Downscaling Forcing for Lake Garda using Neural Networks: Present & Future

Famke Kovacs
IMAU, Utrecht University, The Netherlands
Correspondence: Famke Kovacs (famkekovacs@hotmail.com)

Abstract.

Deep mixing events (DMEs) are an important aspect of the ecosystem in Lake Garda. To predict when DMEs will happen in the future, especially when considering climate change, a hydrodynamical model is used (such as Delft3D). The input for such models has to be atmospheric forcing with a high resolution. The northern part of Lake Garda lies between high mountain ranges, where high-resolution observational atmospheric data is inaccessible due to sparse observational posts. Therefore, data from the models of the National Centers for Environmental Prediction's Global Forecasting System (NCEP GFS) for present-day, and EC Earth (future climate scenarios model) are needed. These datasets, however, have too low of a resolution compared to the lake size. NCEP has a resolution of 0.25° (17 km in width and 27 km in height), and EC Earth has a 25 km resolution. Both datasets need to be dynamically downscaled, normally with a hydrodynamical model such as the Weather Research and Forecasting (WRF) to 3 km or 4 km resolution. This, however, takes a lot of time; here, neural networks could prove to be very useful. After a long time training the network (just under two weeks), downscaling new data takes significantly less time (a couple of minutes). For the Lake Garda region, two different neural networks (Generative Adversarial Networks or GANs) will be trained using the work of van Rijk (2022), who based his work on Stengel et al. (2020). The first neural network will use training data from 2017 until May 2018, while the second network will use training data until 2020. The test data is from 2021, and all data consists of the following variables: 2 m above surface temperature (T2m) and 10 m above surface wind velocity (U10m and V10m). With the second neural network, a future test dataset will be downscaled from EC Earth 25 km resolution to a 3 km resolution and compared to WRF 4 km. The validation metrics of the generated super-resolution, when compared to the WRF data, are the root mean squared error (RMSE) and the structural similarity index measure (SSIM). Improvements to both SSIM and RMSE for all variables are observed with more training, though they do not improve at the same rate. For future data, both are worse than the present-day data.

1 Introduction

Climate change leads to changes in the thermal structures of lakes, which are led by atmospheric changes (Råman Vinnå et al., 2021). Therefore, it is important to study these atmospheric changes to know what will happen in lakes.

A lake is dependent on deep mixing events for the preservation of its ecosystem. Deep mixing events (DMEs) are occasions where the water in a lake is mixed, meaning the rich nutrients, which are often at the bottom of a lake, will be transported towards the surface. Then oxygen will be transported from the surface towards the bottom. For deep lakes, this is especially crucial, because the regular circulation will not provide this mixing. (Biemond et al., 2021)

In the case of Lake Garda (see Figure 1), these mixing events only happen every few years, meaning this lake is oligomictic. In the work of Biemond et al. (2021), the mechanisms of creating DMEs are explained. This paper concludes that thermocline tilting and turbulent cooling can lead to deep mixing.

Figure 1. Lake Garda 1

To predict DMEs, small-scale atmospheric data is needed to force the hydrodynamic models. Lake Garda is at its smallest part 3 km wide with a length of 30 km, and in the shallow sub-basin a maximum of 13 km wide and 20 km in length. Atmospheric data with a smaller scale is needed to force a hydrodynamical model, such as Delft3D. (Amadori et al., 2021)

Most atmospheric data is available from satellites as observations around Lake Garda are sparse. (Amadori et al., 2021) This is mostly due to the increasingly complex topography (mountains) in the north. However, satellite data do not have the preferred grid resolution to use as a boundary condition for hydrodynamical models.

Data from the National Centers for Environmental Prediction's Global Forecasting System (NCEP GFS, from now on called NCEP) is, for example, on a 0.25° by 0.25° global latitude-longitude grid (National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce, 2015). To downscale this data, the Weather Research and Forecasting (WRF) model can be used. This model uses dynamical methods to downscale data. More information on the WRF model can be found in Giovannini et al. (2014).

However, dynamical downscaling can take a long time because it is computationally expensive. (Glotter et al., 2014) Here, machine learning can play an important role.

Machine learning models can do this downscaling faster than dynamically downscaling models after training the machine learning model. This training of the machine learning models can take some time, but afterward, the downscaling of the desired data can be done significantly faster in contrast to the WRF model. (van Rijk, 2022)

Now, the question arises of how accurate these downscaled data are compared to the WRF data. It would also be useful to investigate whether future data (to predict DMEs) can be successfully downscaled with an existing model trained on present-day

data. It will be the topic of this paper.

Data from NCEP and WRF is utilized to train a machine learning model based on the following paper: Stengel et al. (2020). They used Generative Adversarial Networks (GANs) to downscale global resolution data to high-resolution data (downscaling of factor 50). The atmospheric data that will be downscaled consists of temperature and wind velocity.

The neural network architecture from Stengel et al. (2020) will be applied to Lake Garda to investigate the difference between using downscaling based on the physical relations compared to machine learning based models. It is done by training present data and testing present-day and future data. Preliminary work was done earlier by van Rijk (2022), who used and changed the code of Stengel et al. (2020).

A difference compared to the work of Stengel et al. (2020) is the fact that Stengel et al. (2020) used extrapolated data for their training resolution. So, they use high resolution data and upscale this data for the training of their model. For this paper, the training data is already dynamically downscaled instead because for Lake Garda these datasets are available.

2 Methods & Data

2.1 Training

The method used for data downscaling is based on Single-Image Super Resolution in the machine learning world. In this case, a GAN will be downscaling a low-resolution (LR) picture to a high-resolution (HR) picture. Because atmospheric data can be complex, this downscaling is not done in one step. So, just like in Stengel et al. (2020) and van Rijk (2022), a middle step of downscaling is used, which results in mid-resolution (MR) data. It ensures a deep neural network model, thus leading to a resolution with better small-scale phenomena. A dataset of 9km and 3km resolution is needed for the training part of this model because the original data will be at a resolution of 18km. The first downscaling step of LR to MR has a factor of 2, and the second downscaling step of MR to HR has a factor of 3, see Figure 2.

For temperature and wind velocity, two different GANs are trained. These differ in specific variables for the GANs, such as alpha (a tuning variable for the adversarial part of the network), dropout rate, and number of epochs.

The model training is done on a server of the HPC facility of the Utrecht Bioinformatics Community (UBC). Their GPUs are used, which makes the training and testing of models and data go faster compared to when a CPU is used.

More information about the GANs can be found in Appendix A.

The validation metrics of the generated super-resolution (SR), when compared to the WRF data, are the root mean squared error (RMSE) and the structural similarity index measure (SSIM).

2.2 WRF & NCEP

The WRF dataset is a dynamically downscaled dataset by Giovannini et al. (2014) where the NCEP GFS dataset is used as a boundary condition.

For Section 3.1, the training dataset will be shorter in size than for Section 3.2. For Section 3.1, the temporal range of the

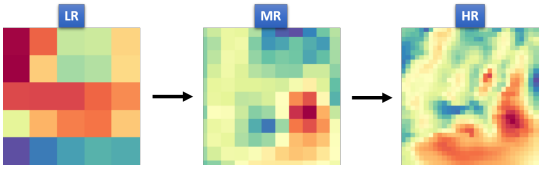


Figure 2. Process of downscaling. The domain is 45.2917178-46.108105°N and 10.100739-11.273682°W. Number of pixels for LR = 5x5, MR = 10x10 and HR = 30x30.

Source	NCEP GFS	EC Earth	WRF	WRF	WRF
Spatial Resolution	0.25°	25km	9km	4km	3km
Temporal Resolution	3 hours	3 hours	3hours	3 hours	3hours
Training Size	11688	-	11668	-	11668
Test Size	2920	2920	2920	2920	2920

Table 1. Specifics of the datasets for the variables for Section 3.2 and 3.3

training data is from 1-1-2017 until 31-5-2018, while for Section 3.2 the training data ranges from 1-1-2017 until 31-12-2020. In Table 1 the characteristics of the datasets of NCEP GFS and WRF can be found. The difference between this dataset and the one used for Section 3.1 is that the training size is smaller (4128). There is also a difference in the boundaries of the WRF datasets and the spatial resolution for the standard test case and the more time case. The domain for the training with more years lies between 45.2917178-46.108105 degrees North and 10.100739-11.273682 degrees West. For the standard test case, the area of downscaling is the same as for van Rijk (2022), with a pixel spatial resolution of 96x96 (HR), 32x32 (MR), and 16x16 (LR). For the case of training with more years, the pixel spatial resolutions are 30x30 (HR), 10x10 (MR), and 5x5 (LR). The variables of all these datasets are the 2 meters above the surface air temperature (T2m in Kelvin (K)) and the 10 meters above the surface wind velocity components (U10m and V10m both in meters per second (m/s)).

The data from NCEP GFS can be downloaded from the National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce (2015) website. The dataset of NCEP has the boundaries as van Rijk (2022).

2.3 Future Data

For the future data test case, with the same variables as before (T2m, U10m, V10m), data from the EC Earth model is used as input (LR data). The work of Davini et al. (2017) explains more about the EC Earth model. The validation test data is WRF data, but this time at a 4km scale. The WRF 4km data domain is 10.437653 - 10.961456°W and 45.419144 - 45.998722°N. The WRF 4km has a pixel resolution of 17x11. The super-resolution will be interpolated towards a 4km resolution, while the super-resolution itself will be at a 3km resolution after the downscaling of the EC Earth data. The future year data that will be explored is 2041. The climate scenario under which the EC Earth data was created is RCP8.5 (Smolders, 2022). The grid of the EC Earth data set is 44-48°N and 8-14°W with 3-hour temporal resolution. The spatial resolution is 25km. Other specifics

can be found in Table 1. Napoli et al. (2022) dynamically downscaled the EC Earth data to 4km using WRF, which is used here as validation for the future test case.

3 Results

The results are divided into 3 sections: a standard test case reproducing the work of van Rijk (2022); a case with more training data; and a case where the model of more training (3.6 years extra compared to the standard test case) data is used to test a future dataset.

3.1 Standard test case

In the standard test case, the following results are displaying for each variable: the RMSE and SSIM for the training and test data in space and time, and the best and worst sSR of the test data. For this case, the training data is from 2017 until May 2018, and the test data is from 2021. The results contain only the last step of downscaling (from MR to HR).

3.1.1 T2m

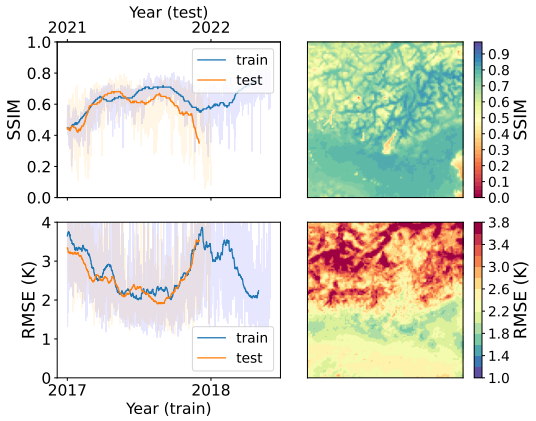
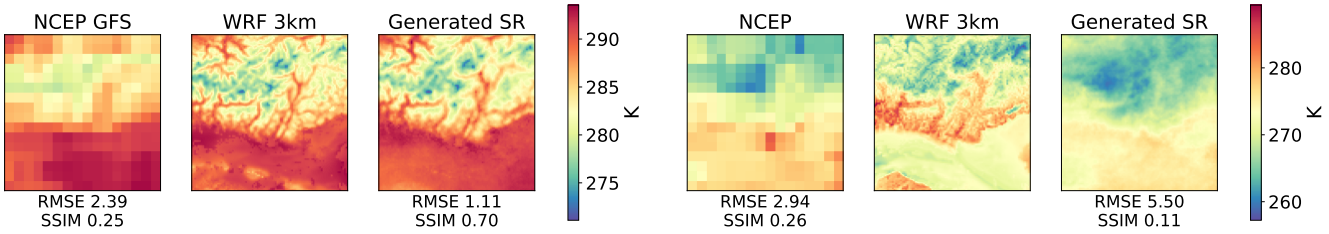


Figure 3. SSIM (upper two figures) and RMSE (lower two figures) for the T2m. The two figures on the left display the SSIM and RSME over time for the training and test data. The two figures on the right are the SSIM and RMSE averaged throughout the test data over the area of interest.

In Figure 3, the SSIM and RMSE are presented for the variable T2m of the generated SR.
The SSIM over time, follows the same pattern for the test and train SR data. The values are often above 0.5, meaning the test and train SR data have structurally similar patterns as the WRF 3km of T2m. Over space, the SSIM has some lower values over Lake Garda (in the middle of the upper right figure), the values are higher in the south where the topography becomes flat. In the north, where the topography is more present/rough, the SSIM seems to switch a lot between high and slightly lower values. The RSME over time is higher at the beginning and end of the year for both test and training SR. The RMSE never dips below

the 2 K mark. Over space, Lake Garda is not as visible as it was for the SSIM. The lowest values of RMSE are in the southern part of the figure, and in the nothern part, the RMSE reaches almost 4 K. In the northern part of the figure, the patterns of RMSE become more complex with detailed structures of valleys and tops of high and low values of RMSE. This may be attributed to the rough topography in the north, however.



(a) Best SR: Snapshot is at 06:00 17-11-2023

(b) Worst SR: Snapshot is at 15:00 23-05-2021

Figure 4. The best and worst SR for T2m. Training period = 2017 - May 2018, testing period = 2021.

In Figure 4a, the best downscaling SR for T2m is displayed over the research area alongside the corresponding NCEP and WRF 3km data. The RMSE is 1.1 K. The SSIM is 0.70, positive, meaning the patterns are one-to-one and not the opposite. In Figure 4b, the worst SR for T2m is displayed together with the corresponding NCEP and WRF 3km data; it is remarkable to see that overall the patterns of SSIM of the SR look vaguely familiar to the NCEP original data, even with almost the same SSIM.

3.1.2 U10m

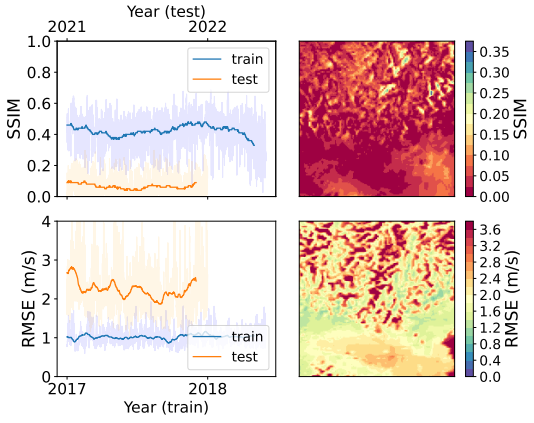
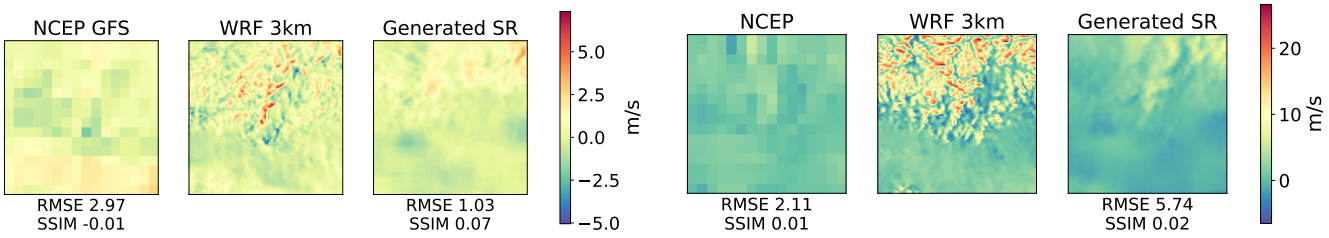


Figure 5. SSIM (upper two figures) and RMSE (lower two figures) for the U10m. The two figures on the left display the SSIM and RSME over time for the training and test data. The two figures on the right display the SSIM and RMSE averaged throughout the test data over the area of interest.

In Figure 5, the SSIM and RMSE are presented for the variable U10m of the generated SR.

The SSIM over time, does not seem to have a pattern. The train SR is quite a bit higher than the test SR. The train SR has an SSIM varying around 0.4, and the test SR is around 0.1. The training SR data is more similar to the WRF 3km data than the test SR. It can be said that the test SR has an SSIM around 0.1, close to zero, which means it is entirely structurally dissimilar from the WRF 3km data. Over space, the SSIM is very low overall, with a max of around 0.3 in the northern region.

The RMSE over time also seems constant, with the training SR being lower (around the value of 1 m/s) than the test SR, which varies between 2 and 3 m/s. Over space, the RMSE is overall high, around the value of 3 m/s. In the left and right side of the middle of the figure the RMSE drops around 1.2 m/s. In the northern part of the figure, the RMSE has more complex patterns because of the different topography in that region compared to the south.



(a) Best SR: Snapshot is at 06:00 12-06-2023

(b) Worst SR: Snapshot is at 18:00 28-01-2021

Figure 6. The best and worst SR for U10m. Training period = 2017 - May 2018, testing period = 2021.

In Figure 6a, the best downscaling SR of the test set is displayed for the 10m-U-component of the wind together alongside the NCEP and WRF 3km data. The RMSE of the SR is even less than the original data of NCEP. The RMSE is just around 1, so there is little error. The SSIM for the SR is around zero. There is total structural dissimilarity.

In Figure 6b, the worst downscaling SR of the test set is displayed for the 10m-U-component of the wind alongside the NCEP and WRF 3km data. Again, just like the worst T2m, the SR looks more like the NCEP than the WRF because the WRF 3km looks very different on the upper side of the grid. Therefore the RSME of the SR is again high, but the SSIM is similar to the NCEP.

3.1.3 V10m

In Figure 7, the SSIM and RMSE are presented for the variable V10m of the generated SR.

The SSIM over time for both train and test SR looks constant. The train SSIM is around 0.4, and the test SSIM is around the value of 0.1, which is low, which means it is almost structurally dissimilar. Over space, the SSIM behavior of the previous subfigure is confirmed. Over the whole subfigure, the SSIM varies between 0 and 0.10, with a few spots of 0.20. Again, the patterns are in the north more complex than in the south.

The RMSE over time for the train SR is more constant around the value of 1 m/s, while the RMSE for the test SR seems to display a more oscillatory behavior, varying between 2 and 3 m/s. Over space, the RMSE is low in the south, around 1.2 m/s.

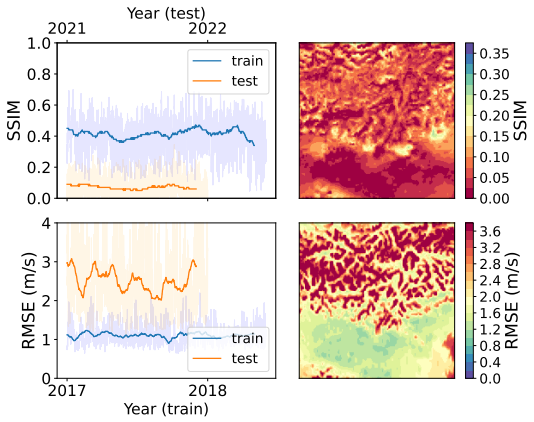
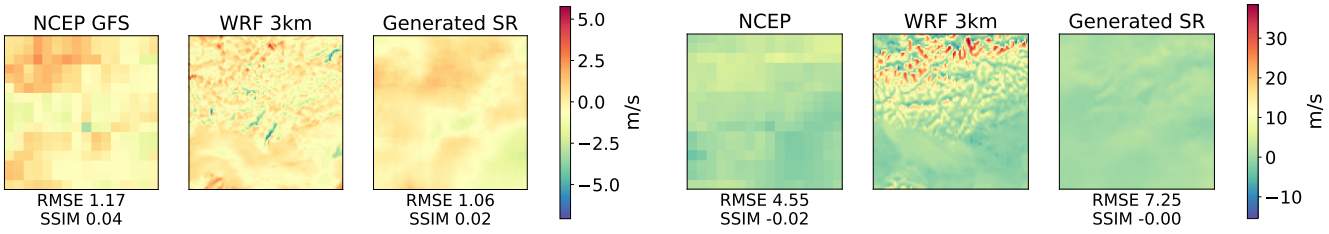


Figure 7. SSIM (upper two figures) and RMSE (lower two figures) for the V10m. The two figures on the left display the SSIM and RSME over time for the training and test data. The two figures on the right display the SSIM and RMSE averaged throughout the test data over the area of interest.

In the north, it has a more complex structure with gradients between RMSE of 3.6 and 2 m/s. There seems to be a noticeable difference between the patterns in the north of RMSE and compared to the SSIM.



(a) Best SR: Snapshot is at 06:00 03-06-2021.

(b) Worst SR: Snapshot is at 21:00 10-05-2021.

Figure 8. The best and worst SR for V10m. Training period = 2017 - May 2018, testing period = 2021.

In Figure 8a, the best downscaling of the test set is displayed for the 10m-V component of the wind alongside NCEP and WRF 3km data. As for the U-component and T2m variables, for the best SR, the RMSE is smaller than the original NCEP. The SSIM of both NCEP and SR are very similar. In the SR, some blue spots are missed compared to the WRF data, indicating the southward wind spots do not generate well.

In Figure 8b, the worst downscaling of the test set is displayed for the 10m-V component of the wind alongside the corresponding NCEP and WRF 3km data. The V-component of SR has the highest RMSE of all the variables for the test data. This behavior is also seen in the RMSE of the original NCEP data. Again, the patterns in the northern parts of the WRF data are not generated in the SR, looking more like the NCEP than the WRF.

3.2 Training with more years

In the training with more years (3.6 years) case, the following results will be displayed for each variable: the RMSE and SSIM for the training and test data in space and over time, and the best and worst SR of the test data. For this case, the training data is from 2017 until 2020, and the test data is from 2021. The results contain only the last step of downscaling (from MR to HR).

3.2.1 T2m

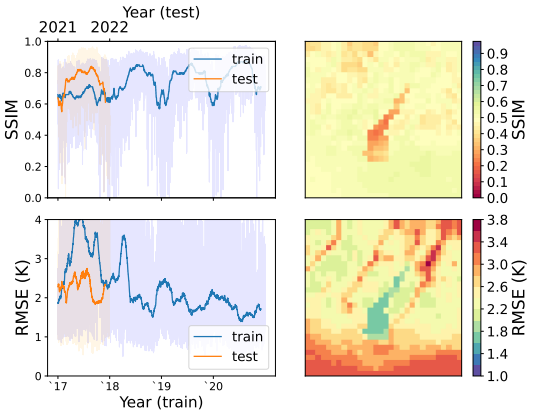


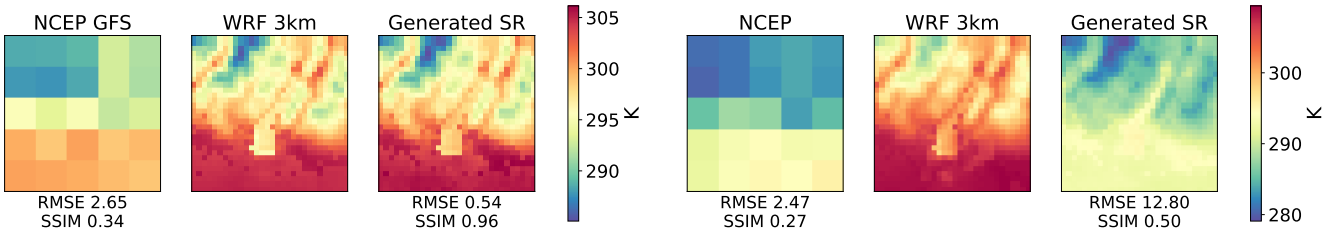
Figure 9. SSIM (upper two figures) and RMSE (lower two figures) for the T2m. The two figures on the left display the SSIM and RSME over time for the training and test data. The two figures on the right display the SSIM and RMSE averaged throughout the test data over the area of interest.

In Figure 9, the SSIM and RMSE are presented for the variable T2m.

The SSIM in time oscillates for both test and train SR data between the values of 0.6 and 0.8. The closer value to 1, the more identical the generated SR is to the WRF data. Over space, the SSIM is low around the Lake Garda region (in the middle of the subfigure, contours of the outlining of the lake are visible), indicating a dissimilar structure around there. Overall, the SSIM seems to be around 0.7, with a few spots in the north indicating a lower SSIM.

The RMSE over time for the training SR data starts high and then lowers, but in an oscillating way. The testing SR data RMSE varies between 2 and 2.5 K. Over space, the RMSE of the testing SR is the lowest over Lake Garda, around a value of 1.6 K. In the south, the RMSE is higher, around 3.2 K. In the north emerges some interesting patterns of higher RMSE aligning diagonally with the value of around 3.2 K, as surrounding areas have a RMSE of 2 K.

In Figure 10a, the best SR test result is displayed alongside the corresponding NCEP and WRF 3km data of the variable. This SR has a RMSE of 0.54 K, which lies close to 0. The SSIM is 0.96, meaning the image is almost identical to the WRF of 3km. The main difference between the generated SR and WRF 3km of the T2m variable is that some areas seem darker, thus more red or blue, compared to WRF 3km. This behavior is seen in the blue area in the left corner, in the lower half of the figure in a red region, and in the middle of the subfigure where approximately Lake Garda lies in the more orange area. Overall, the SR



(a) Best SR: Snapshot is at 12:00 27-06-2021.

(b) Worst SR: Snapshot is at 15:00 16-08-2021.

Figure 10. The best and worst SR for T2m. Training period = 2017 - 2020, testing period = 2021.

does seem to generate the structure very well, with a small error. The warm south has a colder water region depicting Lake Garda and a cold north with more topography of the mountain ranges, leaving vertical lines of warmer and colder temperatures. In Figure 10b, the worst SR is displayed for the variable T2m alongside the corresponding NCEP and WRF 3km data. The RMSE is 12.80 K of the SR, a big difference from the best T2m snapshot. The SSIM is 0.50. It is between structurally correct and structurally dissimilar. Indeed, structure-wise, the same patterns of the WRF are seen in the generated SR: warmer south with a little less warm Lake Garda and cold diagonal strokes in the North with more topography. The RMSE behavior is seen in the generated SR, displaying everywhere colder colors. Comparing the SR with the NCEP however, they seem to be closer in patterns.

3.2.2 U10m

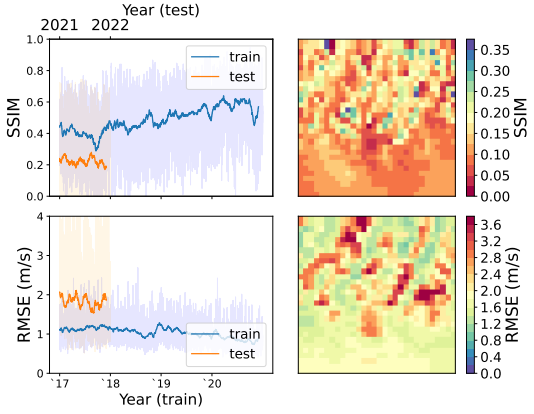
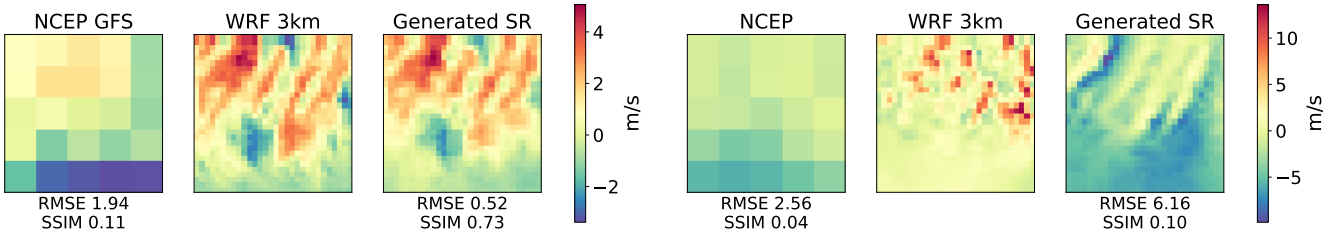


Figure 11. SSIM (upper two figures) and RMSE (lower two figures) for the U10m. The two figures on the left display the SSIM and RSME over time for the training and test data. The two figures on the right display the SSIM and RMSE averaged throughout the test data over the area of interest.

In Figure 11, the SSIM and RMSE are presented for the variable U10m.

The SSIM over time for the training SR data is slightly increasing with some relatively small variations, from a value of 0.4 to almost 0.6. The SSIM of the test SR is lower, varying around 0.2. It means the test data is structurally less similar to the train SR when compared to the WRF data. Over space, the SSIM of the test SR is overall low, and the spots of higher SSIM seem random. In the north, one can find (seemingly random) patches of relatively high values of SSIM.

The RMSE over time for the training SR varies around the value of 1 m/s, while for the test SR, it oscillates around the value of 2 m/s. Over space, the RMSE of the test SR is overall low in the south region of the figure. In the north, high and low values of RMSE are trading off spaces next to each other. These patterns are mostly diagonally. The highest value of RMSE is around 3.6 m/s, and the lowest is around 0.8 m/s.



(a) Best SR: Snapshot is at 12:00 27-06-2021.

(b) Worst SR: Snapshot is at 9:00 22-01-2021.

Figure 12. The best and worst SR for U10m. Training period = 2017 - 2020, testing period = 2021.

In Figure 12a, the best SR for the variable U10m is displayed alongside the corresponding NCEP and WRF 3km data. The RMSE of the SR is 0.52 m/s. The SSIM is 0.73, which is again close to 1 (structurally the same). It is visible in the figure: both colors and structure are very similar. Looking at the pattern, the contour lines going diagonally also appear, just like in the WRF data, with the data in the north sometimes being wider, compared to the WRF data.

In Figure 12b, the worst case of SR for the variable of U10m is displayed alongside the corresponding NCEP and WRF 3km data. The RMSE is 6.16 m/s, and the SSIM is 0.10 of the SR, meaning almost complete structural dissimilarity. It can be seen in the right-most figure of Figure 12b. There is a discrepancy between the SR and WRF, meaning the wind direction is often not similar, especially in the south. In the northern part of the figure, the values are often too low, and the wind direction is better. However, in the WRF data (middle figure), there are visible spots of higher positive values. In the SR data, the spots in the north are more strokes going diagonally.

3.2.3 V10m

In Figure 13, the SSIM and RMSE are presented for the variable V10m.

The SSIM over time for the training SR data oscillates around the value of 0.4 at the beginning of the year, and later steadily increases towards a value of 0.6 with some dips. The SSIM of the test SR oscillates between 0.2 and 0.3, which is quite a bit lower than the training SR. For the SSIM over space, the higher values are placed in the northern parts, while the lower values

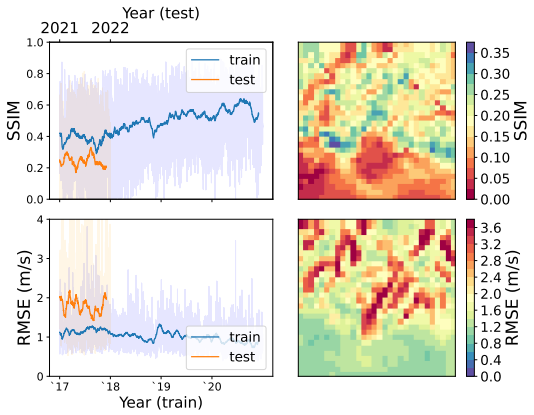
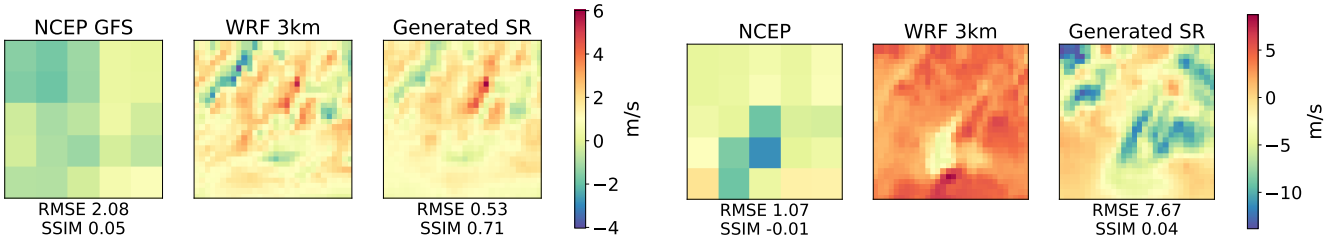


Figure 13. SSIM (upper two figures) and RMSE (lower two figures) for the V10m. The two figures on the left display the SSIM and RSME over time for the training and test data. The two figures on the right display the SSIM and RMSE averaged throughout the test data over the area of interest.

are in the south. In the north, the value of SSIM is the highest around the middle. The value of the test SR is low, meaning the images of the SR are not identical and are more structurally dissimilar compared to the WRF data.

The RMSE over time for the training SR, oscillates a little around the value of 1 m/s. The test SR RMSE oscillates between 1.5 and 2 m/s approximately. Over space, the RMSE of the test SR has relatively low values in the south, while the northern part has some outliers around 3.4 m/s. These patterns are mostly diagonal, and some outliers are in the Lake Garda region.



(a) Best SR: Snapshot is at 12:00 02-03-2021 .

(b) Worst SR: Snapshot is at 16:00 16-06-2021.

Figure 14. The best and worst SR for U10m. Training period = 2017 - 2020, testing period = 2021.

In Figure 14a, the best case of SR for the variable V10m is displayed alongside the corresponding NCEP and WRF 3km data. The RMSE of the SR is 0.53 m/s, and its SSIM is 0.71. The SSIM is again close to structurally correct compared to the WRF data. It is visible in the SR data figure. The same structures of diagonal strokes of first negative and then positive V10m can be seen in the north, except the values have little deviations compared to the WRF data.

In Figure 14b, the worst case of the SR for the variable V10m is displayed alongside the corresponding NCEP and WRF 3km data. The RMSE is 7.76, and the SSIM is 0.04 of the SR. The latter is almost zero, meaning there is next to no structural

similarity with the WRF data. Comparing the two pictures of WRF and SR, for this case structurally (SSIM) and looking at the actual values of V10m (RMSE), this case is very wrong. There are a lot of negative spots in the SR picture, and WRF stayed positive everywhere. The dots and strokes of WRF are not visible in the SR.

3.3 Future forcing

In the future forcing case, the following results will be displayed for each variable: the RMSE and SSIM for the training and test data in space and over time, and the best and worst SR of the test data. For this case, the training data is from 2017 until 2020, and the test data is from 2041. The results will contain the last step of downscaling (from MR to HR).

3.3.1 T2m

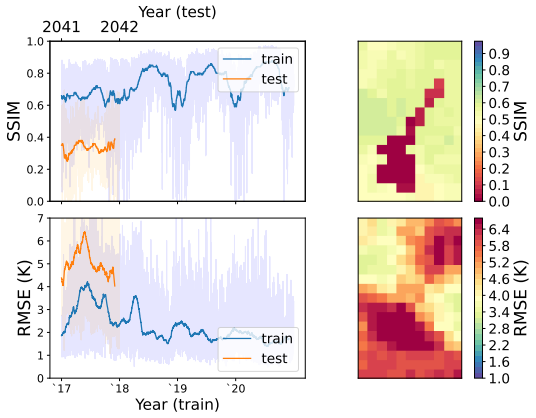
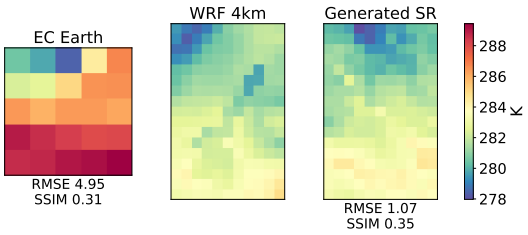


Figure 15. Upper figure: SSIM of the variable T2m through time with average over space with training data of 2017-2020 and test data of 2041 (left) and SSIM averaged over time of the same test data plotted in the area of interest (right).
Lower figure: RMSE of the variable T2m through time with average over space with training data of 2017-2020 and test data of 2041 (left) and RMSE averaged over time of the test data in 2041 plotted in the area of interest (right).

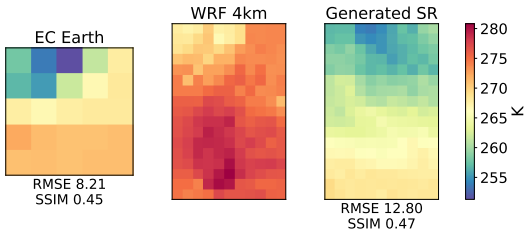
In Figure 15, the SSIM and RMSE for the variable T2m are displayed. The training data in the left figures is the same as from Section 3.2.1.

The SSIM over time of the test SR is lower than the train SR. Values of the test SR vary between 0.2 and 0.4 approximately. Over space, the SSIM of the test SR has the lowest values over Lake Garda, in the middle of the figure. The values there are almost zero. Outside the Lake Garda region, the SSIM values are high, around 0.6.

The RMSE over time of the test SR is higher than the train SR. The RMSE climbs from 4 K to roughly 6.5 K and then decreases again towards 4 K. For the test SR, the RMSE value is high over space, with a minimum value of around 3.4 K, and the highest value of RMSE is approximately 6.4 K. The south of the figure has a high RMSE, and the northeast also has higher values. The northwest and middle eastern parts of the figure have the lowest values of RMSE.



(a) Best SR: Snapshot is at 9:00 11-02-2041.



(b) Worst SR: Snapshot is at 3:00 01-12-2041.

Figure 16. The best and worst SR for T2m. Training period = 2017 - 2020, testing period = 2041.

In Figure 16a, the best case of SR for the variable T2m is displayed alongside the corresponding EC Earth and WRF 4km data. The RMSE is 1.07 K, and the SSIM is 0.35 of the SR. The SSIM is low. It is visible in the figure of the generated SR. Although the SR and WRF have a northeastern cold spot, their locations do not match. Also, in the south of the SR, there is a warm region, which spreads out from west to east. This region can not be found in the WRF data with such warm temperatures.

In Figure 16b, the worst case of SR for the variable T2m is displayed alongside the corresponding EC Earth and WRF 4km data. The RMSE is 12.80 K, and the SSIM is 0.47 of the SR. Interestingly, despite the relatively high SSIM (even higher than the SSIM for the best SR), its RMSE value is incredibly high. It leads to an interesting discrepancy in T2m values, though the overall structure remains similar.

3.3.2 U10m

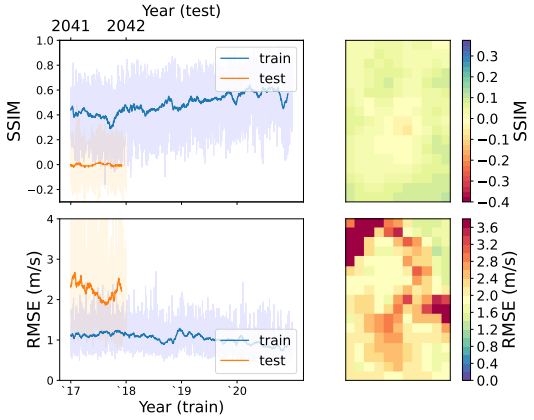


Figure 17. Upper figure: SSIM of the variable U10m through time with average over space with training data of 2017-2020 and test data of 2041 (left) and SSIM averaged over time of the test data of the year 2041 plotted in the area of interest (right).
Lower figure: RMSE of the variable U10m through time with average over space with training data of 2017-2020 and test data of 2041 (left) and RMSE averaged over time of the test data of the year 2041 plotted in the area of interest (right).

In Figure 17, the SSIM and RMSE for the variable U10m are displayed. The training data in the left figures is the same as from Section 3.2.2.

The SSIM over time for the test SR is lower than the train SR. The SSIM for the test oscillates around zero, meaning it is structurally dissimilar to the WRF 4km data. Over space, the SSIM for the test SR is also very low, achieving its lowest values near the center of the figure at around -0.1 to -0.2, whereas the overall highest value is around 0.15.

The RMSE over time for the test SR varies between 2 and 2.5 m/s, with the highest RMSE in the beginning and end times. The lowest RMSE is around two-thirds of the year (2041). Compared to the train SR data, the RMSE is double. Over space, the RMSE of the test SR is the highest in the northwest corner and the lowest in the northeast corner. Also, high values of RMSE are in the east. A vague outline of the lake can be seen, with an RMSE of around 3 m/s.

(a) Best SR: Snapshot is at 15:00 03-11-2041.

(b) Worst SR: Snapshot is at 6:00 06-02-2041.

Figure 18. The best and worst SR for U10m. Training period = 2017 - 2020, testing period = 2041.

In Figure 18a, the best case of SR for the variable U10m is displayed alongside the corresponding EC Earth and WRF 4km data. The RMSE is 0.79 m/s, and the SSIM is 0.13 of the SR. With the SSIM being low, it is visible in the figure structure-wise that the SR and WRF are not the same. The WRF 4km has positive values in the northwestern, southwestern, and eastern parts of the figure. For the SR, this only happens in the east part of the figure, and it is spreading to the northwest from there. Also, some spots with negative values are not taking over into the SR data.

In Figure 18b, the worst case of SR for the variable U10m is displayed alongside the corresponding EC Earth and WRF 4km data. The RMSE is 6.77 m/s, and SSIM is 0.01. The SSIM is so low that the SR and WRF are structurally dissimilar, which is visible in the figure. The WRF has positive values exclusively in the north- and southwest, while the SR has positive values almost everywhere. The negative values of WRF in the middle between the two spots, in the northwest and southeast of the figure, can not be found in the SR image.

3.3.3 V10m

In Figure 19, the SSIM and RMSE for the variable V10m are displayed. The training data in the left figures is the same as from Section 3.2.3.

The SSIM over time of the test SR is low compared to the train SR and varies around 0, meaning structurally, the SR is dissimilar from its WRF counterpart. Compared to the train SR, there is a big difference. Spatially, the SSIM of the test SR is

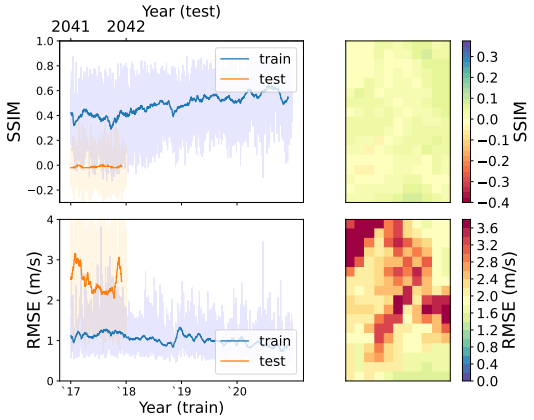
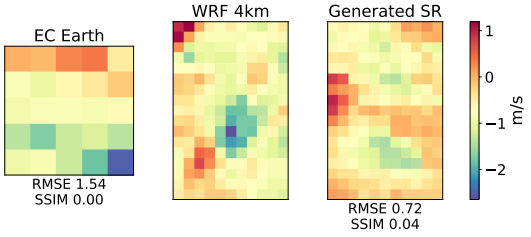


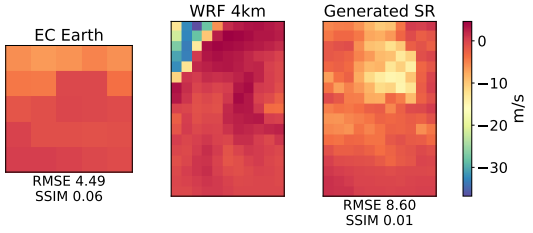
Figure 19. Upper figure: SSIM of the variable V10m through time with average over space with training data of 2017-2020 and test data of 2041 (left) and SSIM averaged over time of the test data of the year 2041 plotted in the area of interest (right).
Lower figure: RMSE of the variable V10m through time with average over space with training data of 2017-2020 and test data of 2041 (left) and RMSE averaged over time of the test data of the year 2041 plotted in the area of interest (right).

also zero, with a few exceptions.

The RMSE over time of the test SR varies between 2.5 and 3 m/s. The highest values are at the beginning and end of the year 2041. Compared to the train SR, the test SR SSIM has almost tripled. Over space, the RMSE of the test SR has its highest values in the northwest and a diagonal stroke over the middle with a small spot at the right in the middle. These values seem to be around 3.6 m/s, while the lowest values of approximately 1.2 m/s are located in the southeast and northeast with some spots.



(a) Best SR: Snapshot is at 21:00 03-08-2041.



(b) Worst SR: Snapshot is at 3:00 26-11-2041.

Figure 20. The best and worst SR for V10m. Training period = 2017 - 2020, testing period = 2041.

In Figure 20a, the worst case of SR for the variable V10m is displayed alongside the corresponding EC Earth and WRF 4km data. The RMSE is 0.72 m/s, and the SSIM is 0.04 for the test SR. The values of the V10m are not that high, leading to relatively small errors compared to the WRF 4km data. However, the SSIM indicates that structures and patterns are wrong because a value of 0.04 is extremely close to zero. The negative values are approximately in the right spot for SR, around the

middle of the figure. The higher positive values of the WRF in the northwest and roughly southwest are not visible in the SR. Overall, the SR has a lot more positive values than WRF.

In Figure 20b, the worst case of SR for the variable V10m is displayed alongside the corresponding EC Earth and WRF 4km data. The RMSE is 8.60 m/s, and the SSIM is 0.01 for the test SR. Structure-wise, this result is even worse than for the best case of V10m. The RMSE is relatively high. However, high values of V10m are present, especially negative, in the WRF data. A spot in the northwest corner is present, with values towards the -30 m/s. This spot is not replicated in the SR figure. The SR has a generally high negative value spot in the middle of the figure. Meanwhile in the WRF data, the lowest values are around 0 m/s.

4 Discussion

In this section, for every variable, the results of Section 3 will be discussed, comparing all three cases to one another. At the end, a summary is available with a few remarks.

4.1 T2m

Comparing the SSIM of Figure 3 to Figure 9 (upper half of both figures), it is visible that for the standard test (ST) case the SSIM over time it is lower than for the more time (MT) case. It holds for both test and train SR. For the MT SSIM, it is the case that the test SR is even higher than the train SR. It does not happen for ST, where the two SRs are in the same range. For the SSIM over space, ST and MT develop the same problem: over the Lake Garda region, the SSIM is the lowest. It is the region of interest, and these values of T2m compute the hydrodynamics of the Lake together with some other variables. Structurally, the patterns are not identical to the WRF data, which models the hydrodynamics.

Comparing the RMSE of Figure 3 to Figure 9 (lower half of both figures), it is visible that for the train SR of ST the RMSE over time only has two peaks of around 4 K at the beginning and end of the training period. For MT, the train SR starts around the same RMSE of 4 K and then decreases over time. The peak of training SR of MT is higher than that of ST. Meanwhile, the test SR of ST follows the same RMSE over time as its training part. For MT, the test SR RMSE is lower. For ST, it varies between 2 and 4 K, while for MT, it varies between 2 and 2.5 K. Spatially, the RMSE of ST is divided over two sections down the middle horizontally, with a high RMSE in the north and a lower RMSE in the south. For MT, this division is not visible. Above and below Lake Garda, the RMSE is higher than at the Lake Garda region itself. This lower RMSE over Lake Garda seem also to happen in the ST.

Further, comparing the SSIM of Figure 9 to Figure 15 (upper half of both figures), the test SR of MT is a lot higher than for the future forcing (FF) case over time. For ST, the SSIM varies between 0.6 and 0.8. For FF, it varies between 0.2 and 0.4. It leads to a less structurally similar SR compared to the WRF data for FF. For the SSIM over space, for both MT and FF, the same behavior happens: Lake Garda has a low SSIM of around 0, and the region around it has an SSIM of around the value of 0.7.

Comparing the RMSE of Figure 9 to Figure 15 (lower half of both figures) over time, the RMSE of MT for the test SR is lower than for FF. For MT, it varies between 2 and 3 K. For FF, it goes up and down from 4 to 6.5 K. Spatially, the RMSE of MT differs much from FF. Over the Lake Garda region, the RMSE of MT is low, around 1.8 K. For FF, it is high everywhere, with some peaks of 6.8 K, even over Lake Garda.

Thus, for T2m, the SSIM improves over time using more training data compared to the standard case, but for future data, the SSIM over time is low. The SSIM over space for all three cases is the lowest over Lake Garda.
Furthermore, for the T2m, the RMSE of the test SR improves using more training data. However, the RMSE of the future data becomes significantly higher. The RMSE over space becomes smaller in the Lake Garfa region after using more training data for the training. This effect is, however, not visible for the future data with high RMSE over that region.

4.2 U10m

Comparing the SSIM of Figure 5 to Figure 11 (upper half of both figures), the train SR has the same range of values for ST and MT over time. Later in the training SR of MT, the SSIM is improving with a higher SSIM than ST. The test SR of ST is lower than that of MT. For both ST and MT, the training SR has the better SSIM. It means the training SR has a higher similarity compared to the WRF data than the test SR. The test SR is structurally dissimilar compared to WRF for ST and close to that structural dissimilarity for MT. The SSIM over space has overall higher values of MT than ST. MT does have a more random pattern, while the SSIM of ST looks more smoothly with some spots of higher SSIM. It is the case because ST has spatially more points than MT. The fact that MT has more training data points in time leads to some improvement in SSIM.
Comparing the RMSE of Figure 5 to Figure 11 (lower half of both figures), both ST and MT test SR is higher than the training SR. The training SR of ST and MT are around the same values of 1 m/s, while the RMSE of the test SR of ST is slightly higher and more varying than for MT. The RMSE over space does have some similarities for ST and MT. In both cases, the RMSE is higher in the northern part and lower in the south. The maximum RMSE is around the value of 3.6 m/s. The area of higher RMSE is higher for the ST.

Comparing the SSIM of Figure 11 to Figure 17 (upper half of both figures), the SSIM over time for FF is lower than for MT. The SSIM of MT over time is around 0.2 while for FF it is around 0 for the test SR. The SSIM over space shows still a few higher spots of SSIM for MT (0.35). For FF, it is mostly around 0.1 and even in the middle negative -0.1.
Comparing the RMSE of Figure 11 to Figure 17 (lower half of both figures), the RMSE over time of the test SR is higher for FF than MT. Where MT has an upper limit of 2 m/s, for FF, that limit is 2.5 m/s, and the lower limit is 2 m/s. Over space, the RMSE is only high in the northern part of the figure for MT, while for FF, this is the case everywhere. For Lake Garda, the error is around the same for both MT and FF.

Thus, for U10m, the SSIM of the test SR is improving over time when using more training data to train the network. However, for future data, the SSIM has decreased again towards SR images being structurally dissimilar from its WRF counterpart. The

SSIM over space seems not to improve over Lake Garda with more training data, and for future data, it is even so low the SSIM becomes negative.

Furthermore, for U10m, the RMSE over time is improving using more training data, but the RMSE is again higher for future data. The RMSE over space around Lake Garda is still after using more training data high. The same happens for future data.

4.3 V10m

Comparing the SSIM of Figure 7 to Figure 13 (upper half of both figures), the SSIM over time for MT is higher than for ST in both cases of training and test SR. The error in SSIM (the lighter blue shaded area) is wider for MT than for ST. Over time, the SSIM for MT increases the whole time. For ST, it stays around the same value. The SSIM over space for ST is low everywhere, whereas, for MT, there are some spots of slightly higher SSIM values (the blue dots) in the south part of Lake Garda. However, MT still has the same SSIM as ST.

Comparing the RMSE of Figure 7 to Figure 13 (lower half of both figures), the RMSE over time of the test SR is lower for MT than ST. For MT, the RMSE varies between 1.5 and 2 m/s, while for ST, it varies between 2 and 3 m/s. For the training SR is no difference from ST and MT. The RMSE over space of ST and MT are largely similar: a high RMSE in the north and a low RMSE in the south. The total area of higher RMSE is for ST larger than for MT, leading to an overall higher RMSE average.

Comparing the SSIM of Figure 13 to Figure 19 (upper half of both figures), the SSIM over time of the test SR is lower for FF than for MT. For FF, it varies around 0, while for MT, it varies around 0.2. The error of FF is smaller than for MT. The SSIM over space is quite different from MT compared to FF. For MT, the SSIM is higher in the north than in the south. There it is around 0. For FF, the SSIM is overall around zero, except in some spots where it is negative.

Comparing the RMSE of Figure 13 to Figure 19 (lower half of both figures), the RMSE over time is again higher for FF than for MT. The RMSE of FF varies between 2.5 and 3 m/s, while the RMSE of MT varies between 1.5 and 2 m/s. The RMSE over space for both MT and FF has the area of highest RMSE around Lake Garda. For MT, the RMSE is also high in the northern parts of the figure, while in the south, the RMSE is lower.

Thus, for V10m, the SSIM over time of the test SR is improving using more training data. The SSIM over time of future data, is again lower. It leads to more SR images being structurally dissimilar to their WRF counterpart. The SSIM over space has some improvement in the northern part of Lake Garda using more training data. However, for future data, everywhere over Lake Garda, the SSIM is low.

Furthermore, for V10m, the RMSE over time of the test SR is improving using more training data. However, for future data, this is not the case. The RMSE over space, especially over Lake Garda, is high, even for using more training data or when downscaling future data.

4.4 Summary & Remarks

In conclusion, improvements to both SSIM and RMSE for all variables are visible using more training, though they do not improve at the same rate. The SSIM and RMSE of future SR data are worse than the present-day SR data for all variables.

It is important to note that going from ST to MT to FF, each case consists of fewer spatial points, even if the training dataset consists of more points. It could lead to more errors and faults in the structure similarity. Also, the region of downscaling becomes more zoomed in with every case. Often, for the worst-case scenarios, the original data (NCEP for ST and MT, EC Earth for FF) looks nothing like the WRF data, leading to a generated SR with no similarity and a high error compared to its WRF counterpart. The generated SR makes logical sense the way it is presented, due to the way the model simulates. It downscales from the original data, which looks nothing like the WRF counterpart. Also, considering the best and worst cases, sometimes the SSIM did improve for a variable going from best to worst. It can happen because the best and worst SRs are chosen from the maximum and minimum values of RMSE. However, the best cases for T2m, U10m, and V10m in the future forcing shows that there is some potential to use this network for future datasets: having small RMSEs. Further, different number of epochs during training were used compared to van Rijk (2022). The number of epochs is 37 for wind and 7 for temperature, while it was the other way around for van Rijk (2022). For the MT models the same number of epochs for the pretraining and other values (alpha, dropout rate) are taking over from van Rijk (2022), while it maybe should have been evaluated what the best parameters were looking at the performance of the MT models of this paper.

For better-downscaled future datasets, future simulated data (like EC Earth data) can be contained in the training of the networks. Then the training dataset could consist of a few years of present data (2017-2020) and a few years of future simulated data (2037-2040, for example). Testing can be done with the year 2041 to see if this leads to better-generated SR. After training with this extra data, future patterns of the different variables may be presented better, because in the future, the structures and values of the variables could be very different than the present-day data due to climate change. A last note, for different variables, maybe different neural networks could be used. For example, Serifi et al. (2021) uses for the variables temperature and precipitation the following network architecture: Residual-Predicting Network (RPN) and Deconvolutional Network (DCN). They conclude there is no single best architecture for all variables by comparing RPN and DCN for their variables. Another example of downscaling using machine learning is presented in Dibike and Coulibaly (2006), where they used a temporal neural network (TNN) to investigate the issues of downscaling present using neural networks. That paper concludes that the TNN is an efficient method for downscaling: for both daily precipitation and daily maximum and minimum temperature series. A last example of a different neural network using downscaling is in Baño-Medina et al. (2021), where they used convolutional neural networks (CNNs) to investigate the suitability of CNNs for downscaling future climate change projections. They concluded that their CNN yields lesser departures from the raw Global Climate Model outputs for the end of the century, making CNNs a plausible application for the downscaling of climate change projections. These three applications of neural networks for downscaling show potential for the future to use different neural networks for different variables, and even for future projections to be downscaled.

Code and data availability. Data: NCEP data is available at National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce (2015).

Code: https://github.com/famkekovacs/Master-Thesis

Competing interests. No competing interests are present.

Acknowledgements. I thank Henk Dijkstra and Bouke Biemond (IMAU) for their supervision. Also, I would like to thank Rene Jansen (HPC system UMC), Martin Marinus (HPC system UMC) and Michael Kliphuis (Info-bio group UU) for their help in navigating the HPC facility of the Utrecht Bioinformatics Community (UBC) and answering all my questions on the usage of it. Further, I thank Emma Smolders (IMAU) for providing the WRF 4km and Anna Napoli (University of Trento) who provided the EC Earth data of 2041. I would like to thank Lorenzo Giovanni (University of Trento) as well, who provided the WRF 9km and 3km data. Moreover, I thank Sebastiaan van Rijk for doing the preliminary work of downscaling over this specific region and having the changed GAN implementation online. And I thank Stengel et al. (2020) for creating and putting their GAN implementation online. At last, I would like to thank my boyfriend and friends whom I have irritated with correcting my English in this paper.

Appendix A: GAN

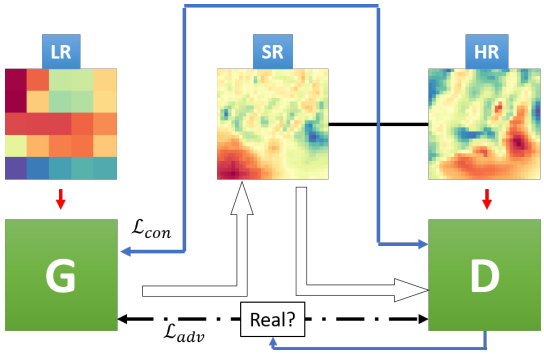


Figure A1. The training scheme for GANs

A Generative Adversarial Network (GAN) consists of two networks working together in an adversarial way: a generator G and a discriminator D, see Figure A1. In this work, a Super Resolution GAN is used based on the work of van Rijk (2022). A neural network consists of layers of nodes, where the layers connect through weights. Convolutional filters extract features from the data. A loss function determines the network performance. One of the hyperparameters in a neural network is the learning rate: it determines the step size of the adjustments to the weights during training. The network training ensures an increase in their performance, which is determined by gradient descent on the dependence of the loss function.

Because a GAN consists of two neural networks, two loss functions exist. The performance measure of the generator, trained by low-resolution data, is the content loss $\mathcal{L}_{con}$. It is calculated by mapping both the created SR and the HR into a feature space and then measuring the distance between the two mapped resolutions in the feature space. This method leads to better performance compared to pixel-based error measures. The second loss function is the adversarial loss function $\mathcal{L}_{adv}$. It is an added penalty for the generator. It is because the generator trains to fool the discriminator. The discriminator trains to distinguish the generated super-resolution.

Some other network hyperparameters are batch size, the learning rate, and the α parameter. This latter parameter is used to determine the strength of the adversarial loss function for the total loss of the generator during the training. It is depicted in Equation A1.

$$\mathcal{L} = \mathcal{L}_{con} + \alpha \mathcal{L}_{adv} \tag{A1}$$

The network is trained in two steps. First, a pretraining of the generator is done with a number of epochs of 39. It happens so that the discriminator will not easily distinguish the SR data from the training data. Otherwise, the training is slowed down. Then, the second step is to train the model in the adversarial fashion of a GAN with a starting generator model with an epoch of 37.

For the T2m variable, the following values of hyperparameters are used: number of epochs = 7; α = 0.001; and the dropout rate = 0.8. For the wind velocity, these parameters are number of epochs = 37; α = 0.001; and dropout rate = 0.9.

Appendix B: Figures

B1 Case More Time Analysis: LR to MR

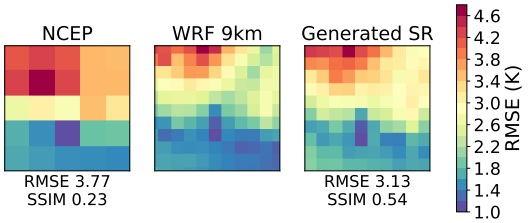


Figure B1. The best super resolution for the variable of T2m in the testing period of 2021. Training period = 2017 - 2020.

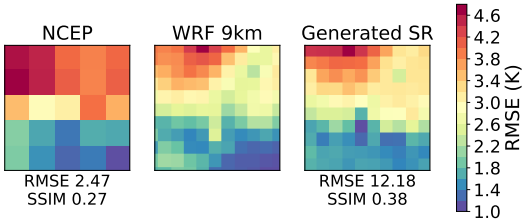


Figure B2. The worst super resolution for the variable of T2m in the testing period of 2021. Training period = 2017 - 2020.

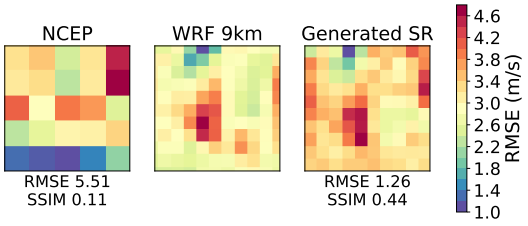


Figure B3. The best super resolution for the variable of U10m in the testing period of 2021. Training period = 2017 - 2020.

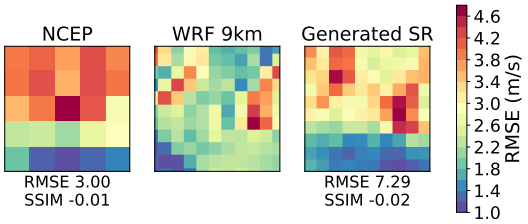


Figure B4. The worst super resolution for the variable of U10m in the testing period of 2021. Training period = 2017 - 2020.

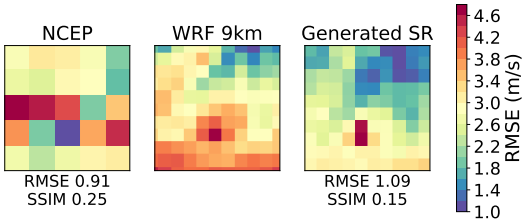


Figure B5. The best super resolution for the variable of V10m in the testing period of 2021. Training period = 2017 - 2020.

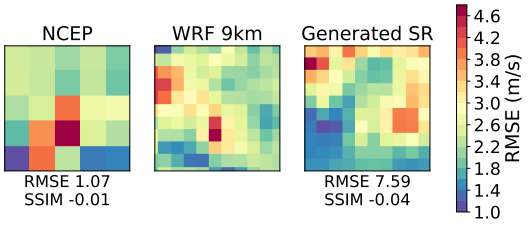


Figure B6. The best super resolution for the variable of V10m in the testing period of 2021. Training period = 2017 - 2020.

References

Wikimedia Commons File:Lake Garda map.png, https://commons.wikimedia.org/wiki/File:Lake_Garda_map.png, accessed: 2023-08-20.

Amadori, M., Giovannini, L., Toffolon, M., Piccolroaz, S., Zardi, D., Bresciani, M., Giardino, C., Luciani, G., Kliphuis, M., van Haren, H., et al.: Multi-scale evaluation of a 3D lake model forced by an atmospheric model against standard monitoring data, Environmental Modelling & Software, 139, 105 017, 2021.

Baño-Medina, J., Manzanas, R., and Gutiérrez, J. M.: On the suitability of deep convolutional neural networks for continental-wide downscaling of climate change projections, Climate Dynamics, 57, 2941–2951, 2021.

Biemond, B., Amadori, M., Toffolon, M., Piccolroaz, S., Haren, H. V., Dijkstra, H. A., et al.: Deep-mixing and deep-cooling events in Lake Garda: Simulation and mechanisms, Journal of Limnology, 80, 2021.

Davini, P., von Hardenberg, J., Corti, S., Christensen, H. M., Juricke, S., Subramanian, A., Watson, P. A., Weisheimer, A., and Palmer, T. N.: Climate SPHINX: evaluating the impact of resolution and stochastic physics parameterisations in the EC-Earth global climate model, Geoscientific Model Development, 10, 1383–1402, 2017.

Dibike, Y. B. and Coulibaly, P.: Temporal neural networks for downscaling climate variability and extremes, Neural networks, 19, 135–144, 2006.

Giovannini, L., Antonacci, G., Zardi, D., Laiti, L., and Panziera, L.: Sensitivity of simulated wind speed to spatial resolution over complex terrain, Energy Procedia, 59, 323–329, 2014.

Glotter, M., Elliott, J., McInerney, D., Best, N., Foster, I., and Moyer, E. J.: Evaluating the utility of dynamical downscaling in agricultural impacts projections, Proceedings of the National Academy of Sciences, 111, 8776–8781, 2014.

Napoli, A., von Hardenberg, J., Pasquero, C., and Parodi, A.: Altitudinal dependence of historical and future extreme events in the Great Alpine Region modelled with WRF, in: EGU General Assembly Conference Abstracts, pp. EGU22–2752, 2022.

National Centers for Environmental Prediction, National Weather Service, NOAA, U.S. Department of Commerce: NCEP GFS 0.25 Degree Global Forecast Grids Historical Archive, https://doi.org/10.5065/D65D8PWK, 2015.

Råman Vinnå, L., Medhaug, I., Schmid, M., and et al.: The vulnerability of lakes to climate change along an altitudinal gradient, Commun Earth Environ, 2, 2021.

Serifi, A., Günther, T., and Ban, N.: Spatio-temporal downscaling of climate data using convolutional and error-predicting neural networks, Frontiers in Climate, 3, 656 479, 2021.

Smolders, E.: Mixing processes in Lake Garda, 2022.

Stengel, K., Glaws, A., Hettinger, D., and King, R. N.: Adversarial super-resolution of climatological wind and solar data, Proceedings of the National Academy of Sciences, 117, 16 805–16 815, 2020.

van Rijk, S.: Downscaled future forcing for Lake Garda, 2022.