Supplemental Material

© Copyright 2022 American Meteorological Society (AMS)
For permission to reuse any portion of this work, please contact
permissions@ametsoc.org. Any use of material in this work that is determined to be “fair
use” under Section 107 of the U.S. Copyright Act (17 USC §107) or that satisfies the
conditions specified in Section 108 of the U.S. Copyright Act (17 USC §108) does not
require AMS's permission. Republication, systematic reproduction, posting in electronic
form, such as on a website or in a searchable database, or other uses of this material,
except as exempted by the above statement, requires written permission or a license
from AMS. All AMS journals and monograph publications are registered with the
Copyright Clearance Center (https://www.copyright.com). Additional details are provided
in the AMS Copyright Policy statement, available on the AMS website
1. Procedures for implementing spectral filters

In this section we provide a detailed description of the procedures for Fourier and wavelet decomposition, so that others may reproduce our results and use these methods in their own research. We also provide code (see link in “Data-availability Statement”).

Let $\delta_{\text{min}}$ and $\delta_{\text{max}}$ be the minimum and maximum spatial resolutions allowed through the spectral filter. Thus, $\lambda_{\text{min}} = 2\delta_{\text{min}}$ and $\lambda_{\text{max}} = 2\delta_{\text{max}}$ are the minimum and maximum wavelengths, while $\nu_{\text{min}} = \frac{1}{\lambda_{\text{max}}}$ and $\nu_{\text{max}} = \frac{1}{\lambda_{\text{min}}}$ are the minimum and maximum wavenumbers.

a. Fourier decomposition

For one target field – i.e., the convection mask at one time step – the procedure is described below. See Figure 7 in the main text for a schematic illustration, where the spectral filter allows wavelengths from $\lambda_{\text{min}} = 0.5^\circ$ to $\lambda_{\text{max}} = 2^\circ$.

1. Tapering in spatial domain. The Fourier transform assumes a periodic domain, so if we analyze a finite spatial domain, the Fourier transform implicitly assumes that the domain is toroidal. In other words, the Fourier transform assumes that the top and bottom rows are adjacent, also that the left and right columns are adjacent – causing spurious patterns in the wavelength domain. To prevent these spurious patterns, we taper the data before the forward transform, then undo tapering after the inverse transform (Step 6). Specifically, we triple the grid size from $205 \times 205$ to $615 \times 615$ and fill the new pixels with zeros (indicating no convection).

2. Windowing in spatial domain. Even after tapering, the Fourier transform’s assumption of a periodic domain is problematic. As shown in Figure 8.7 of Stull (1988), a smoothly varying pattern in a finite spatial domain can appear as a sawtooth pattern in the wavelength domain,
containing many spurious wavelengths called “red noise”. In general, if data in a finite spatial
domain are not processed with an appropriate window, the result is leakage in the wavelength
domain, where the estimated amount of signal (i.e., amplitude of Fourier coefficient) at one
wavelength is contaminated by neighbouring wavelengths. See Section 8.4.3 of Stull (1988)
for details.

We use a 2-D, radially symmetric version of the Blackman-Harris window (Harris 1978),
defined as:

$$
\begin{cases}
    w(r_g) = 0.42 - 0.5 \cos(\pi \left[ 1 + \frac{r_g}{R} \right]) + 0.08 \cos(2\pi \left[ 1 + \frac{r_g}{R} \right]), & \text{if } r_g \leq R; \\
    w(r_g) = 0, & \text{if } r_g > R.
\end{cases}
$$

(1)

$r_g$ is the distance of grid point $g$ from the domain center; $R$ is the maximum distance; and

$w$ is the resulting weight. We set $R$ to 307 grid points, the half-width of the 615-by-615
tapered grid. We multiply the weight matrix elementwise with the original field (predictions
or observations).

3. **Forward transform.** Use the Fourier transform to convert the spatial field into the wavelength
domain. For a spatial domain of $N \times N$ grid points, the Fourier transform outputs an $N \times N$
complex-valued coefficient matrix in the wavelength domain. For each spatial direction
this matrix contains one coefficient for the infinite wavelength (zero frequency), $\frac{1}{2}(N - 1)$
coefficients for positive wavelengths, and $\frac{1}{2}(N - 1)$ coefficients for negative wavelengths. We
split the coefficient matrix into two: one containing the magnitudes, the other containing the
phases, of the complex-valued coefficients.

4. **Filtering in wavelength domain.** The naïve approach would be to zero out coefficients at
undesired wavelengths (i.e., use a rectangular filter), but this causes artifacts in the spatial
domain after the inverse transform. Instead, we use a second-order Butterworth (1930) filter,
which decreases the magnitude of all coefficients (the phase remains unchanged) but especially
those at undesired wavelengths. We use a band-pass filter, which is a superposition of low-pass
and high-pass filters. The low-pass filter admits low frequencies (large wavelengths), while
the high-pass filter admits high frequencies (small wavelengths). For each wavenumber pair
\((k, l)\) – where \(k\) is the zonal wavenumber and \(l\) is the meridional wavenumber, both in units
of \(m^{-1}\) – the gain of the Butterworth filter is defined as\(^1\):

\[
\begin{align*}
g_{\text{low}}(k, l) &= \frac{1}{1 + \left( \frac{\sqrt{k^2 + l^2}}{\nu_{\text{max}}} \right)^{2\alpha}}; \\
g_{\text{high}}(k, l) &= 1 - \frac{1}{1 + \left( \frac{\sqrt{k^2 + l^2}}{\nu_{\text{min}}} \right)^{2\alpha}}.
\end{align*}
\]

\(\sqrt{k^2 + l^2}\) is the total wavenumber; \(\alpha = 2\) is the order of the filter; \(g_{\text{low}} \in [0, 1]\) is the gain for
the low-pass filter; and \(g_{\text{high}} \in [0, 1]\) is the gain for the high-pass filter. For each wavelength
pair we multiply the magnitude of the Fourier coefficient by \(g_{\text{low}}\) to implement the low-pass
filter, then \(g_{\text{high}}\) to implement the high-pass filter.

5. **Inverse transform.** Use the inverse Fourier transform to convert the coefficients back to
the spatial domain. If the coefficients were not filtered, the inverse transform would exactly
recover the original spatial field. With filtered coefficients, the inverse transform recovers the
part of the spatial field corresponding to the desired wavelengths.

6. **Undo tapering.** Remove zero-padding from the reconstructed spatial field.

\(b.\) **Wavelet decomposition**

For one target field – *i.e.*, the convection mask at one time step – the procedure is described
below. See Figure 8 in the main text for a schematic illustration, where the spectral filter allows

\(^1\)https://www.originlab.com/doc/Origin-Help/2DFFT-Filter-Algorithm#Butterworth
wavelengths from $\lambda_{\text{min}} = 0.1^\circ$ to $\lambda_{\text{max}} = 0.4^\circ$. For more information on wavelet decomposition, see Kumar and Foufoula-Georgiou (1994), Vidakovic and Müller (1994), and Versaci (2021). For wavelet decomposition involving 2-D images, we especially recommend the illustrations in Versaci (2021), although this paper serves only partly as a theoretical treatment of wavelet decomposition, devoting much space to technical documentation for the WaveTF software library.

1. **Tapering.** Taper the field by zero-padding. The wavelet transform needs both grid dimensions to be an integer power of 2, so we taper the grid from $205 \times 205$ to $256 \times 256$. As in Fourier decomposition, we fill the new pixels with zeros (indicating no convection).

2. **Forward transform.** Use the wavelet transform to convert the tapered field from the spatial domain to the wavelength domain. We use the WaveTF library (Versaci 2021) in Python, which offers the choice between two wavelet types: Haar and Daubechies. We use the Haar wavelet because it is less computationally expensive. For a spatial domain of $2^K \times 2^K$ grid points, the wavelet transform outputs coefficients at $K$ levels. At each level there are four sets of coefficients: LL (representing the low frequency, or mean, in both the vertical and horizontal), LH (representing the mean in the horizontal and high frequency [details] in the vertical), HL (the opposite), and HH (representing details in both the vertical and horizontal). At each successive level, the grid size halves while the wavelengths represented double.

3. **Filter out coefficients at undesired wavelengths.**

   (a) Zero out LL coefficients for all wavelengths $> \lambda_{\text{max}}$. In Figure 8 of the main text, this is done for levels 5 through 8 (where LL coefficients are for wavelengths of $0.8^\circ$, $1.6^\circ$, $3.2^\circ$, and $6.4^\circ$).
(b) At each level $k$ with wavelengths in the allowed range, starting with the deepest level, use
the inverse wavelet transform to reconstruct LL coefficients at level $k$ from all coefficients
at level $k + 1$. In Figure 8 of the main text, this is done for levels 4, then 3.

(c) At each level $k$ with wavelengths < $\lambda_{\text{min}}$, starting with the deepest level, use the inverse
wavelet transform to reconstruct LL coefficients at level $k$ from all coefficients at level
$k + 1$, then zero out all detail coefficients (LH, HL, and HH) at level $k$. In Figure 8 of
the main text, this is done for levels 2, then 1.

4. **Inverse transform.** Use the inverse wavelet transform to convert the filtered level-1 coef-
ficients back to the spatial domain. The inverse wavelet transform, like the inverse Fourier
transform, would exactly recover the original spatial field if coefficients were not filtered. With
filtered coefficients, the inverse transform recovers the part of the spatial field corresponding
to the desired wavelengths.

5. **Undo tapering.** Remove zero-padding from the reconstructed spatial field.

2. **Additional case study**

Figure S1 shows a winter case, with two time steps on Jan 25 2018: 1230 and 2230 UTC.
This case is almost trivial, as both time steps feature only one discrete storm – both examples of
convective initiation, where the storm exists at the valid time ($t_0 + 1$ hour) but not at the forecast-
issue time ($t_0$) – and no probabilities $\geq 0.05$ from any model. However, this case illustrates two
important characteristics of all models, including the pixelwise model and those with SELFs. First,
the models are poor at capturing weak isolated storms, especially for convective initiation. Second,
the models have excellent sharpness for low probabilities – *i.e.*, when there is little to no convection
in the domain, they are capable of predicting negligible probabilities (< 0.05) everywhere.
3. Additional experiment (2-hour lead time)

We conduct an experiment similar to that shown in the main body, but with a lead time of 2 hours instead of 1 hour. We hypothesized that SELFs would have a greater benefit at longer lead times, because longer lead times involve more advection (storm motion), which causes a greater spatial offset between predicted and observed convection, thus making the double penalty a bigger problem.

We train U-nets with 120 of the 336 loss functions used in the 1-hour experiment: those involving the FSS, Brier score, or cross-entropy. Based on results from the 1-hour experiment (Section 6a in main body), we believe that the other scores (IOU, Dice coefficient, CSI, Peirce score, Heidke score, and Gerrity score) are not worth considering here; including them would needlessly consume computing resources.

As for the 1-hour experiment, we separate our analysis into three steps: the preliminary, intermediate, and final analyses. In the preliminary and final analyses, where a conclusion is supported by a single time step, we have ensured that the conclusion is representative of the entire corresponding day.

a. Preliminary analysis

In the 1-hour experiment we used the preliminary analysis to dismiss a large swath of loss functions (those involving the IOU, Dice coefficient, CSI, Peirce score, Heidke score, or Gerrity score) as inappropriate for the task of predicting convection, thus paring down the number of models for the intermediate analysis. Here, in the 2-hour experiment, we use the preliminary analysis to gain a qualitative understanding of the properties of each model, before conducting the fully quantitative intermediate analysis.
Figure S2 shows predictions made by models trained with neighbourhood loss functions, valid at 2200 UTC 2 Jun 2017, the same time step used in the preliminary analysis for the 1-hour experiment. Analyzing the models’ predictions for this day yielded the following key results:

- Some models (panels i, j, l, o) make very poor predictions, because the models did not converge. Specifically, the four models in question reached their minimum validation loss after (1, 2, 2, 2) epochs respectively, despite being trained for 30 epochs thereafter. In the interest of computing time, we left such non-converging models “as is,” instead of training with different random seeds until we obtained a converging model. We tried the latter approach and found that models may need to be trained up to 30 times before convergence. In the interest of fairness, if we trained *some* models 30 times and took the best version, we would need to retrain *all* models 30 times and take the best version. This would involve training $120 \times 30 = 3600$ models, which would take weeks on our computing resources (2000 CPU cores and 800 GPU cores).

- In the 1-hour experiment we did not find non-converging models. We believe that non-convergence is a problem for the 2-hour experiment because the prediction task is more difficult, leading to fewer acceptable local minima.

- The spatially averaged forecast probability generally increases with neighbourhood size, *i.e.*, along each row of Figure S2. We believe that this is because larger neighbourhood sizes more

---

2The loss function depends on all NN parameters (weights and biases), and each U-net in this experiment has 542 053 parameters. In this very high-dimensional space, the loss function has many local minima, some of which are better than others. The same NN with different random seeds – *i.e.*, different initializations of the 542 053 parameters, leading to different starting points in the loss-function space – can reach very different local minima. When a model reaches a poor local minimum early in the training procedure and gets stuck there, we say that the model “did not converge”.
effectively avoid the double penalty, which is a bigger problem at 2-hour lead time than at 1-hour lead time, due to the amount of storm motion that the U-net must implicitly forecast.

- Other than the non-converging models, all models produce reasonable probabilities – i.e., no large areas of near-zero probability that coincide with a lot of convection, no large areas of near-100% probability that coincide with no convection at all.

Figure S3 is analogous to Figure S2 but for models trained with scale-separation loss functions.

Key results are listed below:

- The three models with wavy artifacts in the probability field – F0.1-$\infty^\circ$ FSS (panel j), W0-0.1$^\circ$ FSS (panel k), and W0.1-$\infty^\circ$ FSS (panel l) – are all non-converging. They reached their minimum validation loss after (1, 2, 1) epochs respectively, despite being trained for 30 epochs thereafter.

- Models trained with the Brier score (panels a-h) produce negligible probabilities ($< 0.05$) everywhere; models trained with cross-entropy (panels q-x) produce negligible probabilities almost everywhere; and models trained with the FSS (panels i-p) are the only ones that produce large areas of non-negligible probabilities. These conclusions also apply to scale-separation-based loss functions for the 1-hour experiment (Figure 11 in main body).

b. Intermediate analysis

As in the 1-hour experiment, we use the intermediate analysis to select a handful of models for the final analysis, which consists of in-depth case studies. We use the same procedure, yielding one summary score for each of 120 models on each of the 40 spatial filters used in the verification metrics. Results for the 2-hour lead time are summarized below and in Figure S4.
• For 37 of the 40 filters, the best model includes the FSS, rather than the Brier score or cross-entropy, in the loss function. Thus, as for the 1-hour lead time, the FSS appears to be the most appropriate score for producing a skillful model.

• For the other 3 filters, the best model is trained with 5-by-5 cross-entropy.

• For 30 of the 40 filters, the best model is trained with a neighbourhood-based, rather than scale-separation-based, FSS. The specific neighbourhood sizes are $5 \times 5$ (ranks best for 14 verification filters), $9 \times 9$ (best for 8 filters), and $13 \times 13$ (also ranks best for 8 filters).

• For 7 of the 40 filters, the best model is trained with a scale-separation-based FSS. The specific filters are $F_{0.0-0.8^\circ}$ (ranks best for 5 verification filters), $W_{0.4-\infty^\circ}$ (ranks best for 1 filter), and $F_{0.1-0.2^\circ}$ (also ranks best for 1 filter).

For the final analysis, we select the model trained with pixelwise FSS as a baseline, as in the 1-hour experiment, because pixelwise FSS is not spatially enhanced. We also select the following models to compare against the baseline: those trained with 5-by-5 FSS, 9-by-9 FSS, 13-by-13 FSS, $F_{0.0-0.8^\circ}$ FSS, and 5-by-5 cross-entropy.

c. Final analysis

In this section we analyze the properties of forecasts produced by the six selected models – trained with pixelwise FSS, 5-by-5 FSS, 9-by-9 FSS, 13-by-13 FSS, $F_{0.0-0.8^\circ}$ FSS, and 5-by-5 cross-entropy – on the same case studies as for the 1-hour experiment. However, we do not show the winter case study (analogue to Figure S1), because the results are the same – i.e., all models produce negligible probabilities throughout the domain.

Figure S5 shows two time steps during the passage of Tropical Depression (TD) Luis. As in Figure 13 of the main body, for ease of interpretation, we have labeled areas of strong (W, N) and
weak (Y, E, S) convection at the first time step and two areas of strong convection (A, B) at the second time step. At the first time step (0830 UTC; panels a-h), the model trained with 9-by-9 FSS (panel c) best captures the structure of the convective area, while producing substantially higher probabilities for the strong convection than the weak convection. The model trained with 5-by-5 FSS (panel b) produces higher probabilities for the whole convective area, especially the strong convection, but also produces many false alarms, most notably in the southern third of the domain (south of $\sim 22.5^\circ N$). The models trained with 13-by-13 FSS (panel d) and F0-0.8° FSS (panel e) are similar to that trained with 5-by-5 FSS, producing many false alarms in the southern third, but also produce much lower probabilities for the strong convection. Of the two models heretofore not mentioned, that trained with pixelwise FSS produces near-100% probabilities through most of the domain, rendering it useless. Finally, the model trained with 5-by-5 cross-entropy (panel f) highlights the strong convection and is the only model producing negligible probabilities (<0.05) for the weak convection. Overall, for the first time step (0830 UTC) we judge that the model trained with 5-by-5 cross-entropy would be most appropriate for users with a high risk threshold, while that trained with 9-by-9 FSS would be most appropriate for a low risk threshold.

Results for the second time step during TD Luis (2030 UTC; panels i-p) are qualitatively similar. The model trained with 5-by-5 FSS captures the whole convective area and produces the highest probabilities for both the strong and weak convection (except the pixelwise model, which again produces perversely high probabilities), making it most appropriate for users with a low risk threshold. The model trained with 5-by-5 cross-entropy captures the two highlighted areas of strong convection and ignores the mostly-weak convection southeast of Taiwan, making it most appropriate for a high risk threshold.

Figure S6 shows two time steps on Jun 3 2018, a more typical summer day, featuring convection at various scales smaller than a tropical cyclone. At the first time step (0120 UTC; panels a-h), all
models trained with a spatially enhanced FSS (panels b-e) produce negligible probabilities for the QLCS (labeled “Q”) and non-negligible probabilities for the convective-decay event (labeled “D”), constituting a false negative and false positive respectively. Meanwhile, the model trained with 5-by-5 cross-entropy produces negligible probabilities everywhere with defined target values \(i.e.,\) inside the 100-km range rings). Since all convection in the domain is quite weak, we conclude for this time step that the model trained with 5-by-5 cross-entropy would be best for users with a high risk threshold. For users with a low risk threshold, all models are inappropriate, since that trained with pixelwise FSS produces a large swath of perversely high probabilities and the others produce negligible probabilities throughout the convective area. At the second time step (0720 UTC; panels i-p), the structure of the convective area is captured best by models trained with 5-by-5 FSS (panel j), 9-by-9 FSS (panel k), and F0-0.8° FSS (panel m). The structure is mostly captured by the model trained with 5-by-5 cross-entropy as well, which produces substantially lower probabilities than the other models mentioned. Since all convection in the domain is quite weak, again we conclude that the model trained with 5-by-5 cross-entropy would be most appropriate for users with a high risk threshold. For users with a low risk threshold, we deem that the model trained with 5-by-5 FSS would be most appropriate, since it captures the structure of the convective area as well as any other model but produces higher probabilities.

Lastly, Figure S7 shows the attributes diagram and performance diagram for each final model – using pixelwise verification, which is the default for these diagrams – based on validation data. As discussed in the main body, the attributes (performance) diagram handles probabilistic (deterministic) forecasts. We summarize the quality of the attributes diagram with the positively oriented Brier skill score (BSS) and negatively oriented reliability (REL); we summarize the quality of the performance diagram with the positively oriented area under the curve (AUPD).
In terms of probabilistic forecasts (panels a-f), the best model is that trained with 5-by-5 cross-entropy. Comparing this model to the others, all five BSS differences and all five REL differences are significant at the 95% confidence level, according to a two-sided paired bootstrapping test with 1000 iterations. In the case studies explored above, the model trained with 5-by-5 cross-entropy also appeared to be best for users with a high risk threshold; thus, objective spatial evaluation (Supplemental Section 2b), subjective evaluation (earlier in this section), and objective pixelwise evaluation (Figure S7) highlight this model as one of the most attractive options. In terms of deterministic forecasts (panels g-l), the best model is again that trained with 5-by-5 cross-entropy. However, AUPD differences between this model and the other spatially enhanced models (all except pixelwise FSS) are not significant at the 95% level. Lastly, although the model trained with 5-by-5 FSS appeared in case studies to the best for users with a low risk threshold, it has the worst probability calibration of all models trained with a SELF (panels b-f). The four REL differences and BSS differences are significant at the 95% level. However, it is important to remember that Figure S7 uses pixelwise evaluation. Sometimes pixelwise evaluation and subjective evaluation are complimentary, as for the model trained with 5-by-5 cross-entropy, but sometimes this is not the case. Such disagreements highlight the importance of holistic model verification, including both subjective and objective methods.

References


List of Figures

Fig. S1. Predictions made by final 1-hour U-nets (each trained with a loss function involving a different spatial filter on FSS) for winter case. Formatting (black dots and panel titles) is explained in the captions of Figures 10-11 in the main body. [a-f] Forecast convection probabilities valid at 1230 UTC 25 Jan 2018. [g-h] Composite reflectivity and convection mask valid at 1230 UTC 25 Jan 2018. [i-n] Forecast convection probabilities valid at 2230 UTC 25 Jan 2018. [o-p] Composite reflectivity and convection mask valid at 2230 UTC 25 Jan 2018. .............................................. 15

Fig. S2. Predictions valid at 2200 UTC 2 Jun 2017, made by U-nets trained with different neighbourhood loss functions. Formatting is explained in the caption of Figure 10 in the main body. [a-h] Convection probabilities forecast by U-nets trained to optimize the Brier score with different neighbourhoods. [i-p] Same but for FSS. [q-x] Same but for cross-entropy. [y] Composite (column-maximum) radar reflectivity valid at 2200 UTC. [z] Convection mask valid at 2200 UTC. .............................................. 16

Fig. S3. Predictions valid at 2200 UTC 2 Jun 2017, made by U-nets trained with different scale-separation loss functions. Formatting is explained in the caption of Figure 10 in the main body. [a-h] Convection probabilities forecast by U-nets trained to optimize the Brier score with different spectral filters. For example, the filter “F0-0.1°” uses Fourier decomposition to highlight wavelengths ≤ 0.1°, and “W0.8-∞°” uses wavelet decomposition to highlight wavelengths ≥ 0.8°. [i-p] Same but for FSS. [q-x] Same but for cross-entropy. [y] Composite (column-maximum) radar reflectivity valid at 2200 UTC. [z] Convection mask valid at 2200 UTC. .............................................. 17

Fig. S4. Intermediate analysis. Each panel shows the summary for all 120 models for verification metrics with one spatial filter. Formatting is explained in the caption of Figure 12 in the main body. [a] Summary score on metrics using Fourier decomposition of 0-0.1°. [b-l] Likewise but for metrics with different spatial filters. Only 12 sets of metrics (i.e., metrics involving only 12 of the 40 spatial filters) are shown, for the sake of brevity. .............................................. 18

Fig. S5. Predictions made by final 2-hour U-nets (each trained with a different loss function) for Tropical Depression Luis. Formatting (black dots and panel titles) is explained in the captions of Figures 10-11 in the main body. [a-f] Forecast convection probabilities valid at 0830 UTC 23 Aug 2018. [g-h] Composite reflectivity and convection mask valid at 0830 UTC 23 Aug 2018. [i-n] Forecast convection probabilities valid at 2030 UTC 23 Aug 2018. [o-p] Composite reflectivity and convection mask valid at 2030 UTC 23 Aug 2018. .............................................. 19

Fig. S6. Predictions made by final 2-hour U-nets (each trained with a different loss function) for summer case. Formatting (black dots and panel titles) is explained in the captions of Figures 10-11 in the main body. [a-f] Forecast convection probabilities valid at 0120 UTC 3 Jun 2018. [g-h] Composite reflectivity and convection mask valid at 0120 UTC 3 Jun 2018. [i-n] Forecast convection probabilities valid at 0720 UTC 3 Jun 2018. [o-p] Composite reflectivity and convection mask valid at 0720 UTC 3 Jun 2018. .............................................. 20

Fig. S7. Attributes diagram and performance diagram for each final model. Formatting is explained in the caption of Figure 15 in the main body. [a] Attributes diagram for model trained with pixelwise FSS; [b] 5-by-5 FSS; [c] 9-by-9 FSS; [d] 13-by-13 FSS; [e] F0-0.8° FSS; [f] 5-by-5 cross-entropy. [g-l] Same as a-f but for performance diagrams. .............................................. 21
Figure S1: Predictions made by final 1-hour U-nets (each trained with a loss function involving a different spatial filter on FSS) for winter case. Formatting (black dots and panel titles) is explained in the captions of Figures 10-11 in the main body. [a-f] Forecast convection probabilities valid at 1230 UTC 25 Jan 2018. [g-h] Composite reflectivity and convection mask valid at 1230 UTC 25 Jan 2018. [i-n] Forecast convection probabilities valid at 2230 UTC 25 Jan 2018. [o-p] Composite reflectivity and convection mask valid at 2230 UTC 25 Jan 2018.
Figure S2: Predictions valid at 2200 UTC 2 Jun 2017, made by U-nets trained with different neighbourhood loss functions. Formatting is explained in the caption of Figure 10 in the main body. [a-h] Convection probabilities forecast by U-nets trained to optimize the Brier score with different neighbourhoods. [i-p] Same but for FSS. [q-x] Same but for cross-entropy. [y] Composite (column-maximum) radar reflectivity valid at 2200 UTC. [z] Convection mask valid at 2200 UTC.
Figure S3: Predictions valid at 2200 UTC 2 Jun 2017, made by U-nets trained with different scale-separation loss functions. Formatting is explained in the caption of Figure 10 in the main body. [a-h] Convection probabilities forecast by U-nets trained to optimize the Brier score with different spectral filters. For example, the filter “F0-0.1°” uses Fourier decomposition to highlight wavelengths $\leq 0.1°$, and “W0.8-∞” uses wavelet decomposition to highlight wavelengths $\geq 0.8°$. [i-p] Same but for FSS. [q-x] Same but for cross-entropy. [y] Composite (column-maximum) radar reflectivity valid at 2200 UTC. [z] Convection mask valid at 2200 UTC.
Figure S4: Intermediate analysis. Each panel shows the summary for all 120 models for verification metrics with one spatial filter. Formatting is explained in the caption of Figure 12 in the main body. [a] Summary score on metrics using Fourier decomposition to highlight wavelengths of 0-0.1°. [b-l] Likewise but for metrics with different spatial filters. Only 12 sets of metrics (i.e., metrics involving only 12 of the 40 spatial filters) are shown, for the sake of brevity.
Figure S5: Predictions made by final 2-hour U-nets (each trained with a different loss function) for Tropical Depression Luis. Formatting (black dots and panel titles) is explained in the captions of Figures 10-11 in the main body. [a-f] Forecast convection probabilities valid at 0830 UTC 23 Aug 2018. [g-h] Composite reflectivity and convection mask valid at 0830 UTC 23 Aug 2018. [i-n] Forecast convection probabilities valid at 2030 UTC 23 Aug 2018. [o-p] Composite reflectivity and convection mask valid at 2030 UTC 23 Aug 2018.
Figure S6: Predictions made by final 2-hour U-nets (each trained with a different loss function) for summer case. Formatting (black dots and panel titles) is explained in the captions of Figures 10-11 in the main body. [a-f] Forecast convection probabilities valid at 0120 UTC 3 Jun 2018. [g-h] Composite reflectivity and convection mask valid at 0120 UTC 3 Jun 2018. [i-n] Forecast convection probabilities valid at 0720 UTC 3 Jun 2018. [o-p] Composite reflectivity and convection mask valid at 0720 UTC 3 Jun 2018.
Figure S7: Attributes diagram and performance diagram for each final model. Formatting is explained in the caption of Figure 15 in the main body. [a] Attributes diagram for model trained with pixelwise FSS; [b] 5-by-5 FSS; [c] 9-by-9 FSS; [d] 13-by-13 FSS; [e] F0-0.8° FSS; [f] 5-by-5 cross-entropy. [g-l] Same as a-f but for performance diagrams.