma levels) to pressure coordinates (i.e., isobaric levels) is applied.”

44. “Additionally, any missing or corrupted data points represented by “inf” or “nan” values are eliminated from the dataset.” Can you clarify? you mean any sample which has one or more nan/inf value is removed?

Yes, all instances where NaN or Inf values were found in the data were removed from the dataset. This process effectively reduced the number of samples from 37,944 to 33,813.

45. To avoid over-fitting we divide the data into training and testing sets. Three years of data 2018-2020 are used for training (2,4154 samples total). For testing we use the full fourth year, 2021, as well as the first four months (January to April) of 2022.” What about a validation set? How did you tune your hyperparameters?

Thanks for pointing this out. The aforementioned test set includes validation data as well. Specifically, 205 samples from 2021 were used for testing, based on which we have reported the metrics in Table 1. Although the remainder of 2022 data was left as validation data, there was no hyperparameter tuning involved as the EDM framework is auto-tuned. Note that we adopted most of the hyperparameters from the EDM paper. The advantage of the EDM framework is that it offers an almost tuning-free diffusion originating from ODEs. The critical hyperparameter to tune is the noise standard deviation intensity (p_{mean}) in the forward noising process. We adjust this based on the WRF data resolution and dynamic range, following the guidelines proposed in the Simple Diffusion paper by Hooeboom et al. (2023) ⁹.

To clarify this issue, we revised the manuscript to include:

“Our target (WRF) dataset spans 2018 through 2021 at hourly time resolution. We use 2018 through 2020 for training. For testing we used 205 random samples from 2021, and the rest used for validation. We additionally use several days of typhoon data from 2023 and some snapshots of a coherent frontal weather system from 2022 for testing case studies.”

For the choice of hyperparameters we add:

⁹Hooeboom, Emiel, Jonathan Heek, and Tim Salimans. “Simple Diffusion: End-to-End Diffusion for High Resolution Images.” International Conference on Machine Learning. PMLR, 2023.

“We adopt the hyperparameters based on the guidelines proposed in EDM, Karras et al., (2022), including the same optimizer hyperparameters and learning rate schedule. EDM offers a physics-inspired design space based on ODEs that can be auto-tuned to our scenario (see table 1 in Karras et al., (2022)). The only relevant hyperparameter that is tenable in this framework is P_{mean} of the noise schedule. The value of this parameter was selected based on the resolution and dynamic range of the data as done in Hooeboom et al. (2023).”

46. “To ensure compatibility and consistency, we employ the same UNet architecture used in EDM (Elucidated Diffusion Models) for both the regression and diffusion networks(see S1). This architecture is based on the UNet model proposed in [53].” Repetition

This repetition has been removed.

47. “For time embedding in the diffusion process, we utilize Fourier-based position embedding. However, in the regression network, time embedding is disabled.” Why no time embedding for the U Net? How do we know what benefits in skill are down to diffusion model vs just knowing the time of day?

“Time” in this case refers to the evolution of the diffusion process, not the time of day. We apologize for the ambiguous terminology, and have modified the text to include ‘diffusion-time’ that is distinct from time of day, and explain its origin:

“Here ‘diffusion time’ is a virtual variable used for denoting denoising steps and has roots in differential equations that are derived from Langevin dynamics. To avoid confusion with the time of day, we use diffusion time to denote this variable.”

48. “Additionally, we introduce 4 channels for sinusoidal positional embedding.” Please provide reference if not describing exactly what this is.

We have added position embedding to the UNet architecture to improve spatial consistency, following established practices in the field. Sinusoidal positional embedding, originally introduced in the context of transformers by Vaswani et al. (2017), is a widely adopted method in both CNNs and transformer-based neural networks due to its effectiveness in capturing positional information. This technique has also been utilized in various computer vision tasks to enhance spatial encoding in models like Vision Transformers (Dosovitskiy et al., 2020) and other advanced neural network architectures (Carion et al., 2020). We have clarified this in the revised manuscript and provided appropriate references to support our approach. References:

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems* (pp. 5998-6008).
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-End Object Detection with Transformers. In *European Conference on Computer Vision* (pp. 213-229). Springer, Cham.

49. “During the training phase, we use the Adam optimizer with a learning rate of ...” Need to explain how you chose these numbers; are these all based on ref [27]? Or chosen during hyperparameter tuning? It would be good to specify this at this point.

Yes, we adopt the optimizer hyperparameters from EDM as they perform well and are stable in our setting. We have clarified this in the revised manuscript as follows:

“We adopt the hyperparameters based on the guidelines proposed in EDM, Karras et al., (2022), including the same optimizer hyperparameters and learning rate schedule. EDM offers a physics-inspired design space based on ODEs that can be adapted to our scenario. For instance, we adjust the lognormal noise schedule of EDM by shifting the $p_{mean}(0.0)$ to account for the higher dynamic range and resolution of our dataset.”

50. “The regression network solely receives the input conditioning channels. On the other hand, the diffusion training incorporates the 20 input conditioning channels from the coarse-resolution ERA5 data, which are concatenated with 4 noise channels to generate the output for each denoiser.” This isn’t very clearly written. I believe you mean that both UNet and Diffusion model take 20 input conditioning channels, and the diffusion model also takes in an additional 4 noise channels. But the wording makes this hard to read.

We revised the sentence, please note that there was a typo and the number of input channels is 12 not 20. The sentence now reads:

“The regression network receives only the 12 input channels from the ERA5 conditioning data. In contrast, the diffusion training concatenates these same 12 input conditioning channels from the coarse-resolution ERA5 data with 4 noise channels to generate the output for each denoiser. To further enhance the diffusion conditioning, we also added the mean obtained by the regression model in the first stage. This addition provides more context and global information to the diffusion process, potentially improving its convergence.”

51. “A necessary condition for calibration is that the spread-error relationship be 1-to-1 when averaged over sufficient samples”. It seems strange to discuss spread-error and rank histograms and then not use them. Additionally, something like a spread-error plot would provide useful information about the calibration of the CorrDiff ensemble.

Please see reply to Major Comment 1 and Minor Comment 37.

52. “CRPS is defined as” Please define the step function symbol used.

Revised.

53. “For a deterministic forecast, $F(y)$... needs $\geq$ symbol

Implemented

54. “CRPS is equivalent to the mean absolute deviation.” I think its clearer to refer to it as mean absolute error, since this is the standard way to say this.

Implemented.

55. “To the best of our knowledge ML downscaling of tropical cyclones, which requires accounting for different physics, across many channels and channel synthesis was not shown before.” If this is the case, then I think this needs to be emphasised in the main body of text.

Please see response to Minor Comment 8.

56. “In order to examine the realism of individual samples from CorrDiff and their quality compared with the target data figure shows an animation”

Unfortunately, this animation as part of a PDF file will only work if the PDF is opened with Adobe Reader, we was provided in the revised caption a link as an alternative method for downloading and viewing the animation:

“Comparative analysis of maximum radar reflectivity across multiple samples: Diffusion model predictions versus UNet regression and WRF target simulations for diverse cloud regimes. This figure is presented as an animated visualization, viewable as a video in Adobe-compatible formats. The full animation can be accessed and downloaded from <https://figshare.com/ndownloader/files/48060031>.”

57. “Comparison of simulated Typhoons with historical records”: I’m not sure why the results in this section aren’t more prominent. It’s hard to draw robust conclusions from the case study, but this at least gives some insight into the overall performance of CorrDiff.

We have the same preference, especially for expert readers such as yourself who are well equipped to discern the importance of such sampling. However, for the purpose of communicating foremost to the broad interdisciplinary readership of *Nature Communications*, we chose to focus our main narrative real estate on case studies, as they give an easy to communicate qualitative sense of the performance, complementing the technical depth in the SI. We are grateful to the reviewer for recognizing the importance of such sampling, which has inspired us to add even more to the Appendix: We extended our case study analysis to 10 time instances spanning two separate typhoons. Meanwhile, this response encouraged us to add more nuance to discussion of the one illustrative typhoon time instance in the main narrative, which has been appended as follows:

“For further discussion of typhoon downscaling, we refer the interested reader to SI Section 6.2, where we explore additional date-times for Haikui (2023), investigate an additional typhoon Chanthu (2021), and analyze generated wind statistics across a 600-member ensemble of typhoon-containing time intervals spanning the 1980-2020 period. This extended analysis suggests that, while the main results emphasized for are case study above are qualitatively representative when typhoons are far from land, the diffusion component of CorrDiff frequently plays a stronger role in intensifying typhoon axisymmetric structure, and CorrDiff tends to lead to too much horizontal contraction of cyclone morphology, predicting a radius of maximum winds that is statistically too small.”

58. “One imperfection of CorrDiff is that it shrinks all storms, including those with the correct size or those that were too small in the ERA5 input data (panel a).” This seems like a key finding that should be reported in the main text.

We agree and please note that this has been intentionally included (see reply to Minor Comment 57).

59. Y labels on Figs. S2 c and e are m/s, inconsistent.

Revised, with thanks.

60. Figure S3 seems to be missing?

Revised, with thanks.

61. “Panel (a): abscissa: mean (dot) and one standard deviation (bars) of predicted radius of maximum winds for the ERA5 (red) and CorrDiff (orange) storms, grouped by their observed values; ordinate: corresponding observed values.” Written in quite a confusing way. Referring to abscissa and ordinate seems redundant here. E.g. could be written “Panel (a) mean (dot) and one standard deviation (bars) of predicted radius of maximum winds for the ERA5 (red) and CorrDiff (orange) storms, grouped by their observed values”.

We have re-written this caption as follows:

“Comparison of predicted vs. observed (a) radius (km) and (b) value (m/s) of maximum axisymmetric windspeed (m/s) for all typhoons found in the domain with windspeed of 30 m/s or more spanning the years 1980-2020. Predictions are compared via the mean (dot) and one standard deviation (bars) of all predictions falling within each observational bin of the reference JMA observations. Panel (c) shows the log(PDF) of the predicted windspeed flattened across all of these typhoon instances.”

3 Reviewer 2

The authors propose a novel approach to downscale global reanalysis or forecast data to a reliable high-resolution ensemble with a horizontal resolution of 2 km. They have trained the combination of a U-Net and Corrector Diffusion-network on a dataset of 4 years of ERA5 data and the corresponding high-resolution WRF forecasts (with radar DA). The area under study was a region surrounding Taiwan, and multiple variables were used as input (20) and output (4) channels. Precipitation was not explicitly added as an input channel, but only as an output channel.

The authors highlight the fidelity of their method in reproducing WRF’s spectra and distributions for rainfall and 10m wind. They show its enhanced performance for CRPS for the test dataset, the improvement for wind speed for historical typhoon data as well as the performance for a “live” forecast of typhoon Loinu. The method substantially reduces the required computational power (1.300 times more energy efficient than the traditional forecasts).

3.1 Overall assessment and recommendations

The proposed method is an important, timely and original contribution to ML-driven weather prediction, potentially enabling the generation of fast and reliable high-resolution forecasts even with a modest training dataset. The authors position their approach accurately within the field. Their approach to separate the learning process into a U-Net for the mean and a CorrDiff-network for the more localized residuals appears to be highly effective. Moreover, the methodology has a lot of potential not just in the context of downscaling, but also e.g. for the fast generation of large ensembles from high-resolution NWP forecasts, without the need

for an ensemble training data set.

I recommend that the authors make the architecture or code available to the scientific community. This would greatly increase the impact of this work and, most importantly, warrant the research results' reproducibility.

The manuscript is well-written apart from some minor typographical errors, some of which are listed below. The structure of the manuscript, in which the results and discussion precede the details of the methodology, is unconventional but reads well.

We thank the reviewer for their favorable assessment of our work and are happy to reassure that the architecture, training pipeline, training dataset, and checkpoints needed to reproduce the results are all definitely being made available to the scientific community via NVIDIA's open-source Modulus product.

3.2 Specific remarks

There are some details in the methodology which were not entirely clear to me, so I would appreciate if the authors could clarify these points:

1. P2: You mention the path towards an end-to-end ML downscaling system. What do you mean by that, exactly (and how is this more end-to-end than previous ML-based downscaling techniques)? Do you mean it generalizes to both reanalysis and different forecast models?

We were trying to convey that that the same pipeline that we have demonstrated can downscale from ERA5- to WRF-scales demonstrated here could be extended to downscale instead from ML global forecast models that were trained on ERA5 to equivalent. But, realizing this could be misconstrued or distracting, have opted to clarify the open challenges that would need to be resolved to arrive at end-to-end forecasting systems in the concluding discussion.

2. P2: What was the update cycle / frequency of the CWA's forecasts, and which lead times were used as target for the training? Please clarify this in the manuscript or S.I. The forecast skill is not constant in leadtime and may also suffer from spin-up effects, how was this dealt with?

We trained on analysis states not forecast states from the CWA, which were available at hourly frequency.

3. P3 and P9: Is there any potential for a pretrained model to be used over a different domain? Does it learn the local statistics, or the actual physics of downscaling? Latitude, longitude, orography, land-sea mask, etc. are not provided as input channels, so I suppose not. Would such a more generic model be feasible and what would be required? This would be highly valuable, considering the cost of retraining such a model.

We share the reviewer's curiosity about these issues and are actively exploring the question of generalizability through precisely these same boundary conditions, which indeed were not included as inputs in this initial proof of concept. Orography and land-sea mask were not represented explicitly, although horizontal location was provided via learnt spatial embeddings. It would be inappropriate to expect this initial version of CorrDiff to generalize to new regions. It is too early to say anything definitive on this generalizability topic as there is considerable research yet to do on it. For now, the hypothesis that appropriate boundary control of learnt local dynamics using extensions of CorrDiff is an interesting open question.

4. P3 and further: you refer to “generated rainfall reflectivity field”. Do you use a radar simulator to generate these fields? Why not just consider the (log) total precipitation (intensity or accumulation)? Reflectivity depends not just on precipitation intensity but also precipitation type, distance from the radar, etc. so may not be the most useful variable to look at. Moreover, it suggests that you are comparing to observations in Fig. 3, which you are not, if I understand correctly; you compared to WRF output.

The reviewer is correct that we are comparing composite radar reflectivity as “generated” by CorrDiff (via ML) against the equivalent from the WRF analysis states used for training, which in turn was indeed “generated” (via WRF’s radar simulator). This radar target can be viewed as an arbitrary variable choice to exemplify the challenges of emulating convective variables (please see reply to Minor Comment 42 by Reviewer 1), and we agree that the proof of concept could also have been done using precipitation or integrated cloud water as the target rather than synthetic composite radar reflectivity. An exciting future extension of CorrDiff would be to train on direct sensor nets of measured rainfall, or actual observed radar, which we have added as a ripe opportunity in the concluding discussion:

“

- **Synthesizing sub-km sensor observations:** To achieve effective resolutions beyond what is possible to simulate today, and sidestep issues of numerical simulation, it would be interesting to test whether variants of CorrDiff can be trained to generate raw sensor observations where dense networks exist. Our demonstrated ability to synthesize an observable as challenging as radar reflectivity from column water vapor should embolden such efforts.”

P11: time here is not related to time in the forecast, please make sure this distinction is clear.

We assume you mean the section titled ‘SDE formulation’, where we added:
 “Here ‘diffusion time’ is a virtual variable used for denoting denoising steps and has roots in differential equations that are derived from Langevin dynamics. To avoid confusion with the time of day, we use diffusion time to denote this variable.”

P12: only here it becomes clear that the precipitation-related output variable is in fact 1hr maximum radar reflectivity; please mention this earlier. Why look only at hourly maxima instead of accumulations, which are more useful to hydrologists and to calculate longer accumulations (e.g. 24h rainfall)?

Only the 1h maximum radar reflectivity was made available to us by CWA. The reviewer is correct that we should have made that distinction from 1h accumulated radar on its first introduction. We revised the manuscript to explicitly state that 1h maximum radar reflectivity is the radar channel and will be referred as ‘radar reflectivity’ hereafter.

P12: Is any other data than radar data assimilated in the CWA-WRF model – such as station observations, useful for correcting T2m? P1 mentions “radar and other sensors”.

Yes, the data assimilation includes surface observations that can recover some of the surface information radar often misses. The full process is best described in Chen et al., (2020) that is cited in the manuscript.

No real mention is made of the fact that this downscaling approach bypasses, or even replaces, data assimilation. Do I understand correctly that this approach yields an improvement wrt coarse global forecasts compa-

rable to WRF downscaling with DA, even though this approach does not assimilate the high-resolution data? One would assume that the most recent observations are required to fill in the missing information needed to improve model skill at higher resolutions.

The reviewer is correct that this approach bypasses regional DA but not the global DA. Since regional DA is vital to actual use cases of km-scale weather prediction we have removed the case study of operational weather forecasting in the revised manuscript to avoid distraction. Please also see our response to your next comment below.

P13: It is not clear to me from the architecture of the CorrDiff method how you ensure continuity, or realistic temporal autocorrelations, between the subsequent generated high-resolution fields (for a single member forecast). Realistic time correlations are important e.g. for hydrological applications. Could you clarify this?

The reviewer is correct in noting that issue. At the moment, there is no explicit method in which we ensure continuity in time, apart from that which is ensured by the conditioning data (ERA5). To better communicate these limitations that would prohibit a seamless application to weather forecasting, including the reviewer's previous comment, we have added the following paragraph to the concluding discussion:

“To become useful for km-scale weather prediction, extensions of CorrDiff are encouraged that include temporal coherence, such as via video diffusion or learnt autoregressive km-scale dynamics; as with super-resolution these must be formulated as stochastic machine learning tasks. Currently, beyond the coherence of the large scale conditioning given from ERA5 there is no guarantee that CorrDiff's km-scale dynamics will be coherent in time. Additional integrations with km-scale data assimilation are also essential for this use case. Our current demonstration relies only on the global data assimilation (DA) used to produce the ERA5 dataset. Unlike the target data it is trained on, CorrDiff effectively bypasses the regional DA, which for CWA is of similar computational cost as running the operational numerical model (WRF) model for 13h.”

Please explain the meaning of your acronyms (e.g. SDE, EDM...) and variables (ω in $d\omega(t)$...) upon first use.

Revised as suggested, with apologies:

“In our approach, we adopt the Elucidated diffusion model (EDM).. ”.

“Such models learn stochastic differential equations (SDEs hereafter) ...”

“Here ω in the forward SDE is a Wiener process, while the backward SDE comprises two terms: a deterministic component representing the probability flow ODE with noise degradation, and noise injection via the Wiener process.”

Typographical and other minor remarks:

P1: weather hazard -> weather hazards

Revised - thank you.

Good estimates ... exist-s-

Revised - thank you.

P2: Building from -> Building upon/on

Revised - thank you.

P3: Missing “and” in “The large uncertainty....”

Paragraph restructured.

P6: same calendar date: and time / hour?

Revised to “same date and time” - thank you.

P7 caption of figure 4:

“true wind vectors” do you mean “target”?:

Removed “true”.

"arrows are along front wind component" is a bit confusing, please rephrase

Rephrased to:

“Top to bottom: 2 meter temperature (arrows are wind vectors), along front wind (arrows are along-front component of the wind vector) and across-front wind (arrows are across-front component of the wind vector).”

P8: “it[is] most accurate” - sentence removed in revision.

“CorrDiff seem[s] to struggle” - sentence removed in revision.

P10: downscaling - revised thank you.

P12: Missing reference: “see ??” Revised.

hydrometers -> hydrometeors - revised thank you.

P13: “Subsequently, vertical interpolated from sigma levels to isobaric levels then save as a custom CWA DMS format.” - this sentence has grammatical errors and contains jargon without explanation.

Revised to:

“Initially, the target data is provided in the NetCDF format, which is the output of the WRFDA assimilation process. A vertical interpolation from hybrid coordinates (i.e., sigma levels) to pressure coordinates (i.e., isobaric levels) is applied. As part of the preprocessing steps, the data is converted to the Hadoop Distributed File System (HDFS) format.”

The supplementary information still has a lot of typographical and grammar errors and I suggest the authors carefully go through this document before resubmitting it.

Thank you for the constructive feedback, we have revised the SI section.

S3: I cannot see the animation, I tried multiple PDF readers. I am interested to see how realistic the time evolution is. Would it be possible to share this video another way? I've used TIB (<https://www.tib.eu/en/publishing-archiving/digital-preservation/>) in the past.

We apologize. We have added a link for download and edited the caption “The full animation can be accessed and downloaded from <https://figshare.com/ndownloader/files/48060031>.”

As for the caption, it is misleading to refer to WRF as “truth” - rather call it target (as you do elsewhere).

We agree and have revised truth to target throughout the text.

P22: You mention the improved efficiency of the inference, but this is only a part of the total picture. How many MWh did the training take, and how often do you estimate you will need to retrain?

While the reviewer is correct that training of this prototype of CorrDiff incurs substantial compute. There is a difficulty in adding this to the comparison for several reasons: First, one can argue that it is akin to the R&D stage of any numerical modeling system. Second, it is not clear how often it would need to be done, as for future improvement, smaller, fine-tuning approaches could suffice. Third, we expect at least an order of magnitude reduction in training expense as optimizations are implemented in our code.

Response to the Reviewers - Round 2

November 25, 2024

Dr. Morteza Mardani
NVIDIA Corporation
Santa Clara, CA

Email: mmardani@nvidia.com

Dear Editor:

Thank you for the opportunity to revise our manuscript for Nature Communications Earth & Environment. We have carefully addressed the remaining reviewer comments and completed the requested editorial adjustments to ensure compliance with the journal's requirements. We appreciate your guidance and look forward to finalizing this submission.

1 Reviewer 1

The authors have clearly put in a lot of effort to address concerns from the previous round of reviewing. Whilst the authors note they are unable to incorporate many samples due to time and space restrictions, they do include estimates of sampling uncertainty which demonstrate their results are robust. I also appreciate the addition of rank histograms and spread-skill plots which provide a clearer picture of how the ensemble behaves. Overall then I would recommend this revised version is accepted for publication.

Response. We appreciate the reviewer's positive feedback and are pleased that our revisions have strengthened the manuscript. We are glad that the sampling uncertainty estimates, rank histograms, and spread-skill plots provided clarity on ensemble behavior. Thank you for recommending this version for publication.

One comment I would make is in relation to my 3rd major comment in the previous review (which I can see probably wasn't that clear in retrospect). What I felt was lacking previously was a clear message about the relative utility of the UNet versus the Diffusion model for this application. Whilst there has been much improvement here, I would say there is still room to improve the strength of the message. For example, the power spectra seem to show that UNet and CorrDiff perform similarly on 2m temperature and 10m wind speed, so I think it would be illuminating to see the UNet represented in the Frontal System Case Studies so that readers can get an idea of the differences of these approaches.

Response. Thank you for highlighting this. We added the UNet to the analysis of the Frontal System Case Studies and modified the text accordingly:

"Examining the UNet cross sections (blue lines in all these figures) we find that its predictions often fall within the range of one-sigma (standard deviation) from the mean of the CorrDiff predictions. However, in some specific cases the diffusion step does offer additional sharpening of the frontal wind shear (see Fig. S5)."

Section 3.2: I'm not sure why the CorrDiff MAE is taken for the ensemble mean, why not just a single member ensemble? To my mind that seems the more natural quantity to report on, as the limit of the CRPS as the ensemble goes to one member. Or perhaps worth reporting on both? The same applies to Table S3.

Response. The main point of looking at the ensemble-mean MAE is to compare the best deterministic forecast made by CorrDiff—its ensemble mean—with the deterministic baselines. From this we conclude that the mean of CorrDiff is a bit better than the Unet. Since we have no ensemble baselines, all of them estimate the ensemble mean rather than sample individual ensemble members, the MAE of an individual ensemble member would not be strictly comparable to any of the other rows in the table, and is not a proper scoring rule. In effect, every row in the table is CRPS. Since CRPS is given by “average MAE – spread” and the spread vanishes for a single-member ensemble. That said, the mean MAE would be the correct metric for comparing CorrDiff to a single ensemble member sampled by a method that samples from the distribution rather than estimating its mean, such as a single high resolution WRF downscaling simulation. So it would enable future comparisons to report it.

Fig. 3: Just checking that these definitions are consistent with Leutbecher and Palmer, Ensemble Forecasting, 2008. - “the RSME of mean prediction for the...” just checking you are referring to ensemble mean here? (also note typo RSME -> RMSE). - “The standard deviation is adjusted with a factor $\sqrt{1 + 1/n}$ ” : My understanding is that you are using the unbiased estimator of the standard deviation, hence the difference compared to the correction factor of $\sqrt{(N+1)/(N-1)}$ in Leutbecher and Palmer. Just checking this is the case.

Response. Thank you for these observations. The typo has been corrected to "RMSE." Yes, "the RSME of mean prediction" refers to the RMSE of the ensemble mean and we are using equation (15) in Fortin et al., (2014) applying the factor $\sqrt{1 + 1/n} = \sqrt{(n + 1)/n}$ as noted in the text. We added the exact reference in the caption of the figure: “(see equation 15 in Fortin et al., 2014)”.

In their response to my point on line 45 (previous manuscript), the authors say they have updated the text but looks like this part of the updated text hasn't come through: ... We use 2018 through 2020 for training. For testing we used 205 random samples from 2021, and the rest used for validation. ...”

Response. Thank you for catching this oversight. The text has been updated in the revised manuscript to reflect the training, testing, and validation data details as mentioned.

Section 2: “Our input y ... is a low-resolution meteorological data with a forward and a backward processes working in tandem.”

Response. Revised. Thank you.

2 Reviewer 2

In this revised version of the manuscript, the authors present their combined U-Net and Corrector Diffusion-based downscaling approach with a correction of a bug in their code, which originally led to an overestimation of the model's performance. I appreciate the authors' scientific integrity and effort in admitting and correcting this bug. Even with a slightly reduced skill, I believe the model is a valuable advancement upon the state of the art. I also appreciate the intention of the authors to make the architecture openly available.

Response. We greatly appreciate the reviewer's understanding and positive feedback regarding the bug correction. We believe this update enhances the robustness of our findings and maintains the model's relevance

as a valuable contribution to the field. Additionally, we are committed to openly sharing the architecture, as mentioned, to support further research and development.

The authors have implemented nearly all the requested changes and added a valuable new supplementary section containing a statistical analysis which discusses the significance of their improvements.

Response. We thank the reviewer for acknowledging the additional statistical analysis and the implemented changes. We believe these additions strengthen the manuscript and provide a clearer understanding of the significance of our improvements.

The authors answered my question regarding the lead time of the WRF-CWA forecasts: the reference corresponds to the analysis, not to a forecast. However, this is still not clear from the manuscript itself. Moreover, this makes me wonder how the 1-hour maximum radar reflectivity was computed. This is (based on) a prognostic variable, meaning that only the analysis is insufficient, and you need to run the WRF model forward in time to compute it. Could the authors please clarify this?

Response. There seem to be a miss-communication with our collaborators in CWA over the exact nature of the variable, and we highly appreciate the reviewers questions that help us uncover it. The radar reflectivity used in this study is the maximum value of the vertical profile of the radar reflectivity in each column and at each time (i.e. instantaneous) and it is diagnosed from the model's prognostic variables during the data assimilation stage (3Dvar). We revised the manuscript accordingly to:

“We hypothesize that the significant distribution shift between the input variables and challenging target variables, particularly the vertical maximum of the derived radar reflectivity (hereafter referred to as radar reflectivity), necessitates high noise levels.”

and later:

“We sub-selected 4 channels (variables) as the target variables that are most relevant for practical forecasting: temperature at 2 meter above the surface, the horizontal winds at 10 meter above the surface and the maximum value of the derived radar reflectivity at each column (radar reflectivity hereafter) - a surrogate of the expected precipitation.”

*Please correct: "radar reflectivity - a much more complex observable that relates to the *sixth moment of the cloud water droplet distribution*" -> *sixth moment of the raindrop size distribution.*

Response. Revised. Thank you.

There are some missing capitals in the References section (e.g., taiwan, era5, ...)

Response. This has been corrected in the revised manuscript.

The manuscript and SI still contain some typographical/editing errors (e.g., "metrics reported in 1" -> "metrics reported in Section 1")—please review carefully.

Response. Thank you for noting this. We have thoroughly reviewed and corrected these issues.