Measuring Sharpness of AI-Generated Meteorological Imagery

Imme Ebert-Uphoff[a,b] , Lander Ver Hoef[a] , John S. Schreck[c] , Jason Stock[d] , Maria J. Molina[e,c] , Amy McGovern[f] , Michael Yu[f] , Bill Petzke[c] , Kyle Hilburn[a] , David M. Hall[g] , David John Gagne II[c] , William F. Campbell[h] , Jacob T. Radford[a,i] , Jebb Q. Stewart[i] , Sam Scheuerman[j]

[a] *Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO, USA.*

[b] *Electrical and Computer Engineering, Colorado State University, Fort Collins, CO, USA.*

[c] *NSF National Center for Atmospheric Research, Boulder, CO, USA.*

[d] *Computer Science, Colorado State University, Fort Collins, CO, USA.*

[e] *Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD, USA.*

[f] *School of Computer Science and School of Meteorology, University of Oklahoma, Norman, OK, USA.*

[g] *NVIDIA, Santa Clara, CA*

[h] *U.S. Naval Research Laboratory, Marine Meteorology Division, Monterey, California*

[i] *Global Systems Laboratory, Oceanic and Atmospheric Research, National Oceanic and Atmospheric Administration, Boulder, Colorado, USA*

[j] *Mathematics, Colorado State University, Fort Collins, CO, USA*

Corresponding author: Imme Ebert-Uphoff, iebert@colostate.edu

Early Online Release: This preliminary version has been accepted for publication in *Artificial Intelligence for the Earth Systems*, may be fully cited, and has been assigned DOI 10.1175/AIES-D-24-0083.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT: AI-based algorithms are emerging in many meteorological applications that produce imagery as output, including for global weather forecasting models. However, the imagery produced by AI algorithms, especially by convolutional neural networks (CNNs), is often described as too blurry to look realistic, partly because CNNs tend to represent uncertainty as blurriness. This blurriness can be undesirable since it might obscure important meteorological features. More complex AI models, such as Generative AI models, produce images that appear to be sharper. However, improved sharpness may come at the expense of a decline in other performance criteria, such as standard forecast verification metrics. To navigate any trade-off between sharpness and other performance metrics it is important to quantitatively assess those other metrics along with sharpness. While there is a rich set of forecast verification metrics available for meteorological images, none of them focus on sharpness. This paper seeks to fill this gap by 1) exploring a variety of sharpness metrics from other fields, 2) evaluating properties of these metrics, 3) proposing the new concept of Gaussian Blur Equivalence as a tool for their uniform interpretation, and 4) demonstrating their use for sample meteorological applications, including a CNN that emulates radar imagery from satellite imagery (GREMLIN) and an AI-based global weather forecasting model (GraphCast).

SIGNIFICANCE STATEMENT: AI-based estimates of meteorological images, e.g., for forecasting applications, often lack sharpness, but there are no well established metrics to measure sharpness of meteorological imagery. This manuscript seeks to close this gap by exploring sharpness metrics for meteorological imagery, analyzing their properties, and providing guidelines for their interpretation. We hope that the tools provided here will aid the development of AI algorithms that provide more realistic meteorological imagery.

1. Introduction

Neural networks (NNs) are increasingly used to generate meteorological imagery for numerous meteorological applications, ranging from the generation of synthetic radar imagery (Hilburn et al. 2020) to global weather forecasting tasks (Bonev et al. 2023; Bi et al. 2023; Lam et al. 2023). A concern with many of these models, especially many convolutional neural networks (CNNs), is that they produce imagery that is considered too blurry to be realistic (Blau and Michaeli 2018). Newer AI models, in particular Generative AI models - which are discussed in Section 1c - can yield imagery that contains much more detail and thus appears to be much ”sharper”. The emergence of Generative AI provides vast new opportunities to customize AI models to satisfy the specific requirements of an application, e.g., the need to provide detailed meteorological features. However, navigating this extended AI model space also presents new challenges, as sharper images are not always better images. For example, optimizing sharpness by itself may result in a decrease of other performance criteria, such as those measured by traditional forecast verification metrics. Vice versa, optimizing traditional forecast verification metrics, such as root mean square error, tends to decrease sharpness, an effect often described as reduced *effective resolution* of many AI-based weather prediction models (Subich et al. 2025; Selz et al. 2025). To be able to effectively navigate such trade-offs, it is essential to assess the images produced by AI models using both traditional forecast verification metrics, as well as measures for sharpness.

a. Relationship to existing forecast verification metrics

There is a rich body of literature on forecast verification metrics, for an excellent overview of such metrics see for example Gilleland et al. (2009, 2010), Jolliffe and Stephenson (2012), or Dorninger et al. (2018). For use of such metrics in practice, see Turner et al. (2020), which describes a framework for NWP model verification, and Weather Bench 2 (Rasp et al. 2024), which describes a framework to compare the performance of AIWP and NWP models.

Many of these forecast verification metrics apply to spatial fields. Such metrics include pixel-based methods (such as root mean square error; RMSE), neighborhood methods (such as fractions skill score; FSS; Roberts and Lean (2008)), scale-separation methods (Briggs and Levine 1997; Buschow and Friederichs 2020), feature-based methods (Davis et al. 2006; Brown et al. 2007), and field deformation methods (Gilleland et al. 2009). However, none of those focus specifically on evaluating sharpness. For example, pixel-wise comparison of two images, such as calculating the average RMSE, only evaluates pixel-wise match-ups, and reveals nothing about the level of detail included in either image. Similarly, an estimate of the pixel-wise displacement vectors that map one field to another, as can be obtained using field deformation methods, on its own tells us nothing about sharpness of either image. Neighborhood-based methods, such as FSS, require that a continuously-valued image must first be discretized (often binarized), and during that step many details relevant to sharpness, such as *gradual* transition from a low to high values and the texture details in the images, are removed before the FSS can be applied. Similarly, feature-based verification metrics, such as the Method for Object-based Diagnostic Evaluation (MODE; Brown et al. (2007)), require to first map an image to a discrete set of objects, and most details are removed during that step. Finally, scale-separation methods assess similarity of images at specific scales in spectral space, typically applying either a Fourier or wavelet transformation to the images first (Gilleland et al. 2009; Briggs and Levine 1997; Buschow and Friederichs 2020). While those metrics do not directly measure sharpness either, they share the property of performing image comparisons in spectral space with the spectral sharpness metrics discussed in Section 2c.

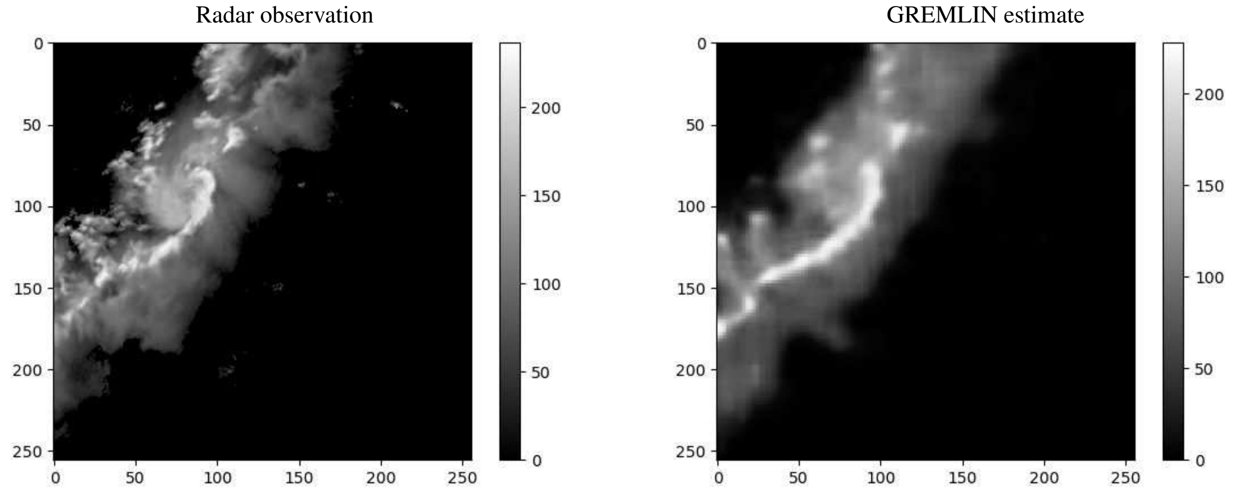
Probably the concept most closely related to sharpness is *effective resolution*. Effective resolution, in the context of weather prediction models, is defined as the smallest spatial scale where atmospheric structures are reproduced with realistic amplitudes (Selz et al. 2025). While well established in numerical weather prediction (Skamarock 2004), it has recently seen a resurgence for the comparison of AI-based weather prediction models to numerical weather prediction models (Subich et al. 2025; Selz et al. 2025). Effective resolution of a forecast image is typically evaluated in spectral space, e.g., by evaluating energy spectra (Selz et al. 2025), properties of spherical harmonic modes (Subich et al. 2025), or properties of a Haar wavelet decomposition (Pfreundschuh et al. 2022). However, effective resolution focuses on evaluating the spatial scale of represented features, but does not directly measure the sharpness of those features, such as rapid transitions.

The emergence of highly customizable AI models, especially Generative AI, creates the need to study a wide variety of sharpness metrics that can capture different aspects of sharpness. For this reason we expand our perspective by studying sharpness metrics from other fields for potential use in meteorological applications. It cannot be emphasized enough though that any such sharpness metrics should only be applied *in addition* to any traditional forecast verification metrics relevant to an application, since as stated earlier, we expect there to be trade-offs between satisfying sharpness criteria and satisfying other critical performance requirements.

b. A guiding example

As illustrative example of a neural network for image generation we use the GREMLIN model developed by Hilburn et al. (2020) throughout this manuscript. GREMLIN is short for ”GOES Radar Estimation via Machine Learning to Inform NWP” and is a convolutional neural network (CNN) model for image-to-image translation. It translates images from geostationary satellites to synthetic radar imagery, specifically composite reflectivity. Its purpose is to estimate radar composite reflectivity in regions where radar is not available, such as in mountainous and remote terrain and over oceans. GREMLIN is a standard CNN with a U-net (Ronneberger et al. 2015) architecture and a custom loss function that improves GREMLIN’s ability to predict high intensity events. Figure 1(a) shows a sample observed radar composite reflectivity image (ground truth) and Figure 1(b) shows the corresponding estimate from GREMLIN. Image values for both the observed composite reflectivity and the GREMLIN estimate are scaled – in Fig. 1 and throughout this manuscript – as follows. The radar reflectivity (in dBZ) is divided by 60 dBZ resulting in non-dimensional values within $[0,1]$. The result is multiplied by 255 to get the image pixel values.

It is apparent in Fig. 1 that GREMLIN’s composite reflectivity estimate is much blurrier than the observation (ground truth). We would like to modify GREMLIN to provide sharper features, but at this point we cannot even quantify what exactly that means - as there are no standard metrics to quantify sharpness of meteorological imagery.



(a)

(b)

Figure 1: GREMLIN model for generating synthetic radar imagery: (a) observed composite reflectivity from radar (ground truth); (b) estimated composite reflectivity by CNN model GREMLIN (Hilburn et al. 2020) based on satellite imagery. Scale: both observed and estimated radar reflectivity values are divided by 60 dBZ to obtain values within $[0,1]$, then multiplied by 255.

c. Predictive vs. generative AI models

It is important to distinguish predictive vs. generative AI models for image generation. Predictive AI models, such as GREMLIN and many other standard CNNs that use RMSE-like loss functions for image-to-image translation tasks, produce a single output image that, approximately, represents the average of all possible solutions (Subich et al. 2025; Selz et al. 2025). In contrast, Generative AI models, such as Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) and diffusion models (Sohl-Dickstein et al. 2015), produce one or more output images that each represent a single sample of all possible solutions.

Figure 2, which is derived from an image by Ledig et al. (2017), illustrates the difference between predictive and generative model output for the task of *super-resolution*, i.e., generating a high-resolution image from a given low-resolution image. Super-resolution is an ill-defined task, since infinitely many high-resolution images correspond to the same low-resolution image. Ledig et al. (2017) study the use of GANs for super-resolution. In Fig. 2 the patches with red frames represent an ensemble, i.e., a subset of the infinitely many possible high-resolution images that correspond to a given (not shown) low-resolution image. The patch in blue indicates the predictive model solution which is an approximation of the average of all red patches, i.e., of the ensemble mean. The patch

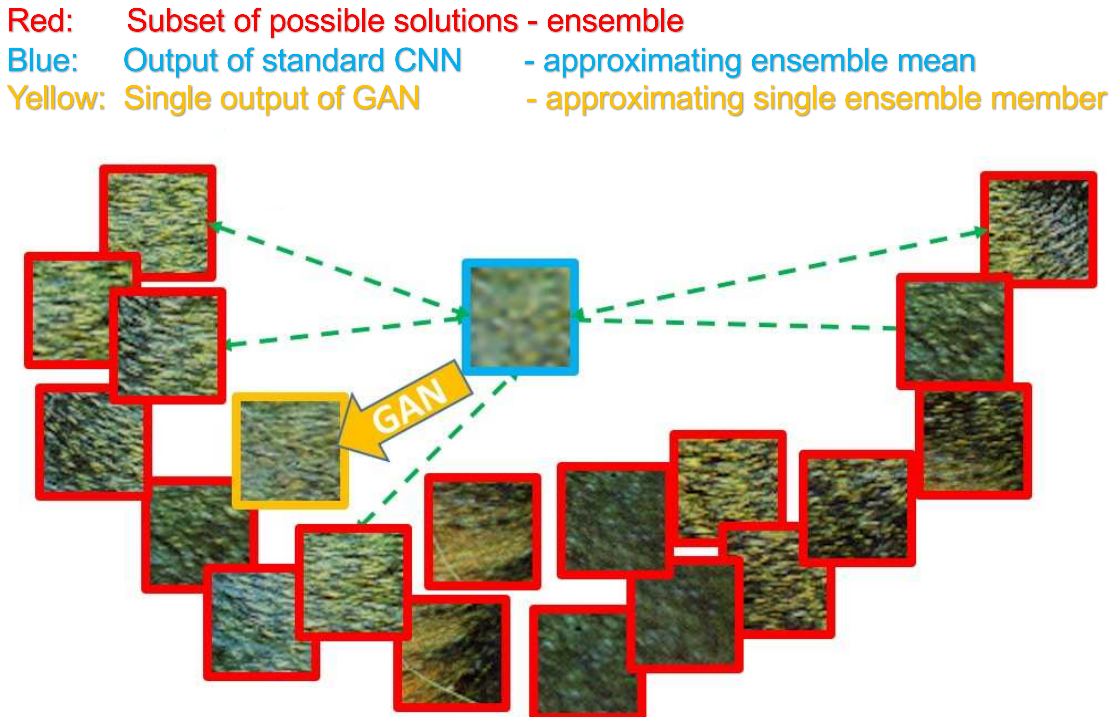

Red: Subset of possible solutions - ensemble
Blue: Output of standard CNN - approximating ensemble mean
Yellow: Single output of GAN - approximating single ensemble member
GAN

Figure 2: Schematic illustrating the typical results obtained by a standard CNN and a GAN for the ill-defined task of generating high-resolution imagery from low-resolution imagery. The patches in red represent a subset of the infinitely many solutions to the problem (ensemble members), the patch in blue represents the solution obtained by a CNN with Mean Squared Error (MSE) loss function (roughly ensemble average), and the patch in yellow represents a GAN solution (approximation of single ensemble member). Image credit: Figure adapted from Fig. 3 in Ledig et al. (2017) - reprinted with permission from IEEE.

in yellow indicates a generative model solution (here from a GAN) which approximates a single ensemble member. As a consequence, the image from the predictive model (in blue) represents the "safe" solution: it typically has the highest possible accuracy (e.g., lowest mean square error; MSE), but it is very blurry and does not itself represent a physically possible solution. In contrast, the image from the generative model (in yellow) is quite sharp and represents a physically possible solution, but at the cost of lower accuracy (e.g., higher MSE), and it might represent an outlier of the ensemble.

d. Effect of AI model uncertainty

Fig. 3 summarizes how model uncertainty is expressed in imagery generated by predictive vs. generative AI models. For a predictive AI model the uncertainty is expressed as increased blurri-

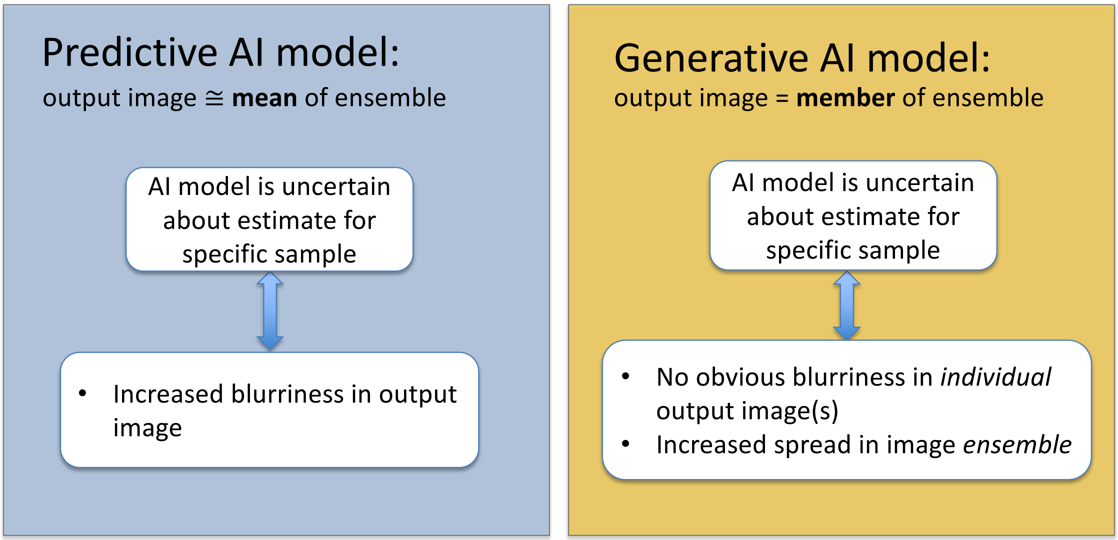

Predictive AI model:
output image ≅ **mean** of ensemble
AI model is uncertain about estimate for specific sample
• Increased blurriness in output image
Generative AI model:
output image = **member** of ensemble
AI model is uncertain about estimate for specific sample
• No obvious blurriness in *individual* output image(s)
• Increased spread in image *ensemble*

Figure 3: How uncertainty is expressed in imagery from predictive vs. generative AI models

ness in the output image, while for a generative AI model the uncertainty is expressed in the spread of the ensemble, while each individual image appears to be sharp.

In the meteorological community a generative model is best understood as serving the function of a probabilistic model, i.e., a model that inherently yields an ensemble forecast and whose output should thus be interpreted as one member of an ensemble. However, in the AI community generative models are not always interpreted that way. They are often used to generate a single output image, and the fact that this image might be an outlier, and thus not representative of the entire ensemble, may be hidden. That is because in many computer science applications, such as animations, an image only has to ”look” realistic. A single member of a generative AI model output fits this requirement of ”looking” realistic. However, that criterion is not good enough for a typical meteorological application, which requires a forecast to be representative of the set of possible solutions.

Thus one needs to be careful when using a single estimate of a generative model. It is recommended to export an ensemble - rather than just a single member - and to check the ensemble's spread. If the spread is small, then using a single ensemble member is fine. If only certain regions of the image have a significant spread, one may choose to visually indicate those regions in the resulting image, or to present several ensemble members. There is an increasing need for the community to explore the most effective ways to communicate the uncertainty information gained from ensembles to different end users, as Generative AI will greatly increase the availability of ensembles in meteorological applications. As a starting point, see the study by Demuth et al. (2020) on how to communicate ensemble information to forecasters of the National Weather Service.

Lastly, we note that the uncertainty discussed here is the total uncertainty of the AI model, i.e. it includes both internal variability of the weather system, aka the aleatory component, and all types of AI model errors, aka the epistemic component. See Haynes et al. (2023) for a detailed discussion of the concepts of aleatory and epistemic uncertainty for machine learning models. The total uncertainty is expressed as blurriness for predictive models and as spread for generative models, regardless of its source.

e. Is a sharper image a better image?

In our quest to make images sharper, we need to carefully consider the specific needs of each application. For example, when predicting precipitation, what is more important, higher spatial accuracy or higher sharpness? The answer depends completely on how the information is supposed to be used, as well as which information is already available. For example, to get a good estimate of the probability that it rains in a particular location, it is likely more important to optimize for spatial accuracy. On the other hand, to get an idea of the severity of the overall rain event, it might be more important to optimize the sharpness of the meteorological features even at the potential cost of more error in the location of the features.

As these examples illustrate, a blurrier but more accurate image may be more useful for analysis purposes, despite its lack of visual appeal. However, there may be instances where having sharp features resolved (even if those features have some degree of spatial or intensity error) is helpful. For instance, if the internal texture of a feature is important, it may not matter whether individual pixels are in the correct location so long as their local features and arrangement is accurate – in such a case, having a sharper but potentially less “accurate” model may be of great utility. A similar scenario where a sharper image is preferred is situations where an over-smoothed solution may blur out small, isolated features which may be crucial to know about, e.g., the presence of isolated storms, even if they are not in the correct position.

In addition, we need to keep track of the aforementioned trade-offs. For example, Blau and Michaeli (2018) find that for the application of image enhancement with neural networks there is a perception-distortion trade-off, i.e. a trade-off between perception-based qualities (such as sharpness) and distortion (i.e. accuracy).

Lastly, there is also the important question whether sharper images, for example a single image produced by a Generative AI model, may hide the model's uncertainty from the user. This could be misleading and instill unwarranted confidence for the user in the model's correctness. We suggest the following actions when using Generative AI to produce meteorological imagery:

- Educate users about the fact that individual forecasts obtained using Generative AI may look extremely detailed and realistic, even if the model is not terribly confident in its own prediction, i.e. the model uncertainty is hidden.
- Consider calculating ensembles and communicating uncertainty by other means if presenting only a single output image to the user.
- We urgently need studies on the effect that sharp forecast images, obtained using Generative AI, have on the user's perception of the model's confidence, especially for users who may base critical decisions on such imagery, such as forecasters.

For all of these reasons, we encourage readers to critically examine the actual benefits sought from sharper images. In particular, care must be taken when using Generative AI to generate sharper images, as outlined above. On the other hand, there is no doubt that Generative AI - in particular diffusion models - provide a powerful technique that, if used carefully, has the potential to greatly advance the field of meteorological forecasting by providing more detailed forecasts, ensemble forecasts, and estimating uncertainty. The potential power of employing diffusion models in meteorological forecasting has been demonstrated recently, for example by the SEEDS (Li et al. 2024) and GenCast (Price et al. 2025) models.

f. Means to increase sharpness and proposed use of sharpness metrics

We need to keep in mind the reasons that cause blurriness for a considered AI model, and thus, whether it is possible, or meaningful, to reduce that blurriness. For example, for a predictive model with an RMSE-like loss function we know that blurriness represents uncertainty, so developers may first look for ways to reduce the epistemic uncertainty by changing the model architecture, loss function (e.g., Subich et al. (2025)), or hyperparameters, then re-evaluating sharpness and other performance criteria. One may also seek to expand the training set, by adding more samples to better represent the sample space, or adding more input variables. In particular, we highlight an often-overlooked source of sharpness: sharpness of the input data. While most of the focus for increasing sharpness has been on loss functions and architectures, the sharpness of the output is directly tied to the sharpness of the input data, and in many cases finding and utilizing sharper sources of data can be an efficient way to add sharpness. Post-processing tools, e.g., applying sharpening filters to the output images may be another way to increase sharpness of predictive model output. Lastly, moving from predictive to generative AI models tends to increase sharpness, with the caveats discussed in Section e.

We hope that the development of sharpness metrics will allow the community to answer important questions, such as:

- How sharp are images generated by generative models compared to predictive models?
- How sharp are images from either type of model in comparison to ground truth?
- Which model changes, e.g., in training data, model architecture, loss function, hyperparameters, or post-processing, improve the sharpness of the model's output?
- What are the common trade-offs between improving sharpness vs. potentially decreasing other important forecast verification metrics?

We foresee two primary ways to use sharpness metrics:

- **Diagnostic use:** Sharpness metrics can be used as a diagnostic tool during model development, i.e., to point developers toward specific shortcomings of a model. The information can be used by the developer, for example, to change the training dataset or model architecture in an attempt to fix the shortcoming. We see much potential for this type of use during the tuning, model-selection, and evaluation phase of AI model development.
- **Use in loss function:** Sharpness metrics can also be added to the loss function of a neural network to provide sharpness-related feedback during training. Many of the sharpness metrics are relatively easy to implement to be included in a custom loss function. For example, calculating image gradients is a standard functionality in neural network programming environments, facilitating the use of the image-gradient based sharpness metrics discussed in Section 2b.

Similarly, taking the spatial Fourier transform of an image is also surprisingly easy in a neural network programming environment (Lagerquist and Ebert-Uphoff 2022), thus enabling the use of metrics such as Fourier-RMSE discussed in Section 2c. What is challenging, however, is to choose a meaningful trade-off in a loss function between optimizing traditional and sharpness-based metrics. We include a few experiments on the use of sharpness-based metrics for neural network training in a vignette in Subsection S1c of the supplemental document. A deeper study of how to choose such trade-offs is suggested as a topic for future research.

g. Image assumptions, scope and organization of this paper

The scope of this paper is to explore several metrics to evaluate the sharpness of any type of meteorological image – whether it comes from observations, from a physics-based model, or from an AI-generated model – along with guidelines on how to interpret them.

For simplicity, we make the following assumptions for images to be evaluated:

1. We assume images to be two-dimensional. Many of the concepts discussed here also apply to higher dimensional images, but for ease of explanation, we restrict our discussion to 2D images with the two dimensions denoted as x and y.

2. We only consider single-channel (i.e., gray) images. For multi-channel (or multi-color) images, a metric can be applied to each channel separately, followed by taking the min, max, or average value across all channels. We suggest a deeper exploration of multi-channel images as a topic for future research.

3. We assume images to have no missing or undefined values (no NaNs).

The remainder of this paper is organized as follows:

- Section 2 provides an overview of metrics included in this study. All of these metrics have been used before in computer vision. We also include a description of *heatmaps*, one of the main visualization tools we use.

- Section 3 introduces the new concept of the Gaussian Blur Equivalent (GBE) and presents two case studies that illustrate its practical use in comparing sharpness: i) GREMLIN estimates vs. observed radar, and ii) weather forecasts from an AI-based model (GraphCast) vs. a traditional Numerical Weather Prediction (NWP) model (GFS).

- Section 4 discusses properties of sharpness metrics that are important for their use, ranging from their computational complexity to the impact of white noise on the sharpness values.

- Section 5 provides a final discussion and suggests topics for future work.

- Section S1 of the supplementary document contains vignettes that illustrate the evaluation of sharpness using the original metrics from Section 2 - rather than the GBE values introduced in Section 3 - for several meteorological applications.

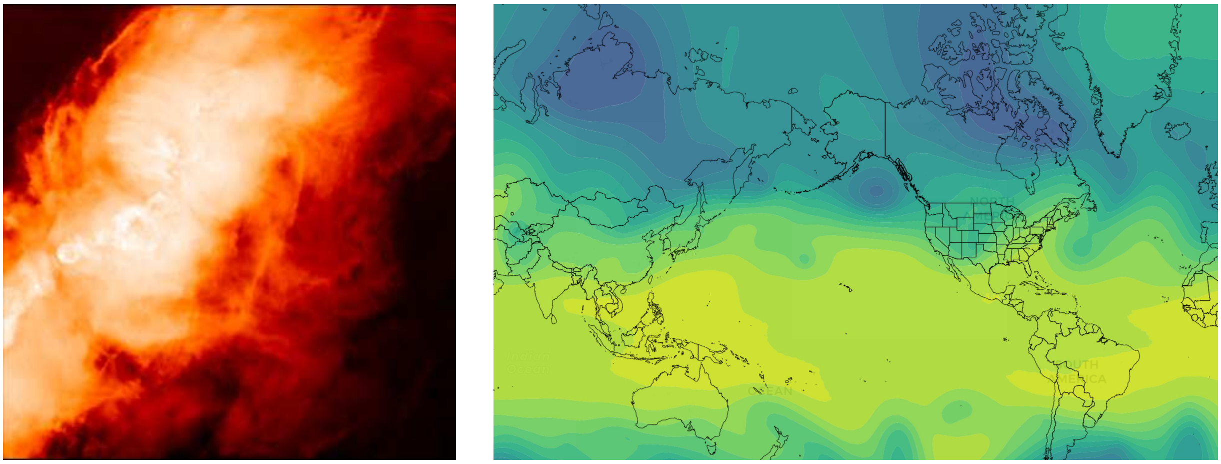
- Section S2 of the supplementary document provides the definitions of all metrics included in this manuscript, along with mathematical proofs for the metrics' properties that are presented in Section 4.

Accompanying code is available on Github, see the Data availability statement for details.

2. Sharpness Metrics

The term *image sharpness* is used extensively in literature, but it is difficult to find a consistent definition. Early definitions of sharpness can be found in photography, where sharpness is often defined *as the acuity, or contrast, between the edges of an object in an image* (SLR Lounge 2023). Note that this definition, and many others in photography, assume the presence of clearly defined edges in the image. However, meteorological imagery, such as the two examples shown in Fig. 4, may not include any such edges. Furthermore, considering the cloud in Fig. 4(a), it is clear that the perception of sharpness in this satellite image mostly comes from the level of detail provided inside the cloud, i.e., the cloud's texture, rather than by the sharpness of the cloud's boundaries. One may even debate the exact location of the cloud's boundaries. The field of photography also offers metrics analyzing image frequencies through the use of the Fourier transform. However, the field of computer vision has developed a wider and more suitable set of sharpness metrics for our purpose, which are discussed next.

Vu et al. (2011) provide an excellent overview of sharpness metrics from computer vision and classify them into three categories: edge-based, pixel-based and transform-based metrics. The list below discusses our selection of metrics, which is based on that classification by Vu et al. (2011). We emphasize that for this first study we selected a set of metrics that i) have simple mathematical



(a)

(b)

Figure 4: Two examples of meteorological imagery: (a) satellite image from the Geostationary Operational Environmental Satellite #16 (GOES-16) showing clouds; (b) forecast of 500mb geootential height at 5 days lead time from GraphCast (Lam et al. 2023) initialized on 4-23-2024 (00Z).Image credit for (b): GraphCast visualization from our Real-time visualization website for purely AI-based weather models (Radford et al. 2025). The richness of texture in the cloud image and the lack of clear boundaries in the geopotential height forecast illustrate why using *only* sharpness measures that focus on edges, i.e., boundaries between ”objects”, would disregard many important details in meteorological imagery.

equations and are easy to understand, ii) appear useful for meteorological applications, and iii) cover a wide range of different concepts.

1. *Edge-based metrics* first identify edges, then analyze their properties. We do *not* cover these metrics due to their underlying assumption that images must have well-defined edges. Furthermore, if there are sharp edges present in an image, their effect on sharpness will be detected by gradient-based metrics anyway, which are included in the next category.

2. *Pixel-based metrics*, aka *spatial metrics*, include gradient-based methods. We include several gradient-based metrics here. Pixel-based metrics also include methods based on eigenvalues/singular value decomposition (SVD) of images (Wee and Paramesran 2008), which are much more abstract and less commonly used. Those are not included here and may be added in future studies.

3. *Transform-based metrics*, aka *spectral metrics*, include methods based on Fourier or wavelet transforms. We include metrics based on both transforms.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-24-0083.1.

4. *Neural network based metrics* were not yet discussed by Vu et al. (2011), because they did not yet exist. These metrics utilize the values of certain internal states (latent space representation) of trained neural networks to assess image properties, e.g., see Zhang et al. (2018). Those metrics are *not* covered here, as their functionality is too opaque (i.e., black box character) for this first study.

We distinguish between *univariate* metrics, which take a single image as input at a time, and *bivariate* metrics, which require two input images at a time. Accuracy metrics, e.g., RMSE, are always bivariate, as one always needs a ground truth for comparison to assess the accuracy of an image. Bivariate sharpness metrics also compare one image to another, but instead of comparing the similarity of the image itself, they compare the similarity of the image's sharpness. In contrast, univariate sharpness metrics are applied to a single input image and assess the sharpness of just that image. To compare the sharpness of two images, one calculates the univariate metric for each image and then analyzes their difference. There is a key difference between univariate and bivariate sharpness metrics when used to compare two images. Bivariate sharpness metrics compare sharpness locally, i.e. they compare whether the images have the same sharpness at individual locations. In contrast, univariate sharpness metrics compare sharpness between images without taking the location of sharp features within an image into account. For example, when comparing two square images, one can rotate one image by 90 degrees without changing the results.

For simplicity we refer to each metric as being computed across an "image," but they can each also be computed on smaller subsets of an image, as we will see in Section 2d, to generate sharpness heatmaps.

We discuss the metrics in three groups in the following subsections: 1) standard forecast verification metrics, 2) sharpness metrics based on total variation and image gradients, and 3) sharpness metrics in spectral space. For each metric we provide a short description, the abbreviated name of its implementation in the GitHub repository (in parentheses), and whether the metric is univariate or bivariate. We emphasize again that none of these metrics are new. Group 1 metrics are already in use for meteorological imagery. The metrics in Groups 2 and 3 are not very common for meteorological imagery, but have all been used in other fields. **Definitions of all metrics are provided in Section S2a of the Supplemental material.**

a. Group 1: Standard forecast verification metrics

While the purpose of this study is to evaluate the sharpness of imagery, it is important to consider sharpness and standard verification metrics in tandem, namely to make sure that increasing sharpness does not come at the expense of drastically reducing other important metrics. We chose the three simple metrics below to represent a few common verification metrics. *These metrics are only a sample selection for Group 1 and developers should replace them with any forecast verification metrics deemed relevant for their considered application.*

1. **Image Intensity / Dynamic Range [univariate]:** We keep track of the min, mean, and max intensity value of each image, because the dynamic range of an image has a significant effect on its apparent sharpness. An easy way to increase many sharpness metrics of an image would be to just increase its dynamic range - which is typically not what we want. This motivates us to keep track of the intensity of images.

2. **Root Mean Squared Error (RMSE) [bivariate]:** RMSE is the square root of the mean squared error (MSE) between two images and is a commonly-used similarity metric for the training of neural networks. We keep track of RMSE to make sure we do not drastically reduce the similarity of image estimates while trying to make them sharper.

3. **Structural Similarity Index Measure (SSIM) [bivariate]:** SSIM is a similarity measure between two images based on a weighted combination of three simpler comparisons: luminance (intensity), contrast, and structure. The product of these measures gives SSIM. An important note is that SSIM acts on a patchwise rather than pixelwise basis, and as such can capture more spatial information than pixelwise methods like RMSE. SSIM values range between 0 and 1, with SSIM = 1 indicating identical images and values approaching 0 indicating increasingly dissimilar images. SSIM is often cited to better represent image similarity - as perceived by humans - than, for example, RMSE. For details, see Wang et al. (2004).

b. Group 2: Sharpness metrics based on total variation and image gradients

Since sharp boundaries result in sharp gradients, it is intuitive to use properties related to the gradient of an image to assess its sharpness. Total variation is very similar to gradient-based methods and is thus included here. We expect this group of metrics to respond strongly to sharp edges in an image.

1. **Total Variation (TV) [univariate]:** Total variation measures how much an image changes if it is shifted slightly. This can measure the sharpness of edges because when a sharp edge is shifted slightly it will cause a larger difference than if a smoother edge is shifted the same amount. TV values close to 0 indicate very smooth images, while sharper images will have larger TV values. It is important to note that we follow common convention in not normalizing TV by image size, so TV values for images (or blocks) of different sizes are not comparable, and it is normal to get TV values that are very large compared to most other metrics described here.

2. **Mean Gradient Magnitude (Grad-Mag) [univariate]:** At each pixel, we can compute gradients in both the horizontal (x) and vertical (y) directions; the *magnitude* of the gradient at that pixel is then the norm of the vector formed by those directional gradients. The Grad-Mag is the mean of these gradient magnitudes across the image, and as such gives a summary statistic that reports, on average, how rapidly intensity changes occur within the image. More rapid intensity changes generally correspond with sharper images, so higher Grad-Mag values indicate a sharper image, with Grad-Mag = 0 indicating a completely uniform image with no variation.

3. **Gradient Total Variation (Grad-TV) [univariate]:** Gradient total variation is the total variation of the gradient magnitude map, where the gradient map is described in Grad-Mag above, and total variation is as described in TV. Because both TV and gradients measure sharpness, the gradient TV is really giving information about how sharp the sharpness map is - i.e., are areas of rapid change (associated with sharpness) themselves sharp. In practice, this second-order sharpness seems to correspond with sharpness.

4. **Gradient RMSE (Grad-RMSE) [bivariate]:** In this bivariate metric, we compute the RMSE not between two images directly, but between two gradient magnitude images. We compute the gradient magnitudes as in Grad-Mag above, but rather than averaging those across a single image to obtain a statistic, we compute the RMSE between the gradient maps for two distinct images. As in general for RMSE, values closer to 0 indicate more similarity, while larger

values indicate more dissimilarity. By taking the RMSE of gradient magnitude maps, we are measuring how closely aligned regions of rapid change are between the two images; i.e., measuring how well sharp edges correspond between the two images.

5. **Laplace RMSE (Laplace-RMSE) [bivariate]:** Laplace RMSE is very similar to gradient RMSE, but instead of taking the magnitude of the gradient vector at each pixel, we compute the divergence of the gradient at each pixel, which is a way of quantifying the local shape of the gradient vector field. By taking the RMSE of two such divergence maps, we are computing how similar the shapes of edges are between two images. As with any of these RMSE measures, values close to 0 indicate that the two images have very similar Laplacian maps, while larger values indicate larger differences.

c. Group 3: Sharpness metrics in spectral space

The last set of metrics seeks to analyze the sharpness of images in spectral space. The idea is to first apply a Fourier or wavelet transformation, and then to analyze image properties in the corresponding spectral representation of the image.

1. **Fourier RMSE (Fourier-RMSE) [bivariate]:** When taking the 2D Fourier transform, the resulting complex-valued phase space can be reduced down to the *power spectrum* by taking the absolute value of the complex values at each frequency, which gives another real-valued 2D array. Fourier RMSE is then the RMSE between the power spectra of the two images being compared. Note that in the power spectrum, spatial coordinates correspond to frequencies, which are all weighted evenly in this RMSE computation.

2. **Fourier Total Variation (Fourier-TV) [univariate]:** We once again start with the power spectrum, but instead of comparing two power spectra, we take the Total Variation of the power spectrum for a single image. The power spectrum contains information about sharpness (as high-frequency information can be interpreted as "sharp"), and TV measures how sharp the power spectrum is, so like Grad-TV, we have some degree of second-order sharpness.

3. **Spectral Slope (Spec-Slope) [univariate]:** As mentioned in Fourier-TV, the power spectrum of an image, in particular, the distribution of high vs low-frequency information, contains information on how sharp an image is and spectral slope seeks to capture this information.

The definition of spectral slope is based on the fact that the magnitude spectrum of almost all natural and model images decreases inversely with frequency and that in a logarithmic plot this decrease can be roughly approximated by a line (Vu et al. 2011). The slope of the line is called the spectral slope of the image. It is very sensitive to blurring, while also being entirely invariant to uniform changes in intensity, i.e., rescaling the image intensity does not change the spectral slope value. Outside of mostly carefully constructed (artificial) examples, values of spectral slope are all negative, with more negative values indicating less sharp images. Spec-Slope's invariance to uniform changes of an image's intensity is advantageous for some applications, but this invariance also creates undesired side effects. Because spectral slope is invariant to intensity, it tends to return very high sharpness results in regions of low intensity and contrast for even miniscule signals, such as noise.

4. **$\mathbf{S_1}$ (S_1) [univariate]:** S_1 is derived from Spec-Slope, designed to compensate for Spec-Slope's aforementioned problem with low intensity signals by adding a minimal contrast requirement. Namely, the S_1 metric computes the contrast for a considered image (or image patch). If the contrast is below a certain threshold, it returns a null sharpness value. Otherwise, it returns the value of Spec-Slope. Thus, S_1 returns non-zero sharpness values only for regions that have been deemed (by setting the contrast threshold) to have sufficient variation to justify consideration. We note that this contrast threshold is a hyperparameter that must be set with care depending on the data type, range, and analysis needs of each application. As a default parameter we use the value 25 (chosen as 10% of the max intensity value, 255, of most images in this manuscript) throughout this manuscript, with the exception of results provided for GraphCast in Section 3c.

5. **Wavelet Total Variation (Wavelet-TV) [univariate]:** Wavelet-TV is based on the wavelet transform, which takes in an image and (for one level) yields a set of four output arrays: the approximation coefficients and three sets of detail coefficients. The detail coefficients contain information about variation in the image at various scales and orientations, while the approximation coefficient contains information about average intensities, so by summing the absolute value of all of these coefficients, we arrive at a notion of total variation in the image utilizing wavelets. Like Total Variation, we view increasing values of Wavelet-TV as having

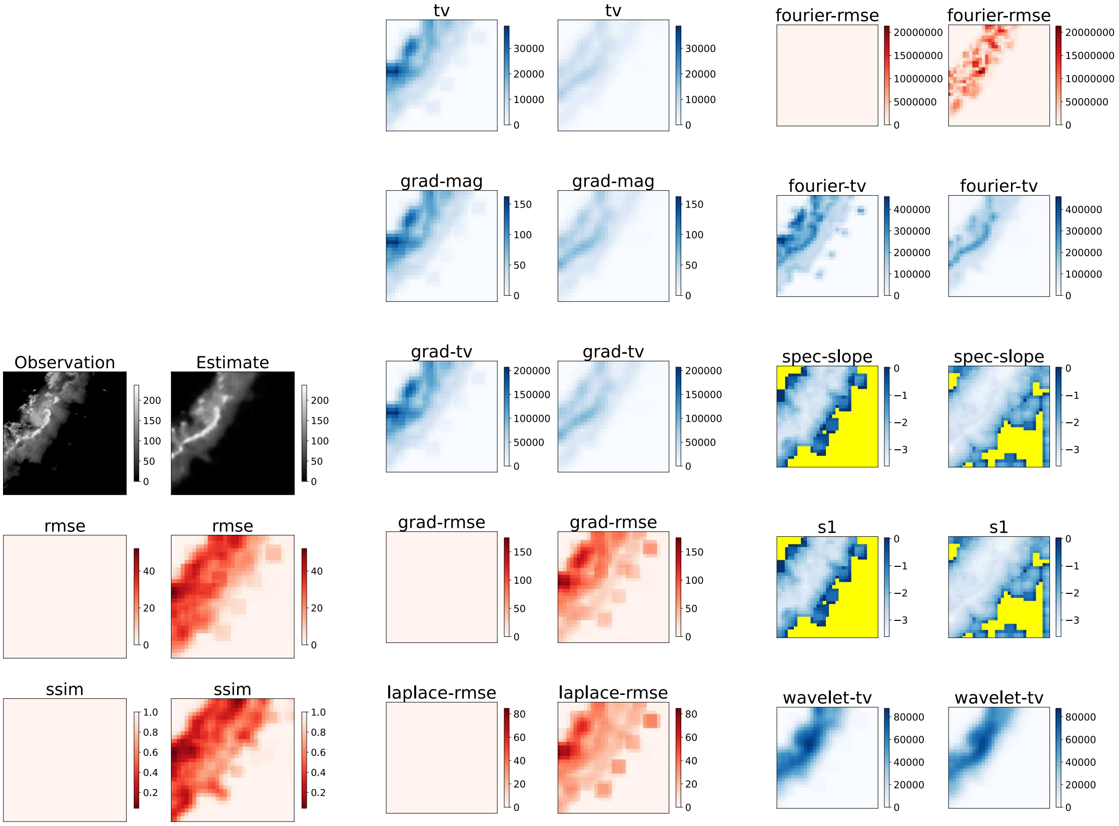
higher sharpness and note that Wavelet-TV is also not normalized by the size of the image, so Wavelet-TV values for different image (or block) sizes are not comparable.

d. Visualizing local sharpness using heatmaps

Since meteorology is a very visual field, we believe it is essential for all of the metrics to not only provide a single number for quantitative assessment, but also a visual representation of which features in an image are perceived to be particularly accurate or sharp. To provide such visual feedback we generate *heatmaps* by evaluating small patches of each image and displaying the resulting local information as an image, i.e., the heatmap of an image for a specific metric, as illustrated in Figure 5. Sharpness heatmaps have been used before (Vu et al. 2011), but to the best of our knowledge they have never been applied for meteorological imagery.

Each patch is a small square block with edges that are $1/8$th the length of the horizontal edge length of the input image. We calculate and visualize the metric values for all blocks - that represents the heatmap. Using disjoint blocks can lead to edges lying along the border between two blocks not being detected. Thus we use overlapping blocks. For most experiments, adjacent blocks overlap 75% of their area, but for blocks smaller than 8×8 pixels the overlap may be less than 75% because we enforce a minimum block stride of 2 pixels. The output heatmap reports the values for each block on the central pixels of that block but because of the overlap each block includes information from a larger region than its value is outputted to. For all metrics that utilize the Fourier transform, we implement windowing using the Hann window (following Vu et al. (2011) who used similar heatmaps) on each block to minimize the edge effects on the Fourier transform. Each heatmap can be shown on its own or used as an overlay over the input image(s) to indicate areas with very high or low values of each metric. We use the following colors to indicate the different types of heatmaps and the occurrence of NaNs:

- **Gray** indicates the original image, i,e., image intensity.
- **Blue** indicates values of univariate metrics, i.e., metrics that are calculated from an individual image (no comparison).
- **Red** indicates values of bivariate metrics, i.e., metrics that compare two images. Throughout this paper, all bivariate metrics indicate the comparison of each image to the original image,



(a) (b) (c)

Figure 5: Heatmaps comparing observed radar (reference) with the corresponding GREMLIN estimate (evaluation image) for (a) Metric Group 1, (b) Metric Group 2, and (c) Metric Group 3. The colormap for SSIM is inverted because SSIM indicates stronger similarity by a higher value - in contrast to standard similarity measures, such as RMSE. Yellow indicates pixels with invalid values.

which is always shown on the top left. Thus bivariate metrics for the original image with itself are identical to zero for all metrics except SSIM, and identical to one for SSIM.

- **Yellow** indicates individual pixels with invalid values (NaNs). We have observed NaNs only for the spectral slope metrics (Spec-Slope and S_1), since spectral slope is undefined in areas of an image that have no signal (constant value). The min/mean/max values of the heatmaps used in the stats plots are calculated across all valid pixels, i.e., pixels with NaNs are currently ignored.

As a convention we always plot the reference image in the left most column, i.e. on the left side of Fig. 5(a), (b), and (c). Bivariate heatmaps (those in red) compare an image to the reference image. This explains why the red heatmaps for the reference image are all constant, i.e. 0 for most bivariate metrics and 1 for SSIM, since SSIM has a value of 1 for identical images and 0 for maximal dissimilarity. Note also that the color scale for SSIM is reversed in these heatmaps, to indicate stronger differences by darker red.

Let us illustrate the interpretation of Fig. 5 for this GREMLIN example. Group 1 metrics indicate that the reference and evaluation image differ most strongly in image regions of strong intensity. Group 2 metrics tell us that the reference image is sharper than the evaluation image (blue heatmaps), and that the difference in sharpness is strongest in areas where the images are most different, as the red heat maps in Group 2 are in similar locations as in Group 1. Most Group 3 metrics, Fourier-RMSE, Fourier-TV and Wavelet-TV tell a similar story as Group 2 metrics, but indicate smaller, more focused regions of sharpness and sharpness difference. Spec-Slope does not appear useful here, as it indicates greatest sharpness in regions with extremely small intensity. The S_1 metric seeks to limit these regions by applying a minimal contrast treshhold, but with limited success: the S_1 metric still highlights areas of little interest since they might just be noise. Note that whatever the contrast threshhold, the S_1 metric tends to have the highest values right at the cutoff, because that is where image contrast has the lowest allowed value to pass the Spec-Slope value through, and thus where small amounts of white noise tend to have the largest effect on S_1.

e. Discussion

An important note about these metrics is that while there are some that are specifically for similarity or for sharpness, others measure some combination of the two. All of the metrics in Group 1 are about either intensity (for the raw image itself) or similarity. On the other hand, Groups 2 and 3 include metrics that either measure sharpness alone, or a combination of sharpness and similarity. The univariate metrics in these groups can be seen as directly measuring the sharpness of the images being input. The bivariate metrics measure not just sharpness, but the local alignment of sharp features (except for Fourier-RMSE, which measures the local alignment of phase space features). This captures both similarity and sharpness, because there are two ways for sharp features to come out of local alignment: one is that there is a lack of sharp features in one image to align, and another is for there to be sharp features that are out of position. Both perspectives are useful, particularly when taken in combination, as the univariate metrics can help disambiguate whether the bivariate metrics show differences due to a lack of sharp features or because of their misalignment.

Note the huge differences of the dynamic range of the various metrics. For example, for the sample images in Fig. 5 the dynamic range of univariate sharpness metrics varies between 4 units (Spec-Slope, S_1) and 400,000 units (Fourier-TV). Similarly, the dynamic range of the bivariate sharpness metrics varies between 80 units and 20 million units. Some of these differences could be reduced using normalizing factors related to image size, but even then large variations remain. We introduce a more meaningful approach in Section 3.

3. Introducing the Concept of Gaussian Blur Equivalent

As seen in Fig. 5 the range of values greatly varies across the metrics. This makes it hard to interpret the raw metric values. What constitutes strong sharpness for each metric? How do values of different metrics compare? Rather than determining how to interpret each metric individually, we propose instead a calibration method that generates a uniform scale for image comparison across all metrics, the Gaussian Blur Equivalent (GBE). We introduce the concepts of **GBE plot** and **GBE values** in Section 3a, then illustrate their use for a GREMLIN example in Section 3b and for two GraphCast examples in Section 3c. We conclude with a discussion in Section 3d.

a. GBE for image comparison

The GBE is designed for the comparison of two images. The *reference image* is typically the ground truth, e.g., the observed radar in the case of GREMLIN. The *evaluation image* is the image to be evaluated, e.g., the GREMLIN estimate. The core idea of GBE is to compare the evaluation image to a series of images, starting with the reference image and increasingly blurred versions of the reference image, as illustrated in Fig. 6. For each metric we compare which level of blurriness of the reference image corresponds most closely to the blurriness level of the evaluation image.

The steps of calculating GBE are described below, along with illustrations for the GREMLIN example in Fig. 7.

1. Panel (a): Create a sequence of increasingly blurred copies of the reference image by applying Gaussian blur with standard deviation, σ, increasing from 0 to some chosen value, $\sigma_{\max}$.

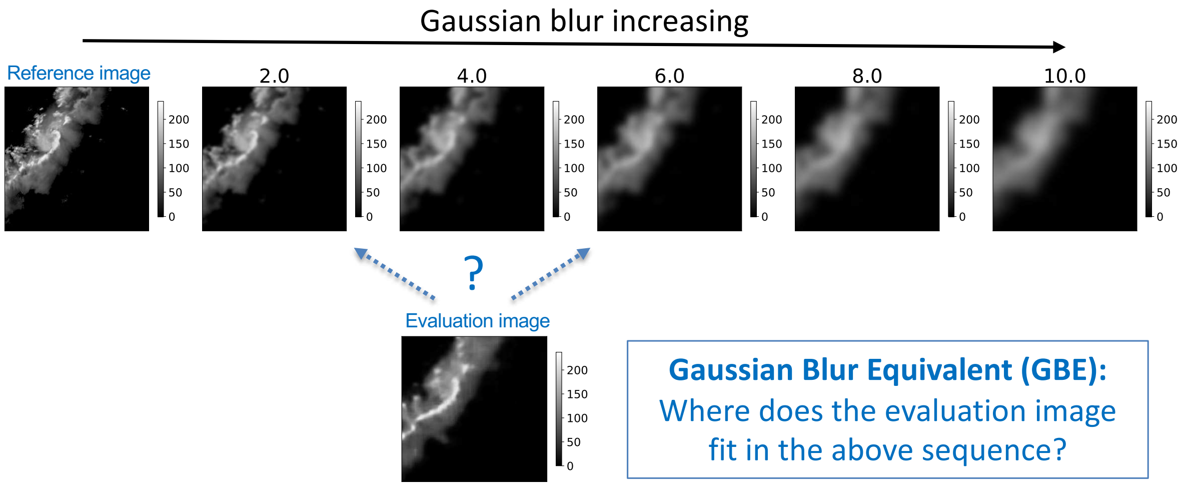

Gaussian blur increasing
Reference image
2.0
4.0
6.0
8.0
10.0
200
150
100
50
0
?
Evaluation image
Gaussian Blur Equivalent (GBE):
Where does the evaluation image
fit in the above sequence?

Figure 6: Core idea of GBE to calibrate a sharpness metric: How much do we have to blur the reference image to get the same (min, mean or max) sharpness value as for the evaluation image?

2. Panel (a): Calculate all metrics for all blurred versions of the reference image with respect to the reference image.

3. Panel (c): Generate a plot of the min, mean, and max metric values for the blurred reference images. Min, mean and max, are indicated in yellow, red, and blue, respectively, with each image resulting in a yellow, red, and blue, data point in the plot in the same order (left to right) as the heatmaps in Panel (a). Subsequent data points are connected by dashed lines. These are the calibration curves for the chosen metric for a specific reference image, aka raw metric value (RMV) plot.

4. Generate heat maps (Panel (b)) and calculate metric values (Panel (d)) for the reference and evaluation image.

5. Panel (e): Construct the **GBE plot** by combining the RMV plots in (c) and (d) as follows. Use the RMV plot (Panel (c)) and add horizontal lines for min, mean, max of the evaluation image (values are shown as the yellow/red/blue end points on the right of Panel (d)).

6. Panel (f): To calculate the GBE values, identify in the GBE plot, for each metric the smallest blur value at which the yellow/red/blue curve and same color horizontal line intersect. These values are called the **Gaussian Blur Equivalent (GBE) values**,

$$\sigma_{\text{GBE}}^{m,s}(\text{reference image, evaluation image}),$$

where m denotes the metric and s the statistic (min, mean, or max).

Interpretation: The Gaussian Blur Equivalent value, $\sigma^{m,s}_{\text{GBE}}$(reference image, evaluation image), is the standard deviation value of the Gaussian blur operation that, when applied to the reference image, would yield an identical metric value as the evaluation image. The GBE values provide the desired interpretation for comparison - using a uniform scale - across all metrics.

b. GBE comparison for GREMLIN vs. observed radar

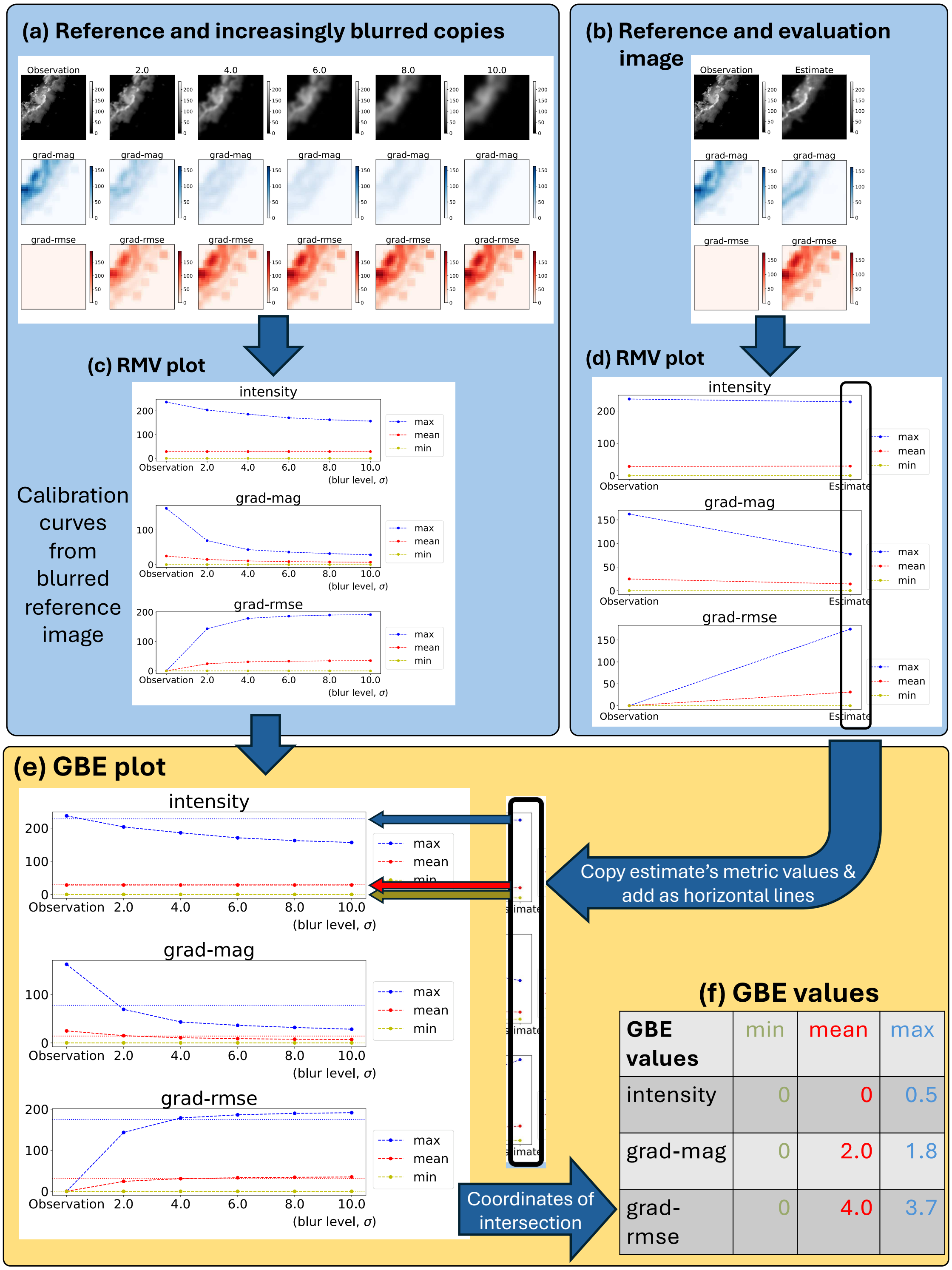
Fig. 8(a) shows the GBE plots for the GREMLIN example from Fig. 5 for all metrics. Fig. 8(b) shows the estimated GBE values, obtained by visual inspection from Fig. 8(a). In future work we plan to add automatic calculation of GBE values, but those algorithms need to properly deal with degenerate cases and alert users to those. Furthermore, visual inspection is required either way, for reasons discussed in Section 3d.

GREMLIN results (Fig. 8): Focusing on the univariate sharpness metrics, the gradient-based sharpness metrics place the image estimate at about the blurriness of applying a Gaussian filter with σ^{mean} around 2 to 2.5, while the spectral metrics consider the estimate to be much sharper, namely corresponding to σ^{mean} around 0.5 to 1.

We should emphasize that we have not established any guidelines for *ideal* GBE values. Such guidelines are bound to be highly application dependent. For example, for the GREMLIN application, it turns out that the reference images (radar observations) contain considerable noise, which, as discussed below in Section 4, can inflate the sharpness values of the reference image. In those cases, we would *not* want the GREMLIN estimates to reach the sharpness values of the reference images, since we do not want to reproduce the noise of the reference image. In contrast, if the ground truth is obtained from an NWP model (as in the following subsection) that is meant to model observations, then we would want the estimates to be even sharper than the ground truth (as NWP model output tends to be too smooth) which would yield a negative GBE value, but the current set-up does not provide for estimating negative GBE values.

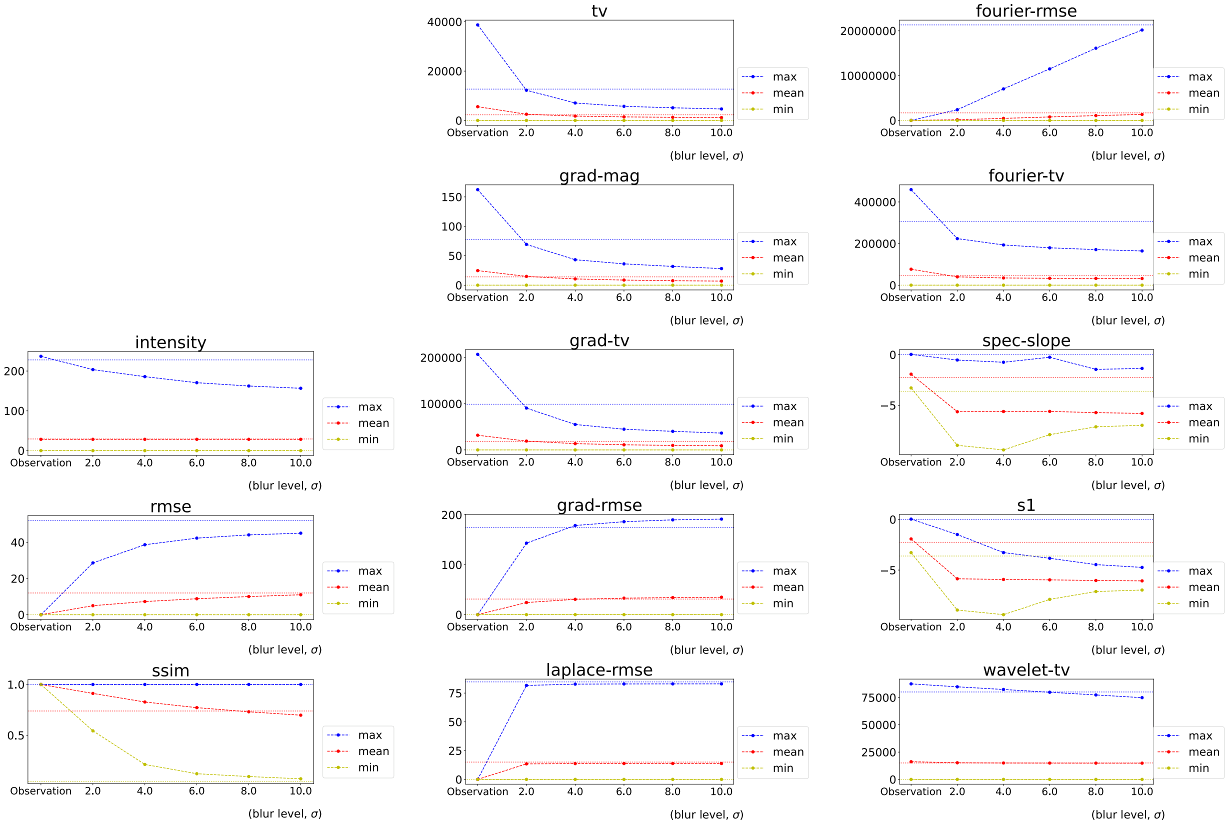
c. GBE comparison for GraphCast vs. GFS

One of the key motivations for this paper was to be able to compare the sharpness of the new generation of purely AI-based global weather prediction models (AIWP) to forecast imagery gen-


(a) Reference and increasingly blurred copies
Observation
2.0
4.0
6.0
8.0
10.0
grad-mag
grad-rmse
(b) Reference and evaluation image
Observation
Estimate
grad-mag
grad-rmse
(c) RMV plot
(d) RMV plot
intensity
grad-mag
grad-rmse
max
mean
min
(blur level, σ)
Calibration curves from blurred reference image
(e) GBE plot
Copy estimate's metric values & add as horizontal lines
Coordinates of intersection
(f) GBE values

GBE values	min	mean	max
intensity	0	0	0.5
grad-mag	0	2.0	1.8
grad-rmse	0	4.0	3.7

Figure 7: Calculation of GBE plot and GBE values, illustrated using GREMLIN example and three sample metrics, intensity, Grad-Mag and Grad-RMSE.



(a)

	Intensity	RMSE	SSIM	TV	Grad-Mag	Grad-TV	Grad-RMSE	Laplace-RMSE	Fourier-RMSE	Fourier-TV	Spec-Slope	S_1	Wavelet-TV
$\sigma^{\text{mean}}_{\text{equivalent}}$	≈ 0	> 10	≈ 8	≈ 2	≈ 2	≈ 2	≈ 4.0	> 10	> 10	≈ 2	≈ 0.2	≈ 0.2	≈ 2

(b)

Figure 8: (a) GBE comparison for radar observations (reference) vs. GREMLIN estimate (evaluation image) for the sample in Fig. 5 using metric groups 1, 2, and 3. Metric S_1 here uses the default contrast threshold, 25. (b) Mean GBE values in (a) estimated visually from red curves and lines.

erated by traditional NWP models. Here we compare an AIWP model, namely GraphCast (Lam et al. 2023), to an NWP model, namely the Global Forecast System (GFS) by the National Centers for Environmental Prediction (NCEP; NCEP (2024)). We do so for just two sample cases, where the GraphCast forecasts (evaluation images) are compared to the corresponding GFS forecasts (reference images), using the GBE approach. We use Release 1 of GraphCast (version 0.1 of Jan 5, 2024, SHA 8debd72, at `https://github.com/google-deepmind/graphcast`), initialize it with GFS initial conditions at t, run a forecast for a time ΔT, and compare the results at $(t+\Delta T)$ with a GFS forecast. Note that this GraphCast version was finetuned on the European NWP model, namely the Integrated Forecasting System (IFS; Wedi et al. (2015)) by the European Centre for Medium-Range Weather Forecasts (ECMWF), rather than GFS. Thus GraphCast results initialized by GFS (rather than IFS) are known to be not quite as good, although differences tend to be small. For example, for the forecast of 500-hPa geopotential heights about a half day degradation in skill has been observed for models initialized with GFS rather than IFS data (personal communication with ECMWF's Matthew Chantry). The *exact* results of this comparison should thus not be used to judge the performance of the GraphCast algorithm - using just two samples to judge an algorithm is not advisable anyway. Instead this section serves only as an illustration of how to apply and interpret these metrics for this type of application.

Figure 9 shows the forecasts generated by GFS and GraphCast for two events - a flashflood event forecast 48h out, and an atmospheric river event 148h out. The heatmap and GBE results are shown in Figures 10 and 11 for the flashflood event and in Figures 12 and 13 for the atmospheric river event.

GraphCast flashflood results (Figures 10, 11): The *mean* metric results are very flat for the blurred imagery which is likely due to the fact that most areas of the images are blank. Thus we focus on *max* metric results for all metrics except SSIM, and use *min* metric results for SSIM. We emphasize that the *mean* GBE value should be used wherever possible in these analyses, since it provides a better estimate of overall sharpness of features than just using the value of the sharpest features, i.e., the max GBE value. However, in this case the mean GBE value is not usable.

Note that for this flashflood event the maximal GraphCast intensity is roughly 1/3 of the maximal GFS intensity. As will be seen in Section 4 all metrics decrease linearly with a change in intensity, with the exception of SSIM and Spec-Slope, which are invariant, and S_1, which varies non-linearly. Thus, the low sharpness indicated in Fig. 11 for most metrics is largely due to the lower intensity, rather than lack of sharp transitions. Univariate gradient-based metrics place the GBE value around $\sigma^{\max} = 2.0$, and Fourier- and Wavelet-TV place it higher, around $\sigma^{\max} = 3$ to 3.5. The Spec-Slope heatmap indicates that that metric is not capturing relevant sharpness information here, so we discard it from consideration. On the other hand, by setting the threshold value for S_1 to 0.1,

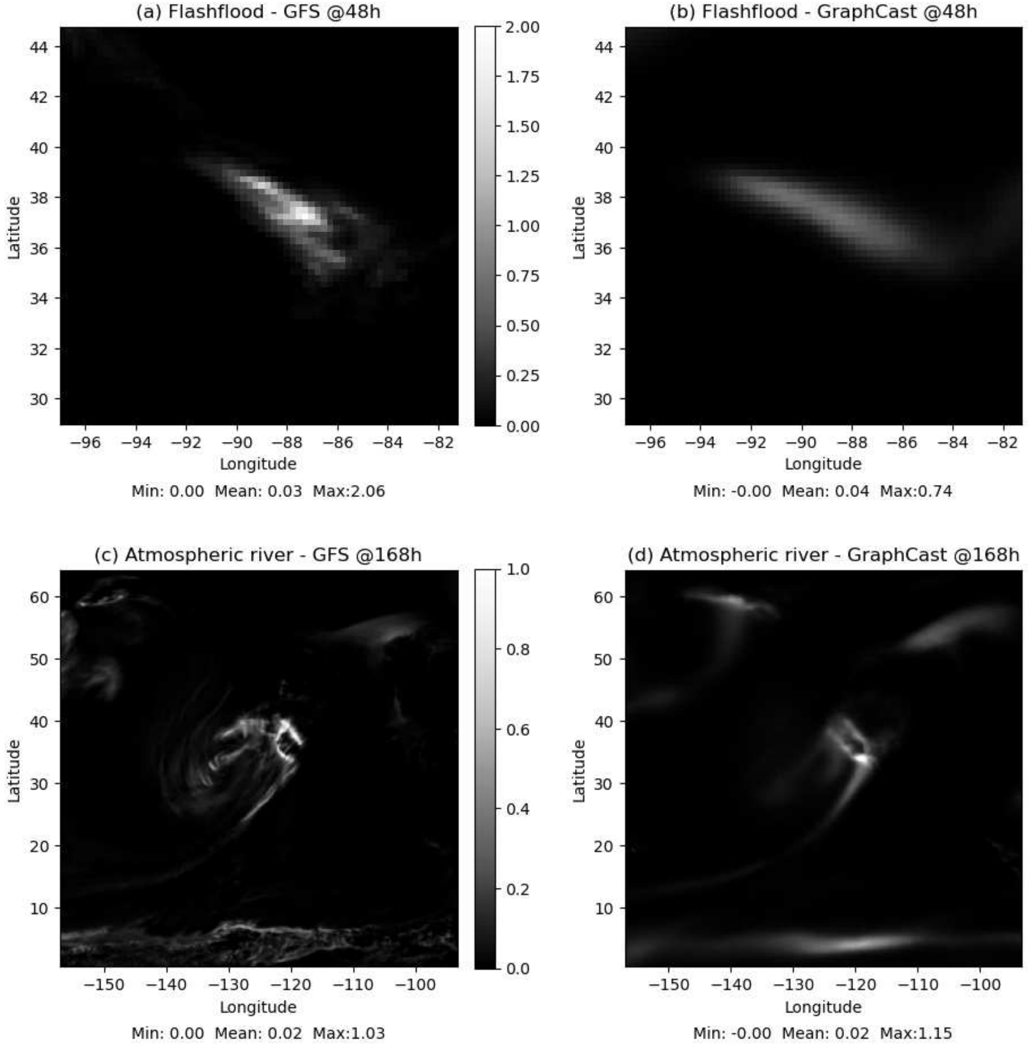


Figure 9: 6h accumulated precipitation (in inches) forecast by GFS and GraphCast for a flashflood event at lead time of 48 hours and for an atmospheric river event at 148 hour lead time. Image size: 64 x 64 for flashflood, 256x256 for atmospheric river.

we were able to extract sharpness information only for the relevant portions of the image, so we retain that as an informative measure.

GraphCast atmospheric river results (Figures 12, 13): For the atmospheric river event the maximal GraphCast intensity is similar, even a bit higher, than the maximal GFS intensity. The univariate gradient-based metrics indicate a $\sigma^{\max}$ value of about 1.8 to 2, and Fourier-TV and Wavelet-TV place it around 1.0 and 1.8, respectively. We hypothesize that this reduction in sharpness despite similar intensity values is due to the high level of texture and sharpness visible in Fig. 9(c) and the relative smoothness of the GraphCast forecast in Fig. 9(d). The Spec-Slope heatmap

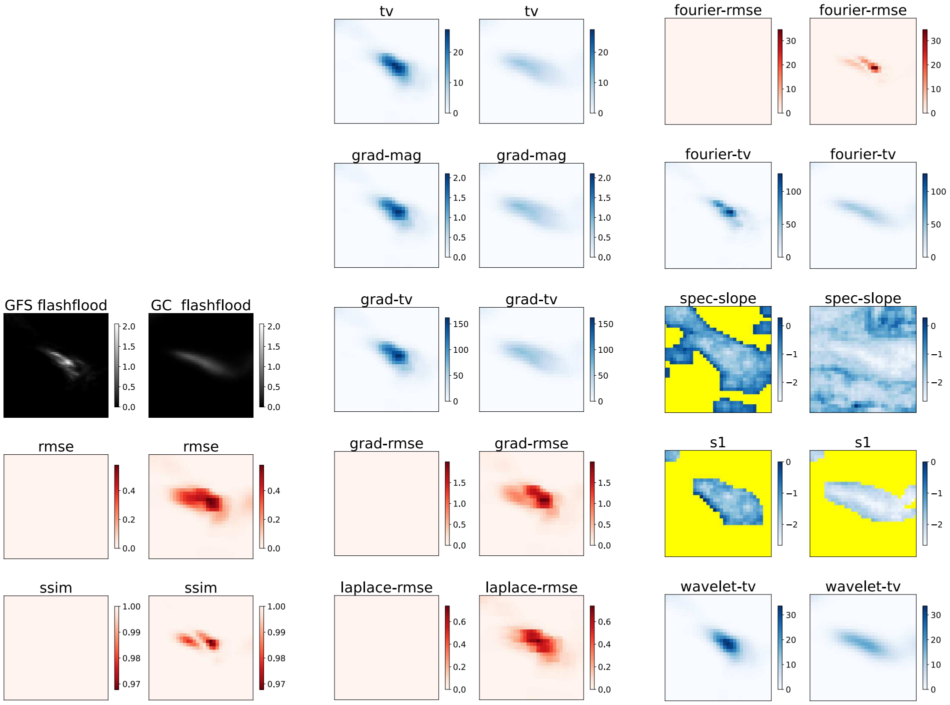
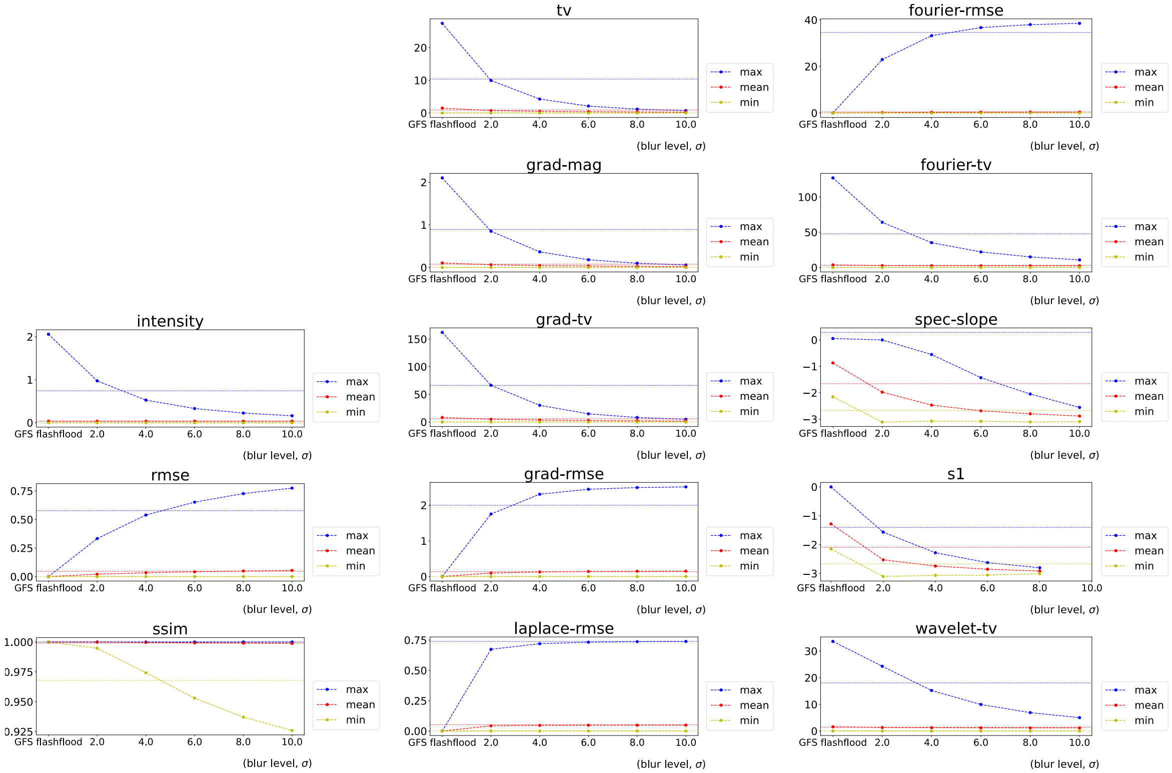


Figure 10: Heatmaps for the flashflood event for GFS vs. GraphCast forecasts of 6h accumulated precipitation at lead time of 48 hours. The contrast threshold for metric S_1 used here is 0.1.

again shows that it is measuring significant sharpness in very dark regions that we are not interested in, so we once again discard it from consideration.

The two examples above compare the sharpness of AIWP-generated imagery to NWP-generated imagery. One should keep in mind that NWP-generated imagery itself comes from a numerical simulation of the real world, and is in fact much smoother than observations, e.g., satellite or radar imagery. A long-term goal of AI algorithms will be to match the sharpness of real observations.



(a)

	Intensity	RMSE	SSIM (min)	TV	Grad-Mag	Grad-TV	Grad-RMSE	Laplace-RMSE	Fourier-RMSE	Fourier-TV	Spec-Slope	S_1	Wavelet-TV
$\sigma^{\text{max}}_{\text{equivalent}}$	≈ 3	≈ 4.5	≈ 4.5	≈ 2	≈ 2	≈ 2	≈ 3	≈ 8	≈ 5	≈ 3	−−	≈ 1.8	≈ 3.5

(b)

Figure 11: (a) GBE plots for the flashflood event, comparing GFS vs. GraphCast forecasts of 6h accumulated precipitation at lead time of 48 hours. The contrast threshold for metric S_1 used here is 0.1. (b) Max GBE values in (a) estimated visually from blue curves and lines. Note that for SSIM the min values estimated from the yellow curves and lines is used. ”−−” indicates a value discarded because of the corresponding heatmap.

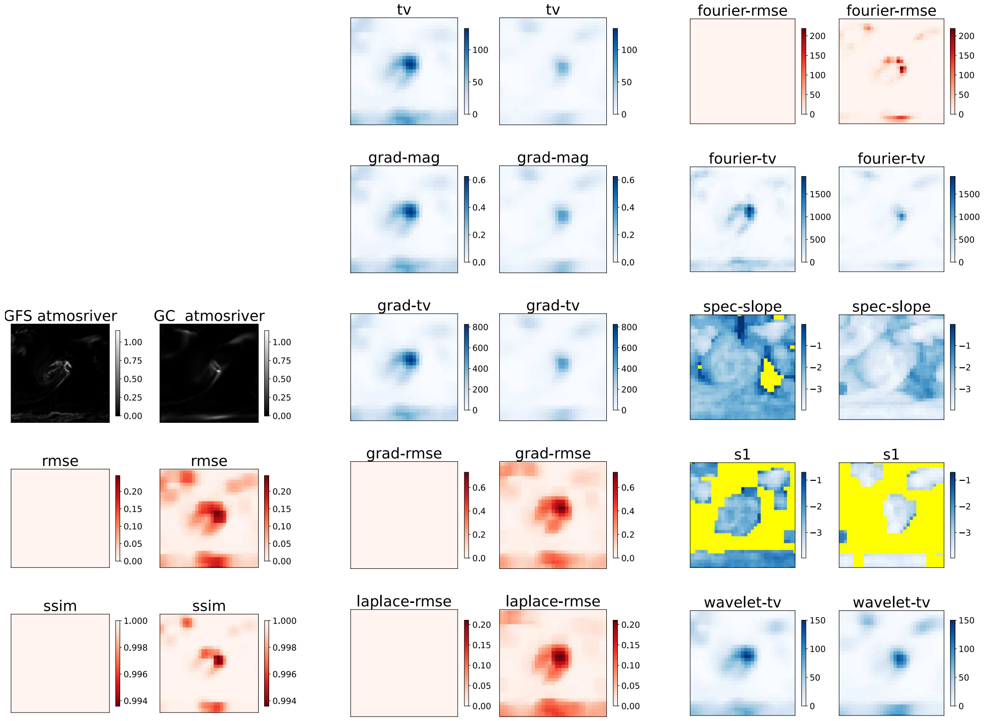
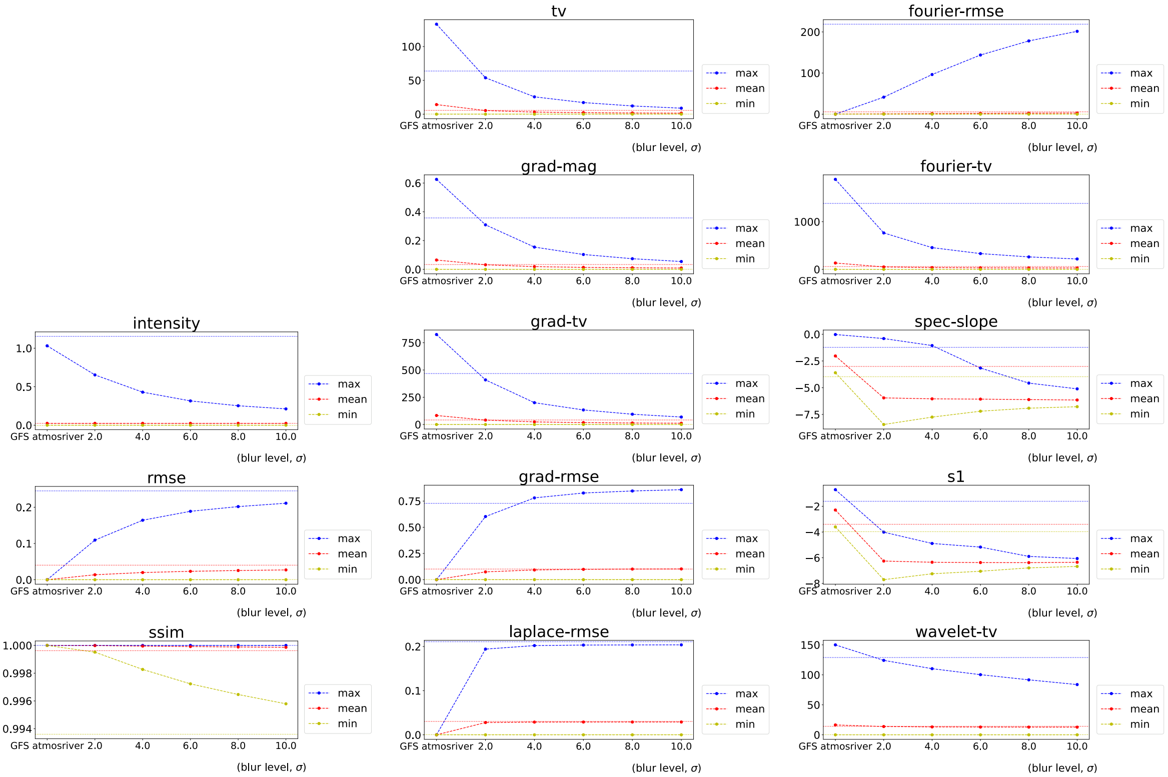


Figure 12: Heatmaps for the atmospheric river event for GFS vs. GraphCast forecasts of 6h accumulated precipitation at lead time of 168 hours. The contrast threshold for metric S_1 used here is 0.1.



(a)

	Intensity	*RMSE*	*SSIM (min)*	*TV*	*Grad-Mag*	*Grad-TV*	*Grad-RMSE*	*Laplace-RMSE*	*Fourier-RMSE*	*Fourier-TV*	*Spec-Slope*	S_1	*Wavelet-TV*
$\sigma^{\max}_{\text{equivalent}}$	< 0	> 10	> 10	≈ 1.8	≈ 1.8	≈ 2	≈ 3.5	> 10	> 10	≈ 1	$--$	≈ 0.5	≈ 1.8

(b)

Figure 13: (a) GBE plots for the atmospheric river event, comparing GFS vs. GraphCast forecasts of 6h accumulated precipitation at lead time of 168 hours. The contrast threshold for metric S_1 used here is 0.1. (b) Max GBE values in (a) estimated visually from blue curves and lines. Note that for SSIM the min values estimated from the yellow curves and lines is used. ”$--$” indicates a value discarded because of the corresponding heatmap.

d. Discussion

From the examples in this section we learned a few lessons regarding the use of these metrics. The overarching theme of these lessons is that no single metric (or even combination of metrics) can serve as a "silver bullet" to determine sharpness in all situations. Each dataset and application will have different requirements and properties that affect how these metrics respond, and care is required in selecting which metrics to focus on. We can utilize the examples in this section as a set of case studies with which we illustrate this analysis process.

One of the first metrics to look at is simply intensity, as illustrated by the GFS vs. GraphCast flashflood example. As we will discuss shortly in Section 4, almost all of these metrics have their response directly correlated with intensity, and as such if the overall intensity ranges do not match, every other metric will be affected. Similarly, before diving into deeper analyses of sharpness, one should always print heatmaps and do a sanity check whether any metrics display undesirable behavior on this particular dataset. For instance, we found that in all three of these examples, the spectral slope metrics focused principally on "sharp" regions in the darkest perceivable regions around the features actually of interest, so we did not focus on them further in these analyses. Finally, the heatmaps can inform which GBE curve (min, mean, or max) to examine. Namely, in datasets in which the features of interest are relatively sparse - for example, the images in two of the three examples considered in this section, namely the two GFS-GraphCast comparisons, contain mostly flat, dark backgrounds with isolated regions of interest - an overwhelming proportion of the heatmap blocks will indicate "no sharpness" or "fully similar", and as such the mean GBE values are saturated with these values and become uninformative. For this reason, we focus on the max GBE curve for all metrics except for SSIM. For SSIM we utilize the min GBE curve, since its interpretation is reversed (its max value means maximal similarity).

Once we have done this pre-analysis, we can begin analyzing the GBE curves themselves. At this point, we are still asking the question of "which metrics are informative for my dataset?", and as such we are looking at the GBE reference curves rather than the GBE values of the evaluation images. An "informative" GBE curve is one that displays consistent drop-off in sharpness as the reference image is blurred more and more – for instance, in Figure 8(a) the mean GBE curves for TV, Grad-Mag, and Grad-TV are all fairly informative, while the mean GBE curve for Wavelet-TV is very uninformative (flat) in this case. We note that Fourier-TV in this case would not be considered particularly informative, because while there is a dramatic drop in the mean GBE curve from blur values 0 to 2, it is likely that a large portion of this is due to observation noise being reduced via blurring, as the mean GBE curve remains more or less flat after this initial drop. In Figures 11 and 13 we decided to get information from the max GBE curves (min for SSIM). There, Wavelet-TV and Fourier-TV have very informative max GBE curves, and while TV, Grad-Mag, and Grad-TV are all reasonably informative as well, they do suffer from more asymptotic flattening at higher blur levels.

Throughout this section so far, we have focused more on the univariate metrics than the bivariate ones. We have done this intentionally as our focus is on the sharpness itself, which is most easily compared via univariate metrics. All of the bivariate metrics include some degree of similarity as well as sharpness, and as such are most useful for asking more complicated questions of the data such as “are the sharp features in these two regions well-aligned?” Answering these types of questions is best done using a combination of univariate and bivariate metrics: the univariate metric response can be used to disambiguate which portions of the bivariate response are due to a lack of sharp features and which portions are due to misaligned sharp features. However, such analyses are by their nature more involved, and are generally beyond the scope of this paper.

We can finally utilize the GBE values themselves to analyze how sharp each of these examples are. Examining the numbers in Fig. 8(b) and focusing on TV, Grad-Mag, and Grad-TV, this case has a mean GBE of approximately 2, indicating that the GREMLIN model is indeed losing some of the sharp features present in the radar observations. However, as discussed in Section 3b, some of this lack of perceived sharpness may be due to a degree of noise in the observed radar data that would not be desirable for GREMLIN to recreate.

Next, examining the numbers in Figure 11(b) for the GFS-GraphCast example of flashflood and focusing on TV, Grad-Mag, Grad-TV, Fourier-TV, and Wavelet-TV, the max GBE values are between 2 and 3.5. However, recall that this flashflood example had significantly lower intensity than the reference image, and as such these blurriness values are exaggerated, and the “true” max GBE value is likely between 1 and 2. This matches visual inspection: the reference GFS image itself does not contain many sharp features, so the GraphCast prediction is proportionally not as blurry. Finally, examining the numbers in Figure 13(b) for the atmospheric river example and focusing on the same set of metrics as in the flash flood example, we obtain consistently smaller max GBE values, ranging from 1 to 2, which roughly matches our intensity-corrected estimate for the flashflood event.

4. Important Properties of Sharpness Metrics

There are several properties of these metrics that are relevant to explore using more abstract analysis techniques. In this section, we present several experiments that illustrate how these metrics respond to:

- What effect does adding white noise have on these metrics? (Section 4a.)
- How do the metrics change if we 1) translate, 2) change intensity, or 3) change the edge sharpness of a synthetic example? (Section 4b.)

In addition to these practical experiments, we also analyzed certain properties of the metrics theoretically. Specifically, we show the effect on metrics resulting from changing resolution, shifting the intensity range by a fixed offset, and scaling the intensity range by a fixed scaling factor. For a quick summary of those theoretical results, see Table 1, and for the full details see Supplemental Sections S2b, S2c, and S2d.

There are a few takeaways from this analysis. First, we note that changing the resolution affects only those metrics that utilize total variation (e.g. TV, gradient TV, Fourier TV, and Wavelet TV) and does so in a predictable way (increases proportionally to the number of pixels), so making sharpness comparisons across differing resolutions should be relatively straightforward. On the other hand, metrics have very different responses to shifting and scaling intensity. Most metrics (aside from SSIM, Spec-Slope, and S_1) scale proportionally as intensity is rescaled (i.e. multiplied by a scaling factor), while all the metrics aside from those three and wavelet total variation are invariant to shifting the intensity range (i.e. adding or subtracting a fixed value). The upshot of this is that when comparing data that needs to be normalized, most metrics can be compared fairly easily (with just the scaling factor taken into account) but additional care needs to be taken for SSIM, spectral slope, S_1, and wavelet total variation.

We also present derivations of the computational complexity of each metric in Supplemental Sections S2e and S2f, with a summary of those results along with practical timings of the computation of each metric on representative examples in Table 2.

Table 1: Summary of Metric Properties. For details, see Supplemental Sections S2b, S2c, and S2d.

Image Change	Invariant Metrics	Non-invariant metrics
Resolution	All others	TV-based metrics increase with square of edge length
Shifting intensity	All others	SSIM, Spec-Slope, S_1, & Wavelet-TV change non-linearly
Scaling intensity	SSIM*, Spec-Slope	All others scale proportionally, except S_1 (non-linear)

*SSIM is invariant to scaling the overall dynamic range of the input type (allowing comparisons between 0-1 and 0-255 scaled data) but not to scalings within a given dynamic range.

Table 2: Computational Complexity of Metrics

Metric	Big O	64×64 wall clock	128×128 wall clock
RMSE	$O(n)$	0.0103 sec	0.0113 sec
SSIM	$O(n)$	0.1379 sec	0.1580 sec
TV	$O(n)$	0.0260 sec	0.0295 sec
Grad-Mag	$O(n)$	0.0447 sec	0.0509 sec
Grad-TV	$O(n)$	0.0414 sec	0.0464 sec
Grad-RMSE	$O(n)$	0.0385 sec	0.0446 sec
Laplace-RMSE	$O(n)$	0.0197 sec	0.0223 sec
Fourier-RMSE	$O(n \log n)$	0.0852 sec	0.0946 sec
Fourier-TV	$O(n \log n)$	0.0807 sec	0.0908 sec
Spec-Slope	$O(n \log n)$	0.9053 sec	1.2101 sec
S_1	$O(n \log n)$	0.9723 sec	1.3772 sec
Wavelet-TV	$O(n)$	0.1662 sec	0.1740 sec

Note: all wall clock times are an average from five similar computations with the same base image.

a. Effect of adding white noise

In this section we consider the question of how the various metrics respond to adding white noise to an image. In Section 3, we discussed an algorithm based on blurring the image to reduce sharpness – in this subsection, we explore how sharpness values can be artificially inflated by the presence of noise. This section serves as an illustration of how these metrics respond to a signal that a human would not regard as making an image "sharper." As an illustrative example we use a satellite image of a cloud, gradually add white noise to it, and track the values of all metrics in this image sequence. The white noise is added separately for each pixel by drawing from a Gaussian distribution with σ set to be a factor times the maximal intensity of the original image. The factor is provided on top of the first row of images in Figures 14.

Observations: All metrics increase with increasing white noise in the image, but to varying degree. Wavelet-TV seems to be least affected by the impact of white noise, with the mean increasing by less than 100% for $\sigma = 1.0$, while all other sharpness metrics increase dramatically with white noise. Thus, Wavelet-TV stands out as being the most invariant to the addition of white noise in an image.

This result highlights a potential issue with utilizing one or more of these metrics inside the loss function of a neural network: **if the network has as one of its goals increasing "sharpness" as measured by these metrics, one way for the network to achieve that goal is to increase the white noise in its output images, which is generally not a desirable property for a neural network.** Thus, great care must be taken in incorporating and properly weighting any term in a loss function incorporating these metrics. In part because of issues like this, a detailed study on the use of sharpness metrics within loss functions is beyond the scope of this paper, but we refer the reader to Section S1c in the supplemental material for a brief example of the impacts of sharpness metrics in a neural network loss function.

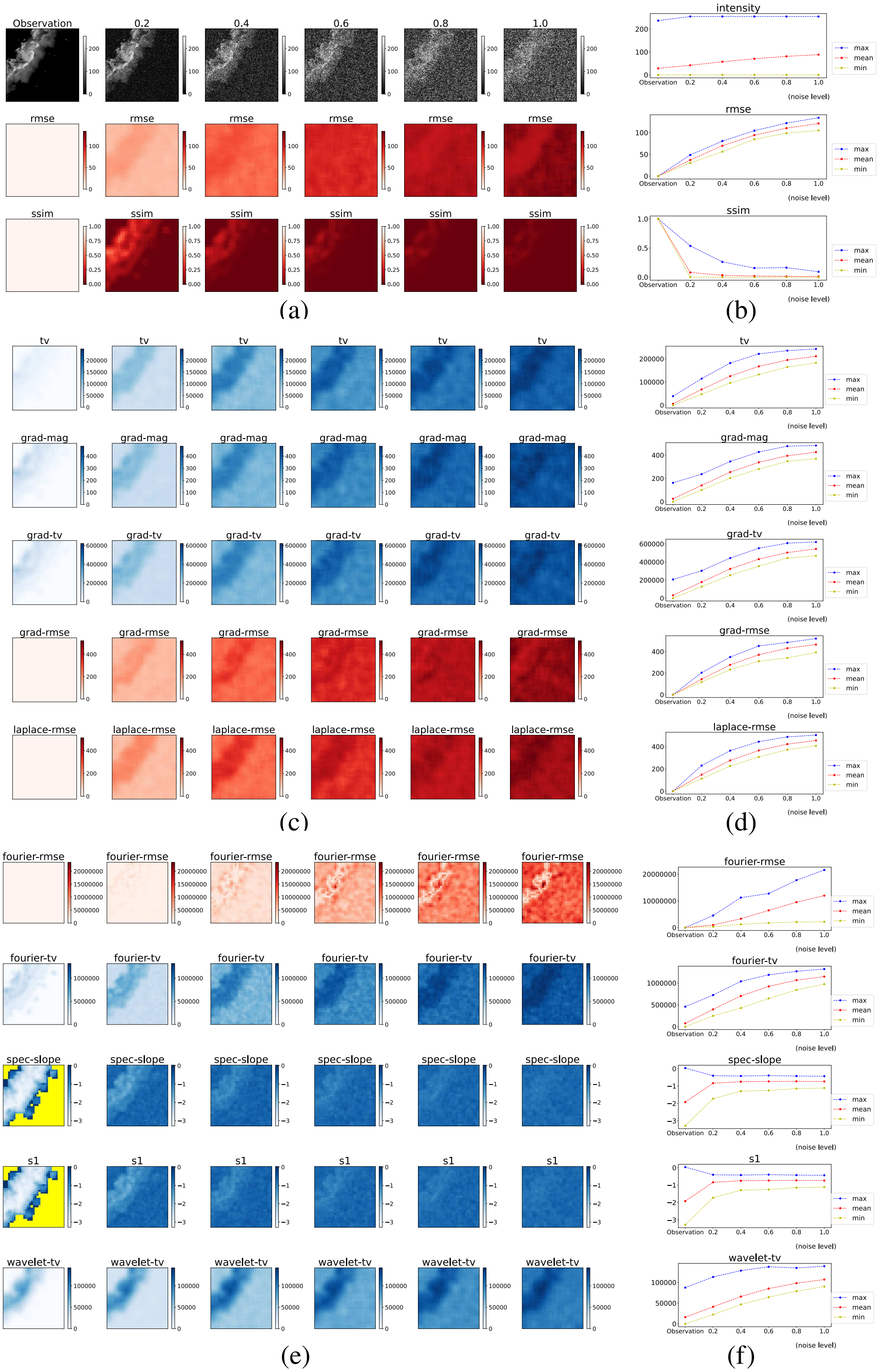


Figure 14: Response of metrics to adding white noise to a satellite image. (a,c,e) show heatmaps and (b,d,f) RMV plots for Metric Groups 1, 2, and 3, respectively.

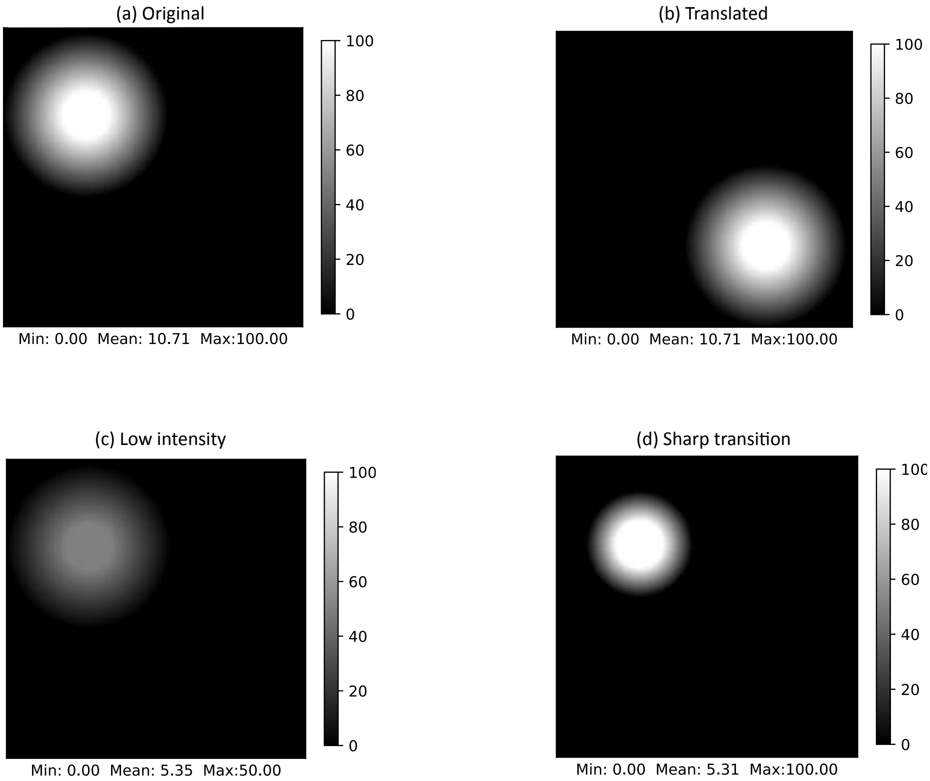

(a) Original
100
80
60
40
20
0
Min: 0.00 Mean: 10.71 Max:100.00
(b) Translated
100
80
60
40
20
0
Min: 0.00 Mean: 10.71 Max:100.00
(c) Low intensity
100
80
60
40
20
0
Min: 0.00 Mean: 5.35 Max:50.00
(d) Sharp transition
100
80
60
40
20
0
Min: 0.00 Mean: 5.31 Max:100.00

Figure 15: Synthetic images for experiments in Fig. 16. (a) Reference image: Background with intensity 0 with a single blob-like feature - a disk of radius 20 pixels, intensity 100 units, and linear drop off of 2 units/pixel from boundary of blob outwards. (b) Translated image: blob is translated to a different, non-overlapping location, (c) Low intensity image: Same as (a) but intensity is divided by 2.0 throughout the entire image, (d) Sharp transition image: size and intensity of central disk of blob is the same as in (a) but rate of drop-off is twice as steep, namely decrease of 4 units/pixel from the disk's boundary outwards.

b. Effect of feature location, feature intensity, and speed of transition

We conduct three experiments that illustrate the effect that specific changes to an image have on the various metrics. Figures 15(a) shows the reference image that has one blob-like feature consisting of a disk with linear intensity drop-off from the disk boundary outward. Figures 15(b-d) show modifications of the image, namely translation of the feature (b), halving intensity of the image (c), and doubling rate of boundary transition from disk to background (d). Fig. 16 shows pair-wise comparisons of the reference image to each of the modified images.

Examining these heatmaps, there are a few things to note. When we translate the disk, all the univariate metrics translate along with it, precisely as expected. In each heatmap block, the bivariate metrics are effectively comparing the presence of a disk vs. a blank background at each location

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-24-0083.1.

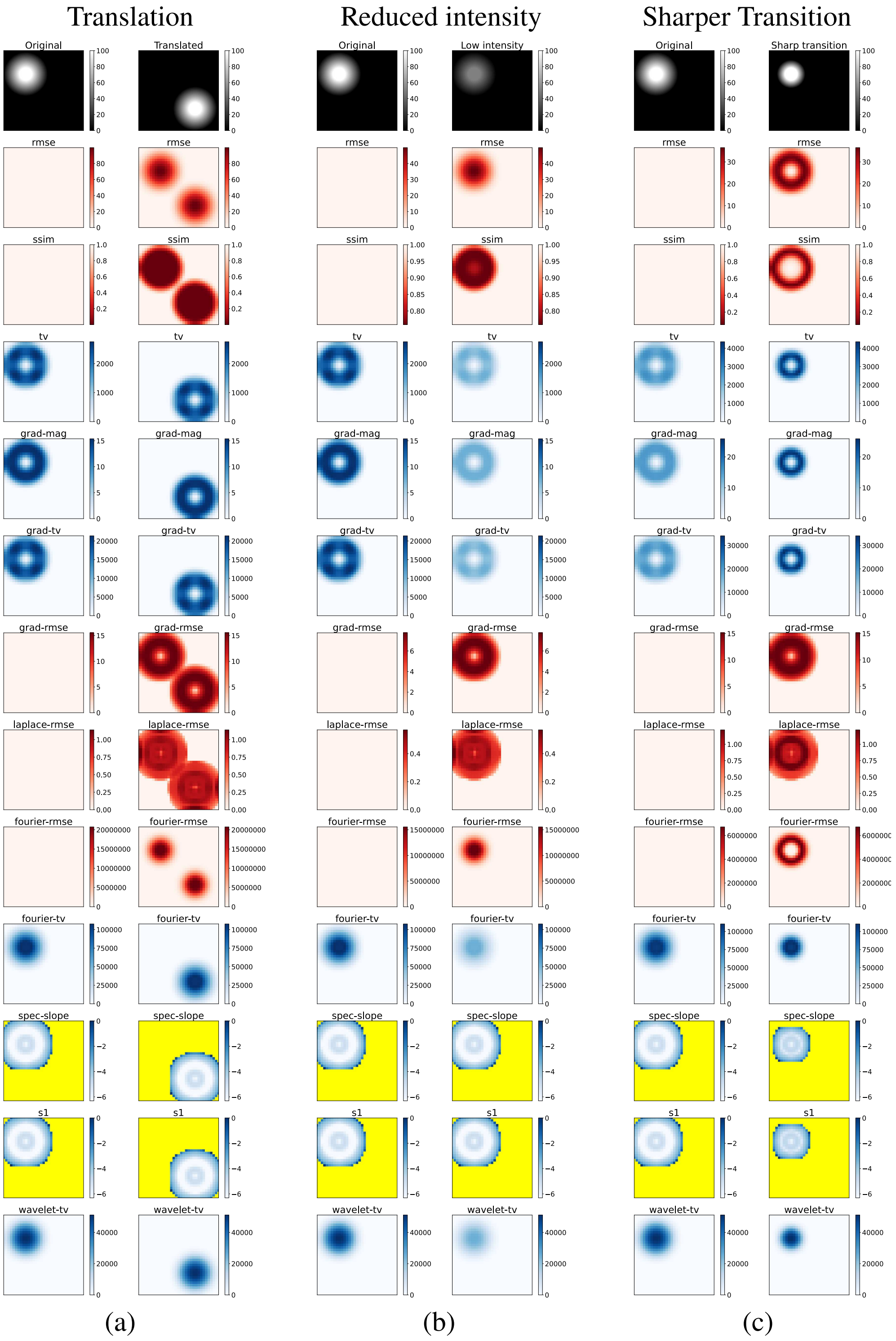


Figure 16: Heatmaps for all metrics demonstrating effect of changing (a) the location of a feature, (b) intensity of an image (including of all features), and (c) the rate of transition from feature to background.

for this image pair (because the old and new location do not overlap here), and as such we obtain two “bumps”: one at the original location, and one at the new location. This is the well-known *double-penalty problem* in the assessment of weather forecasting skills (Zhao and Zhang 2018). It is, however, worth noting how these bivariate metrics fall off: RMSE has a fairly linear fall off with the linear decrease in intensity, as does Fourier-RMSE. On the other hand, SSIM has quite large values anywhere that there is any discrepancy between the images, all the way right out to the edges of the blobs where it finally starts falling off. Finally, Grad-RMSE and Laplace-RMSE exhibit a maximum intensity partway to the edge of the disk, approximately where the linear fall off begins - this makes sense, as that is the location where the gradient is largest, and thus gradient discrepancies.

When we change intensity, we note that our theoretical results reported in Table 1 hold true: All of the metrics except for SSIM, Spec-Slope, and S_1 reduce in intensity by approximately 50%. This is easy to see for the univariate metrics, and we can see it in the bivariate metrics by comparing the values found in column (b) with those in column (a). SSIM, here, displays some degree of difference coming from the change in intensity, with the largest differences being in the fall-off region rather than at the center of the disk. Finally, Spec-Slope remains identical over changes in intensity, but while S_1 shows similar minimum values between intensities, because the contrast threshold was not changed between the two images, more heatmap regions fell within that threshold and were cut out, resulting in a slightly smaller valid region.

Finally, as we make the transition region sharper, we can see this reflected in the metrics. The accuracy metrics (RMSE and SSIM) pick up on the fact that only the transition region has changed, marking the center of the disk as “unchanged.” All of the univariate metrics display similar behavior to one another again, and both constrict the region of detected sharpness and intensify it, corresponding to a narrower, sharper edge. Grad-RMSE has a very similar shape to when the disk was compared against a blank region of the image in column (a), but while the maximum value is about the same, the mean value is much smaller when only the transition rate has been changed. Laplace-RMSE on the other hand displays similar overall statistic values (similar maximum, slightly reduced mean) but the heatmap shows a distinct concentration of high values right around the region of sharper transition that was not present in corresponding location in column (a).

Finally, Fourier-RMSE displays a completely different pattern than it did in column (a), focusing in very precisely on the region of increased transition.

c. Discussion

The metrics presented here are largely invariant to changes in resolution (with the notable exception of TV-based metrics), but are affected, in general, by changes in intensity. These effects are mitigated by utilizing derived tools, like the GBE measure introduced in the previous section, and the standard practice of normalizing data to a common, fixed range (such as $0-1$ or $0-255$) will also ensure that variances due to intensity differences are minimized. While it is generally good practice to share processing details such as normalization steps that were performed, data range information, and analysis resolution, it is particularly critical when these metrics are being used, as that information can enable fair comparison and replication from other groups.

It is worth noting that because of the high sensitivity of these metrics to white noise, applying a denoising step before computing sharpness could be a useful preprocessing step. While analysis of such denoising techniques is beyond the scope of this paper, we suggest that starting with the median filter (Huang et al. 1979) would be a reasonable approach.

We also note that while GBE is theoretically invariant due to its proportional nature, in practice numerical and discretization errors as well as the stochastic nature of some of the transformations discussed here (such as adding white noise) mean that in practice there may be some changes to derived GBE statistics.

5. Discussion and Future Work

We presented here a set of metrics that the community can use to evaluate sharpness and other properties. We hope to have provided an initial understanding of how these various metrics apply to several distinct applications and how their properties affect those results. Some of the key insights we gained include:

- All of these metrics are correlated with and can indicate trends in what we as humans consider "sharp," but none fully capture the subjective, perceptual nature of sharpness. A good illustration of this is the experiment in Subsection 4a: most people would not consider a noisier image to be "sharper," but every metric we tested registered significantly increased sharpness

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-24-0083.1.

in the presence of added noise. **Thus, we should remember that these metrics are measuring idiosyncratic quantities that are each just a proxy of what humans perceive as image sharpness.**

- As a result, there is no “catch-all” metric that will work in every case. Instead, it is always important to test a variety of tools both quantitatively and qualitatively on each dataset and choose the ones that fit.
- There are many ways to assess which of these metrics are compatible with a given dataset. Section 3d summarizes many of our “lessons learned” on understanding how to make those choices. Some of the most important lessons are:
  - Before using any metric one should look at a) intensity, b) heatmaps, and c) GBE curves.
  - Undesired noise in the reference images (e.g., from observations) needs to be taken into account.
  - Any overall difference in intensity must be considered, as that can strongly affect all other sharpness results (with the notable exception of S_1).
- While many of these metrics possess similar properties to one another, there are distinct differences in how they respond to certain changes in the input data, particularly resolution, location, intensity, and sharpness of transitions that should be taken into account when selecting metrics.

We have only scratched the surface of this topic and suggest to expand this study by exploring the following topics:

- A more in-depth study of which kinds of sharpness the various metrics primarily focus on, e.g. from edges vs. texture.
- Adopting a better way to avoid NaNs as output of metrics, such as discussed in Vu et al. (2011) for Spec-Slope.
- Adding some of the metrics discussed in Section 2 that we dropped for this study.
- Testing alternate methods of blurring for techniques derived from GBE, such as bilateral blurring (Tomasi and Manduchi 1998).

- Developing more best practices / protocols for the use of these metrics.
- Conducting experiments to test how much the spherical harmonic loss function recently proposed by Subich et al. (2025) improves sharpness.

Another important topic - one that is only briefly touched on in Subsection S1c of the supplemental document - is the question of how to effectively use these metrics to *train* neural networks, rather than just to *evaluate* them. We hope to explore this topic in a follow-up paper.

Acknowledgments. We would like to thank the following scientists for their excellent feedback for this manuscript: Ryan Lagerquist at CSU, Randy Chase at tomorrow.io, the three anonymous reviewers, and the editor, Mark Veillette. This material is based upon work supported by the National Science Foundation under AI institute Grant No. ICER-2019758. Co-author Molina was supported by the National Science Foundation (NSF) under Grant Number 2425735 and the U.S. Department of Energy, Office of Science DE-SC0022070. Co-author Hilburn was partly supported by GOES-R Program under Grant NA19OAR4320073. Co-author Campbelll acknowledges funding support from the U.S. Naval Research Laboratory (NRL) base project number (N0001423WX00011). Co-authors Schreck, Petzke, Gagne are also supported by the NSF National Center for Atmospheric Research, which is a major facility sponsored by the U.S. National Science Foundation under Cooperative Agreement No. 1852977.

Data availability statement. We provide Python code that implements the metrics discussed here, code for their visualization, and the imagery used for testing, on GitHub at `https://github.com/ai2es/sharpness/`.

References

Bi, K., L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, 2023: Accurate medium-range global weather forecasting with 3D neural networks. *Nature*, **619 (7970)**, 533–538.

Blau, Y., and T. Michaeli, 2018: The perception-distortion tradeoff. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6228–6237.

Bonev, B., T. Kurth, C. Hundt, J. Pathak, M. Baust, K. Kashinath, and A. Anandkumar, 2023: Spherical Fourier neural operators: Learning stable dynamics on the sphere. *International conference on machine learning*, PMLR, 2806–2823.

Briggs, W. M., and R. A. Levine, 1997: Wavelets and field forecast verification. *Monthly Weather Review*, **125 (6)**, 1329–1341.

Brown, B. G., R. Bullock, J. H. Gotway, D. Ahijevych, C. Davis, E. Gilleland, and L. Holland, 2007: Application of the mode object-based verification tool for the evaluation of model precipitation fields. *22nd Conf. on Weather Analysis and Forecasting/18th Conf. on Numerical Weather Prediction*, https://ams.confex.com/ams/pdfpapers/124856.pdf.

Buschow, S., and P. Friederichs, 2020: Using wavelets to verify the scale structure of precipitation forecasts. *Advances in Statistical Climatology, Meteorology and Oceanography*, **6 (1)**, 13–30.

Davis, C., B. Brown, and R. Bullock, 2006: Object-based verification of precipitation forecasts. part i: Methodology and application to mesoscale rain areas. *Monthly Weather Review*, **134 (7)**, 1772–1784.

Demuth, J. L., and Coauthors, 2020: Recommendations for developing useful and usable convection-allowing model ensemble information for nws forecasters. *Weather and Forecasting*, **35 (4)**, 1381–1406.

Dorninger, M., E. Gilleland, B. Casati, M. P. Mittermaier, E. E. Ebert, B. G. Brown, and L. J. Wilson, 2018: The setup of the mesovict project. *Bulletin of the American Meteorological Society*, **99 (9)**, 1887–1906.

Gilleland, E., D. Ahijevych, B. G. Brown, B. Casati, and E. E. Ebert, 2009: Intercomparison of spatial forecast verification methods. *Weather and forecasting*, **24 (5)**, 1416–1430.

Gilleland, E., D. A. Ahijevych, B. G. Brown, and E. E. Ebert, 2010: Verifying forecasts spatially. *Bulletin of the American Meteorological Society*, **91 (10)**, 1365–1376.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, 2014: Generative adversarial nets. *Advances in neural information processing systems*, **27**.

Haynes, K., R. Lagerquist, M. McGraw, K. Musgrave, and I. Ebert-Uphoff, 2023: Creating and evaluating uncertainty estimates with neural networks for environmental-science applications. *Artificial Intelligence for the Earth Systems*, **2 (2)**, 220 061.

Hilburn, K. A., I. Ebert-Uphoff, and S. D. Miller, 2020: Development and interpretation of a neural-network-based synthetic radar reflectivity estimator using goes-r satellite observations. *Journal of Applied Meteorology and Climatology*, **60 (1)**, 3–21.

Huang, T., G. Yang, and G. Tang, 1979: A fast two-dimensional median filtering algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **27 (1)**, 13–18, https://doi.org/10.1109/TASSP.1979.1163188.

Jolliffe, I. T., and D. B. Stephenson, 2012: *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons.

Lagerquist, R., and I. Ebert-Uphoff, 2022: Can we integrate spatial verification methods into neural network loss functions for atmospheric science? *Artificial Intelligence for the Earth Systems*, **1 (4)**, e220 021.

Lam, R., and Coauthors, 2023: Learning skillful medium-range global weather forecasting. *Science*, **382 (6677)**, 1416–1421, https://doi.org/10.1126/science.adi2336.

Ledig, C., and Coauthors, 2017: Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

Li, L., R. Carver, I. Lopez-Gomez, F. Sha, and J. Anderson, 2024: Generative emulation of weather forecast ensembles with diffusion models. *Science Advances*, **10 (13)**, eadk4489.

NCEP, 2024: NCEP products inventory. Website, accessed: 2024-4-30, https://www.nco.ncep.noaa.gov/pmb/products/gfs/.

Pfreundschuh, S., P. J. Brown, C. D. Kummerow, P. Eriksson, and T. Norrestad, 2022: Gprof-nn: A neural-network-based implementation of the goddard profiling algorithm. *Atmospheric Measurement Techniques*, **15 (17)**, 5033–5060.

Price, I., and Coauthors, 2025: Probabilistic weather forecasting with machine learning. *Nature*, **637 (8044)**, 84–90.

Radford, J. T., I. Ebert-Uphoff, J. Q. Stewart, K. D. Musgrave, R. DeMaria, N. Tourville, and K. Hilburn, 2025: Accelerating community-wide evaluation of ai models for global weather prediction by facilitating access to model output. *Bulletin of the American Meteorological Society*, **106 (1)**, E68 – E76, https://doi.org/10.1175/BAMS-D-24-0057.1, URL https://journals.ametsoc.org/view/journals/bams/106/1/BAMS-D-24-0057.1.xml.

Rasp, S., and Coauthors, 2024: Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *Journal of Advances in Modeling Earth Systems*, **16 (6)**, e2023MS004 019.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, **136 (1)**, 78–97.

Ronneberger, O., P. Fischer, and T. Brox, 2015: U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., Springer International Publishing, Cham, 234–241.

Selz, T., W. Bruinsma, G. C. Craig, S. Markou, R. Turner, and A. Vaughan, 2025: On the effective resolution of ai weather prediction models. *Authorea Preprints*.

Skamarock, W. C., 2004: Evaluating mesoscale nwp models using kinetic energy spectra. *Monthly weather review*, **132 (12)**, 3019–3032.

SLR Lounge, 2023: SLR lounge / sharpness. Website, accessed on 05/05/2023, https://www.slrlounge.com/glossary/sharpness-photography-definition/.

Sohl-Dickstein, J., E. Weiss, N. Maheswaranathan, and S. Ganguli, 2015: Deep unsupervised learning using nonequilibrium thermodynamics. *International conference on machine learning*, PMLR, 2256–2265.

Subich, C., S. Z. Husain, L. Separovic, and J. Yang, 2025: Fixing the double penalty in data-driven weather forecasting through a modified spherical harmonic loss function. *arXiv preprint arXiv:2501.19374*.

Tomasi, C., and R. Manduchi, 1998: Bilateral filtering for gray and color images. *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, 839–846, https://doi.org/10.1109/ICCV.1998.710815.

Turner, D., and Coauthors, 2020: A verification approach used in developing the rapid refresh and other numerical weather prediction models. *Journal of Operational Meteorology*, **8 (3)**.

Vu, C. T., T. D. Phan, and D. M. Chandler, 2011: s_3: a spectral and spatial measure of local perceived sharpness in natural images. *IEEE transactions on image processing*, **21 (3)**, 934–945.

Wang, Z., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, 2004: Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, **13 (4)**, 600–612.

Wedi, N., and Coauthors, 2015: The modelling infrastructure of the integrated forecasting system: Recent advances and future challenges. Tech. rep., European Centre for Medium-Range Weather Forecasts. Technical Memorandum No. 760.

Wee, C.-Y., and R. Paramesran, 2008: Image sharpness measure using eigenvalues. *2008 9th International Conference on Signal Processing*, IEEE, 840–843.

Zhang, R., P. Isola, A. A. Efros, E. Shechtman, and O. Wang, 2018: The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhao, B., and B. Zhang, 2018: Assessing hourly precipitation forecast skill with the fractions skill score. *Journal of Meteorological Research*, **32 (1)**, 135–145.