CHALMERS
UNIVERSITY OF TECHNOLOGY

A Bayesian machine learning approach to passive microwave precipitation retrievals

Master's thesis in Complex Adaptive Systems

Teodor Norrestad

Department of Space, Earth and Environment
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2019

Master's thesis 2019

A Bayesian machine learning approach to passive microwave precipitation retrievals

TEODOR NORRESTAD

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Space, Earth and Environment
Division of Microwave and Optical Remote Sensing
Chalmers University of Technology
Gothenburg, Sweden 2019

A Bayesian machine learning approach to passive microwave precipitation retrievals
TEODOR NORRESTAD

Supervisor: Patrick Eriksson, Department of Space, Earth and Environment
Examiner: Patrick Eriksson, Department of Space, Earth and Environment

Master's Thesis 2019
Department of Space, Earth and Environment
Division of Microwave and Optical Remote Sensing
Chalmers University of Technology
SE-412 96 Gothenburg

Typeset in LaTeX
Gothenburg, Sweden 2019

A Bayesian machine learning approach to passive microwave precipitation retrievals
TEODOR NORRESTAD
Department of Space, Earth and Environment
Chalmers University of Technology

Abstract

A machine learning-based approach to precipitation retrievals, using Quantile Regression Neural Networks (QRNNs), is developed for data from the Global Precipitation Measurement (GPM) mission. The retrievals are conducted within a Bayesian framework where the networks are trained to predict quantiles of the posterior distribution of rain rates, conditioned on passive microwave observations. In this way, rain rates are retrieved along with the associated retrieval uncertainties. The effects of including additional spatial information as input to the QRNNs are also investigated. Different QRNNs are trained and tested, first globally over oceans and then over the U.S Great Plains. In both cases, the performance of the QRNNs are compared to the Goddard Profiling Algorithm (GPROF), a state-of-the-art passive microwave retrieval algorithm. The primary results are those over oceans, where the QRNNs show great performance on similar levels as GPROF with respect to point estimate metrics such as the mean squared error. Furthermore, the QRNN retrievals are very fast, taking less than a millisecond per footprint on a standard computer. It turns out that extra spatial information improves the QRNNs, especially on making rain-no rain classifications with fractions of true positives and true negatives exceeding 0.67 and 0.96 respectively. Furthermore, the QRNNs manage to produce well calibrated quantiles, resulting in good confidence intervals to account for retrieval uncertainties. Over the Great Plains, the results are promising but are based on much smaller amounts of data and are thus less significant.

Keywords: machine learning, neural networks, precipitation, retrievals, Bayesian, QRNN, GPM.

Acknowledgements

I would like to thank my main supervisor Patrick Eriksson[1] for great advice and feedback throughout the project. Also, big thanks to Simon Pfreundschuch[1], David Duncan[2], Robin Ekelund[1] and Vasileios Barlakas[1] for helpful discussions and assistance. Furthermore, I would like to thank Christian Kummerow[3] and Veljko Petkovic[3] for being very helpful and answering questions about their work.

Teodor Norrestad, Gothenburg, May 2019

[1]Department of Space, Earth and Environment, Chalmers University of Technology, Gothenburg, Sweden

[2]The European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, UK

[3]Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado, USA

Contents

1 Introduction **1**
- 1.1 Motivation . . . 1
- 1.2 Bayesian retrievals . . . 1
- 1.3 Aim and limitations . . . 2

2 Background **4**
- 2.1 The Global Precipitation Measurement mission . . . 4
  - 2.1.1 DPR . . . 4
  - 2.1.2 GMI . . . 5
  - 2.1.3 GPROF . . . 6
- 2.2 Machine learning . . . 7
  - 2.2.1 Neural networks architecture . . . 7
  - 2.2.2 Loss functions and backpropagation . . . 8
  - 2.2.3 Quantile regression neural networks . . . 8

3 Method **10**
- 3.1 Building the training and testing databases . . . 10
  - 3.1.1 The data files . . . 10
  - 3.1.2 Resolution of the DPR . . . 11
  - 3.1.3 Sampling data . . . 12
  - 3.1.4 Creating training input-output pairs . . . 12
  - 3.1.5 Testing database . . . 13
  - 3.1.6 Databases for land . . . 14
- 3.2 Neural networks . . . 15
  - 3.2.1 Training metrics . . . 15
  - 3.2.2 Structural parameters . . . 15
  - 3.2.3 Training parameters . . . 15
- 3.3 Evaluation metrics . . . 16
  - 3.3.1 Point estimate metrics . . . 16
  - 3.3.2 Rain-no rain classification metrics . . . 16
  - 3.3.3 Error estimation metrics . . . 17

4 Results **19**
- 4.1 Predictions over the ocean . . . 19
  - 4.1.1 Example scenes . . . 21
  - 4.1.2 Quantitative results . . . 23
- 4.2 Predictions over the Great Plains . . . 29
  - 4.2.1 Example scenes . . . 30
  - 4.2.2 Quantitative results . . . 31

5 Conclusions **36**
- 5.1 Results . . . 36
- 5.2 Outlook . . . 37

1 Introduction

1.1 Motivation

The detection of precipitation on a global scale is of great societal benefit for a number of reasons. These include estimating floods, draughts and fresh water availability (Hou et al., 2014). Other examples are of a more economical nature, such as hydropower planning where predictions of water flows, which rely on measurements of precipitation, play an important role. For similar reasons, estimates of recent precipitation is important for agricultural applications, such as irrigation scheduling. Another major application of precipitation measurements are as input to weather forecast models, where better coverage and accuracy can improve predictions of future precipitation and other climate variables, such as temperature and wind speed (Tapiador et al., 2012).

Global coverage of precipitation is also of great scientific use, such as in improving the understanding of the Earth's water cycles and how they are affected by climate change (Hou et al., 2014). Collecting precipitation observations over time may help to predict if, for example, rising temperatures increase the probability of extreme precipitation events and to monitor how the global distribution of rain changes (Kidd and Huffman, 2011).

Ground-based radars are widely used for detecting and measuring the intensity of precipitation but give far from global coverage. Remote areas and oceans are typically not within the reach of the ground radar networks and extensions are generally both costly and impractical. As mountainous regions and especially oceans cover a large portion of the earth's surface, the precipitation in these areas have great effect on the global climate and weather conditions everywhere, making global precipitation measurements crucial (Germann et al., 2006; Islam et al., 2017). An alternative to ground based precipitation measurements is using satellites. Space-borne radar endeavours, such as the Global Precipitation Measurement (GPM) mission (NASA, 2019b), with its constellation of satellites, aim to gain a global precipitation coverage. The GPM Core Observatory, launched in 2014, is equipped with two instruments for precipitation measurements, a Dual-Frequency Precipitation Radar (DPR) and a passive microwave imager (GMI).

1.2 Bayesian retrievals

Measurements obtained directly from sensors, such as radars and microwave imagers, are in general not the quantities that are of interest, like rain rates for example. The process of extracting them is called retrievals and the associated problems of doing so are called inversions. If, for example, the quantity of interest is the rain rate y over a given area and an observation $\mathbf{x}$ of GMI brightness temperatures is made, the inversion problem consists of calculating a value of y that might have caused $\mathbf{x}$. In general, inversions are not unique and rely on the accuracy of the sensors and assumptions of models and are thus usually associated with a great deal of uncertainty. It is therefore crucial for a retrieval to include estimates of the involved uncertainties in some way.

While the rain rate retrievals of the DPR have shown good results with ground

based systems as reference (Kidd et al., 2018), its across track swath width of 245 km is narrow compared to the 885 km swath of the GMI. In addition, all the other satellites in the GPM constellation are equipped with passive microwave instruments similar to the GMI. Retrieval algorithms based on passive microwave measurements alone would thus be a great step towards better coverage. One such algorithm is the Goddard Profiling (GPROF) algorithm, which makes rain rate inversions based on brightness temperature observations from the GMI and other microwave imagers in the constellation. The inversions of GPROF are Bayesian, which means that the sought solution is the posterior distribution, $p(y|\mathbf{x})$, of rain rates y conditioned on the measured brightness temperatures $\mathbf{x}$. According to Bayes' theorem, the posterior distribution is proportional to the product of the a priori distribution $p(y)$ and the likelihood $p(\mathbf{x}|y)$ (Rodgers, 2000). The GPM core observatory is used to build a large a priori database of matched DPR and GMI observations which is used to derive the factors of the posterior distribution. For a given brightness temperature observation, the inversion process consists of searching the a priori database for a set of similar profiles. The corresponding rain rates are weighted according to similarity and averaged over to complete the retrieval. The Bayesian approach deals with the issue of capturing the uncertainty of the retrieval by not only presenting a single value for the retrieval, but rather an entire distribution of it.

1.3 Aim and limitations

The development of a retrieval algorithm in this study is also conducted within a Bayesian framework. Just as for GPROF, the foundation is a an a priori database, used to retrieve the posterior distribution $p(y|\mathbf{x})$ of rain rates given passive microwave observations. A major difference, however, is that instead of searching the database when retrieving, the database is used to train regression neural networks. Neural networks are basically parametric non-linear functions that learn, through training, to map inputs to outputs, which in this case are microwave measurements, $\mathbf{x}$, and rain rates, y, respectively. The training is a process where the network learns iteratively from sampled input-output pairs $\{\mathbf{x}_i, y_i\}$ of the training database. This often requires a lot of training samples in order for the network to successfully learn the mapping and generalise to new cases. Once the training phase is complete, the neural network is decoupled from the database and the retrieval consists in a forward pass through the network. To maintain the Bayesian ideas, a certain type of networks, called Quantile Regression Neural Networks (QRNNs) are used. While a standard regression neural network usually learns to predict a single point estimate, the QRNN learns to predict quantiles (or percentiles) of $p(y|\mathbf{x})$. With enough quantiles predicted, the posterior distribution can be approximated.

Regression neural networks and QRNNs have shown good performance on other retrieval problems, such as the retrieval of cloud top pressure for example (Håkansson et al., 2018; Pfreundschuh et al., 2018), where QRNNs prove to be better suited to capture non-Gaussian errors distributions. Other neural network based retrievals have been developed for the GMI as well (Sanò et al., 2018), not, however, with QRNNs.

The aim of this study is to apply QRNNs to the GPM rain rate retrieval prob-

lem. The goal is to develop a neural networks-based algorithm that uses the GMI observations to accurately predict DPR rain rates and at the same time capture the uncertainties as well. Additionally, the effects of adding extra spatial information from the GMI are investigated. Since the GPROF algorithm relies on searching a database of matched GMI and DPR footprints, extending the size of the input while maintaining efficient searching is far from trivial. For neural networks, this is however less complicated. The idea is that adding GMI measurements from neighbouring footprints can lead to more accurate retrievals, especially when the background emissivity is non-homogeneous.

GPROF is capable of doing both liquid and solid precipitation retrievals globally, over any surface type. The scope of this study is limited to measuring liquid precipitation over oceans and certain more homogeneous land types. In particular, the U.S Great Plains is chosen as a first area to test the QRNNs over land. A neural network based retrieval algorithm for all precipitation and surface types would need a huge training database and require lots of training. While this is possibly doable with enough time and computational power, the focus of this study is to apply and evaluate new retrieval techniques rather than to build a complete precipitation retrieval product.

The first step in the work is to construct an a priori database, based on DPR and GMI data, with extra spatial information included for training. Then QRNNs are trained with different numbers of neighbouring footprints included as input. Once the training process is complete, the different QRNNs are evaluated on a test set and compared to each other and to the GPROF predictions.

2 Background

This section covers some of the basic characteristics of the GPM, the satellite mission from which all of the data is collected. Additionally, brief overviews of neural networks, quantiles and QRNNs are included that are necessary in order to follow the description of the work process.

2.1 The Global Precipitation Measurement mission

The Global Precipitation Measurement (GPM) mission includes a constellation of satellites with the aim of providing global precipitation measurements (NASA, 2019b). As a collaboration between NASA and JAXA, the GPM core observatory was launched in February 2014. With its combination of an active dual-frequency radar, the DPR, and a passive microwave imager, the GMI, the core observatory provides an observational database to unify the retrievals throughout the constellation (Hou et al., 2014). Compared to previous missions, such as TRMM (Huffman et al., 2007), the GPM provides a greater global coverage and more accurate retrievals of light rain and solid precipitation.

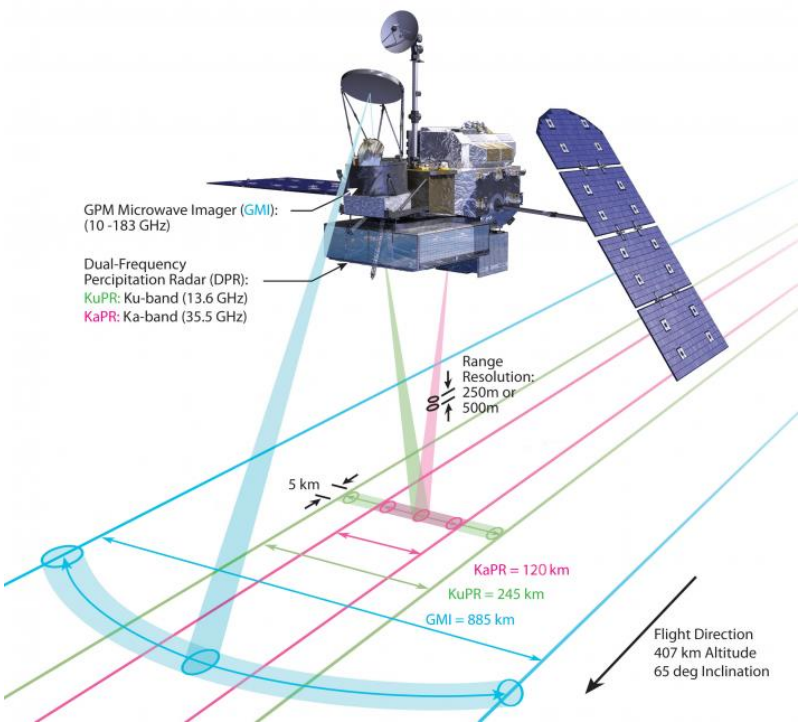

Figure 1: The GPM Core Observatory with its passive and active sensors, the GMI and the DPR with their respective swath widths. Image: NASA.

2.1.1 DPR

One of the main features of the GPM core observatory is the Dual-Frequency Precipitation Radar (DPR). As the name suggests, the DPR consists of two radars, a Ka-band and Ku-band precipitation radar with frequencies 35.5 GHz and 13.6

GHz, respectively. They both have 5 km sized footprints while the Ku-band radar has a wider across track swath (245 km) than the Ka-band radar (125 km). The DPR is an active instrument in the sense that microwave radiation is emitted by the radars. By measuring the backscattered signal as it is reflected by particles in the atmosphere, information can be obtained about precipitation particles. The combined measurements of the two radars give insights to the precipitation Particle Size Distribution (PSD), which plays an essential role in the retrieval of rain rates. The higher sensitivity of the DPR, compared to TRMM for example, allows for detection of rain rates down to about 0.2 mm/h. With a vertical resolution of 250 m, the DPR provides a three-dimensional rain profile. Throughout this work, however, only the precipitation close to the surface is considered. Measurements too close to the surface are however affected by ground clutter effects that disturb the signal (Islam et al., 2012; Rico-Ramirez and Cluckie, 2008). Therefore, precipitation at the surface is estimated from the rain profile.

2.1.2 GMI

Along with the DPR, the GPM core observatory is equipped with a passive radiometer, the GPM Microwave Imager (GMI) (Hou et al., 2014). As a passive instrument, the GMI measures emitted radiation from the Earth's surface and the atmosphere. When radiation passes through the atmosphere, it interacts with particles before it reaches the sensor. It's these interactions, along with emission from the particles, that makes it possible to detect precipitation. How much effect the precipitation particles have on the measured radiation, relative to the surface emission, varies with the surface type. Over oceans, that are rather homogeneous with respect to emissivity, precipitation is more easily detected than over land where the surface emissivity is often high compared to that of the precipitation particles and not as homogeneous.

The GMI has 13 different channels with frequencies between 10 and 183 GHz, divided into two groups with the channels 1-9 in group 1 and 10-13 in group 2. The GMI has a conical across track swath of 885 km at an incidence angle of 52.8° for the group 1 channels. The channels of group 2 has an incidence angle of 49.19°. Because of the different incidence angles, the footprint centres of the two groups are not co-located. For this reason, only the channels of group 1 are used. The details regarding the different GMI channels are summarised in table 1.

Table 1: Characteristics of the GMI.

Frequency (GHz)	**Resolution (km)**	**Group**	**Nominal Earth incidence angle**	**Off-nadir**
10.65 (V & H)	19.4×32.2	1	52.8°	48.5°
18.7 (V & H)	11.2×18.3			
23.8 (V)	9.2×15.0			
36.5 (V & H)	8.6×15.0			
89.0 (V & H)	4.4×7.3			
166 (V & H)		2	49.19°	45.36°
$183.31 \pm 3, \pm 7$ (V)				

2.1.3 GPROF

The development of the Goddard Profiling (GPROF) algorithm started in the 90s, in attempts to make precipitation retrievals based on the microwave imagers of the TRMM (Kummerow et al., 2015). Several versions have been developed since then and successfully adapted for the GPM mission. The latest versions to date are the GPROF2017 versions 1 and 2 (GPM Science Team, 2018). As a Bayesian algorithm, GPROF aims to retrieve rain rates based on the posterior distribution $p(y|\mathbf{x})$ of rain rates conditioned on observed microwave brightness temperatures. According to Bayes' theorem

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)p(y)}{\int p(y', \mathbf{x})dy'}, \tag{1}$$

where $p(y)$ is the a priori distribution which in this case represents what is known about the rain rates without considering any measured brightness temperatures and $p(\mathbf{x}|y)$, the likelihood, is the probability of observing $\mathbf{x}$ given that the rain rate is y. GPROF uses an approach called Bayesian Monte Carlo integration (BMCI) to approximate conditional statistics based on the posterior distribution (Pfreundschuh et al., 2018). The expected rain rate given a set of measured brightness temperatures $\mathbf{x}$ can be obtained by evaluating the integral

$$E[y|\mathbf{y}] = \int y' p(y'|\mathbf{x})dy'. \tag{2}$$

The BMCI scheme consists in rewriting eq. (2) using the Bayesian formulation eq. (1) to get

$$E[y|\mathbf{x}] = \int y' \frac{p(\mathbf{x}|y')p(y')}{\int p(y'', \mathbf{x})dy''}dy', \tag{3}$$

which is then approximated using a database $\{\mathbf{x}_i, y_i\}_{i=1}^N$ that follows the a priori distribution of rain rates. For GPROF this is a large observational a priori database of matched rain rates and brightness temperatures. The brightness temperatures are simulated using forward radiative transfer models to fit the characteristics of microwave imagers on board different satellites within the GPM constellation. Each element in the database is assigned a weight $w_i(\mathbf{x})$, based on its similarity with the observed brightness temperatures, and is calculated as

$$w_i(\mathbf{x}) = \exp\left[-0.5(\mathbf{x}_O - \mathbf{x}_{i,S})\mathbf{R}^{-1}(\mathbf{x}_O - \mathbf{x}_{i,S})\right], \tag{4}$$

where $\mathbf{x}_O$ is the observed brightness temperature vector and $\mathbf{x}_{i,S}$ the simulated brightness temperature vector in the database. $\mathbf{R}$ is the error covariance matrix, describing the observational and forward simulation errors (GPM Science Team, 2018). Using these weights, the right hand side of eq. (3) is approximated by a weighted sum over the database

$$E[y|\mathbf{x}] \approx \frac{1}{C}\sum_{i=1}^{N} w_i(\mathbf{x})y_i, \tag{5}$$

where $C = \sum_{i=1}^N w_i(\mathbf{x})$ is a normalising sum over all the weights. In practice, the entire database isn't used for the retrievals as the operations involved would be very

computationally heavy. In order to make the calculations efficient, ancillary data is added to both database inputs and observations. These include surface temperature, Total Column Water Vapour (TCWV) and surface type and are used to partition the database such that only relevant parts, with respect to the input, are considered. The found subset of the database then constitutes $p(y|\mathbf{x})$ and is weighted and averaged over to obtain the expected rain rate. Additionally, the first and second terciles, i.e. the 0.33th and 0.66th quantiles, are calculated. The considered subset is sorted with respect to rain rates and split into three equally sized parts. The boundaries are where the accumulated weights constitute 33% and 66% respectively of the total sum of weights. The terciles are these two boundary values.

2.2 Machine learning

In general, machine learning refers to algorithms that perform some task without explicit instructions, relying rather on experience. In supervised learning, this experience is obtained from a training set of inputs $\mathbf{x}$ and outputs $\mathbf{y}$. The training usually consists of a parametric model adjusting its parameters to map inputs to desired outputs. The tasks can vary over a broad range of applications from classification to colouring of grey-scale images for example. In this study, the main task is regression, that is, predicting scalar outputs from given input vectors.

2.2.1 Neural networks architecture

One of the most common trends in machine learning in recent year is the use of neural networks (Tang et al., 2007). The fundamental building block of a neural network is the neuron. The neuron is mathematically just a function with parameters $\mathbf{w}$ and θ, called weights and bias respectively, that takes some input and produces an output. The input to a neuron is weighted and passed through an activation function f. To form a network, the neurons are arranged in layers with the first one being the input layer and the last one the output layer. Between the input and output layers are the so called hidden layers. The training input $\mathbf{x}$ is propagated through these layers to produce the output $\hat{\mathbf{y}}$ according to

$$\mathbf{x}_0 = \mathbf{x}, \tag{6}$$

$$\mathbf{x}_i = f_j(\mathbf{w}_j \mathbf{x}_{j-1} + \theta_j), \text{ for } j = 1, ..., n, \tag{7}$$

$$\hat{\mathbf{y}} = \mathbf{x}_n. \tag{8}$$

This process is called forward propagation. Here f_j are non-linear activation functions and n the number of layers. The weights and biases are trainable parameters of the network. The number of neurons in the input and output layers are the same as the dimensions of the input and output respectively. The number of hidden layers and the number of neurons in each layer, usually referred to as depth and width, are design properties of the network that are not trainable, so called hyper-parameters that have to be set in advance of training.

2.2.2 Loss functions and backpropagation

The training of a neural networks is the process of mapping inputs to outputs taken from a training set $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$. This is done by minimising a loss function $L(\hat{\mathbf{y}}, \mathbf{y})$. The minimisation problem is usually solved by using some gradient descent scheme. A key step in the process is called backpropagation, which is basically calculating the gradient of $L(\hat{\mathbf{y}}, \mathbf{y})$ with respect to the trainable parameters of the model, that is $\mathbf{w}_i$ and θ_i for $i = 0, ..., n$. The gradient is calculated numerically by first taking a sample, or batch, of the data, and propagating it forward to obtain $\hat{\mathbf{y}}$. The loss function is then evaluated and the error is propagated backwards to calculate the derivative of the loss function with respect to all the weights and biases. Once these derivatives, $\frac{\partial L}{\partial \mathbf{w}_i}$ and $\frac{\partial L}{\partial \theta_i}$, are calculated, the weights and biases are adjusted according to

$$\mathbf{w}_i = \mathbf{w}_i - \eta \frac{\partial L}{\partial \mathbf{w}_i} \tag{9}$$

$$\theta_i = \theta_i - \eta \frac{\partial L}{\partial \theta_i}, \tag{10}$$

for $i = 0, ..., n$. The learning rate η is another hyper-parameter along with the size of the batches. When the samples included in the batches are chosen at random, this is called stochastic gradient descent. This is usually preferred in order to avoid getting stuck in local minima of the loss function. Processing the entire training set once is referred to as an epoch. One example of a loss function, that is somewhat of a standard, is the mean squared error loss

$$L(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{N}(\hat{\mathbf{y}} - \mathbf{y})^T(\hat{\mathbf{y}} - \mathbf{y}). \tag{11}$$

By minimising this function, the network learns to make the outputs $\hat{\mathbf{y}}_i$ as similar as possible to the training labels $\mathbf{y}_i$ on average, using the L_2 metric. When concerned with regression, the choice of the loss function determines what statistic are to be learned. For example, using the mean squared error, the network learns to predict the mean of $p(\mathbf{y}|\mathbf{x})$, given that the training database reflects the a priori distribution.

2.2.3 Quantile regression neural networks

Some other statistics that are of interest when describing distributions are quantiles. Let X be a random variable with cumulative distribution function (CDF) $F_X(x) = P(X < x)$. The τth quantile of X, x_τ is defined as

$$x_\tau = \inf\{x : F_X(x) \geq \tau\} \text{ for } \tau \in (0, 1), \tag{12}$$

that is, x_τ is the largest lower bound of numbers x such that the probability of observing a value of X smaller than x is at least τ. Anther way to think about it is that the τth quantile is the value x that gives τ when inserted into the CDF of X, i.e. $x_\tau = F_X^{-1}(\tau)$. For example, this means that, given a sample of X, if the 0.1th quantile is 2, approximately 10% of the sample values are smaller than 2. The 0.5th quantile of X is the same as the median of X. Relating this to the probability density function (PDF) of X, the τth quantile is a value x such that a fraction τ of

the area under the PDF is to the left of x. As an example, the 0.1th, 0.5th and 0.9th quantiles of a normal distribution with mean 3 and variance 1, are shown in relation to the CDF and PDF in fig. 2. Note that for a normally distributed random variable, the mean and median are the same. This is however not the case for non-symmetric distributions.

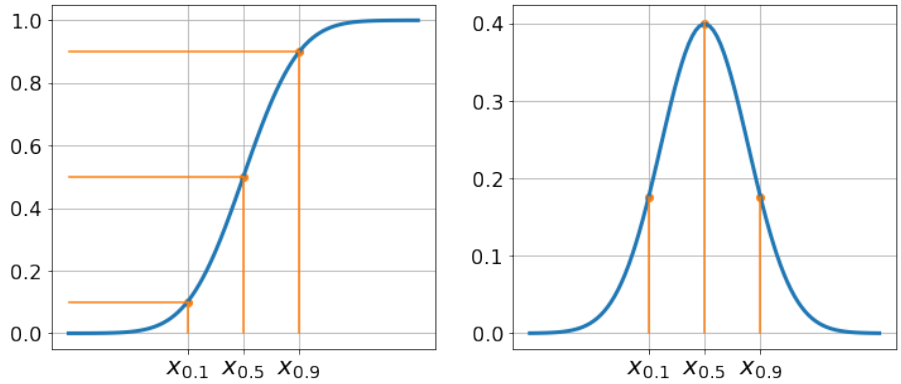

Figure 2: Illustration of the 0.1th, 0.5th and 0.9th quantiles using a normally distributed random variable. The left panel shows the quantiles along with the CDF and the right panel with the PDF.

By training a neural network to minimise the mean of the loss function

$$L_\tau(x_\tau, x) = \begin{cases} \tau|x - x_\tau|, & \text{if } x_\tau < x \\ (1-\tau)|x - x_\tau|, & \text{otherwise,} \end{cases} \tag{13}$$

on the training set $\{\mathbf{x}_i, y_i\}_{i=1}^N$, the network learns to predict the τth quantile of the distribution $p(y|\mathbf{x})$ (Koenker and Hallock, 2001). Thus, a neural network trained to minimise eq. (13) is called a Quantile Regression Neural Network (QRNN). By varying τ in eq. (13), the network can learn to predict multiple quantiles of $p(y|\mathbf{x})$. The CDF, $F_{y|\mathbf{x}}(y)$, can then be approximated using these quantiles. For the quantiles to be predicted correctly, it's very important that the training database properly resembles the a priori distribution of the rain rates.

3 Method

The overall work process is mainly divided into three parts. First, it's the construction of databases using satellite data. Second is the design and training of neural networks, using the training database. Finally, once all networks are trained, they are evaluated on a testing database and compared to GPROF. This section describes these three parts in more detail.

3.1 Building the training and testing databases

A major part of the work consists of constructing appropriate databases for training and testing. As mentioned before, the Bayesian approach relies heavily on the training database's similarity to the true a priori distribution of the data. The process of building these databases are described in more detail, including how data are chosen and what data are used for the different networks.

3.1.1 The data files

All the data files are downloaded from NASA's Earth Observing System Data and Information System (EOSDIS) (NASA, 2019a). Data from all the GPM products have been made available on a daily basis since the launch of the GPM core observatory in 2014. Given a product, for each day there is a batch of data divided into about 16 smaller files, each corresponding to a full orbit of the satellite. For the training database, only the DPR and GMI data files are used. The DPR data are taken from the Level 2A product which contains retrieved rain rates. Along with the rain rate of each footprint, its position (longitude, latitude) is stored. Information about the surface type is also included, indicating whether the footprint is covering land, ocean, coast or inland water. The GMI data are taken from the Level 1C product containing calibrated brightness temperatures for each one of the channels. The coordinates of the GMI footprints are also stored. An example of what the data looks like is shown in fig. 3, where the DPR rain rates are plotted along with brightness temperature for two of the GMI channels.

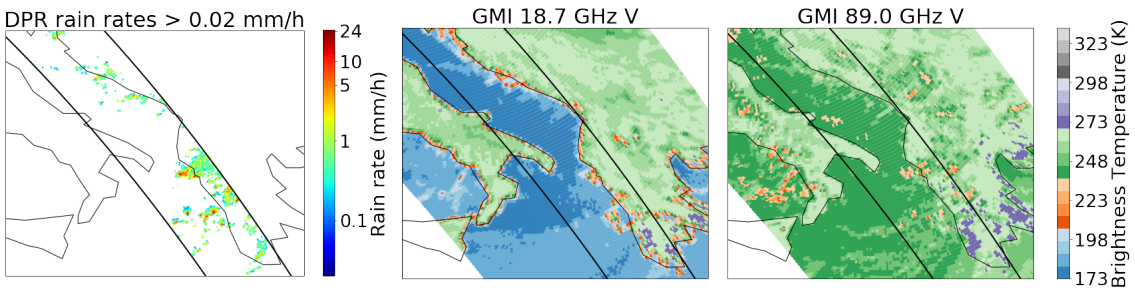

Figure 3: An example of a scene in the Mediterranean area. The left panel shows DPR rain rates while the middle and right panel show GMI brightness temperatures for the 18.7 V and 89.0 V channels respectively.

3.1.2 Resolution of the DPR

All precipitations retrievals are made given some spatial resolution, that is, some area needs to be specified over which measurements are averaged. Since the resolution of the DPR is better than that of most of the GMI channels, a filter is applied to the DPR to make the resolutions comparable. The filter consists of a weighted average over a 9×9 neighbourhood around a given DPR footprint. The weights, shown in fig. 4, are taken from a discretised Gaussian distribution and are the same that are used by the GPROF algorithm. In this way, the resulting predictions of the QRNNs are comparable with those of GPROF.

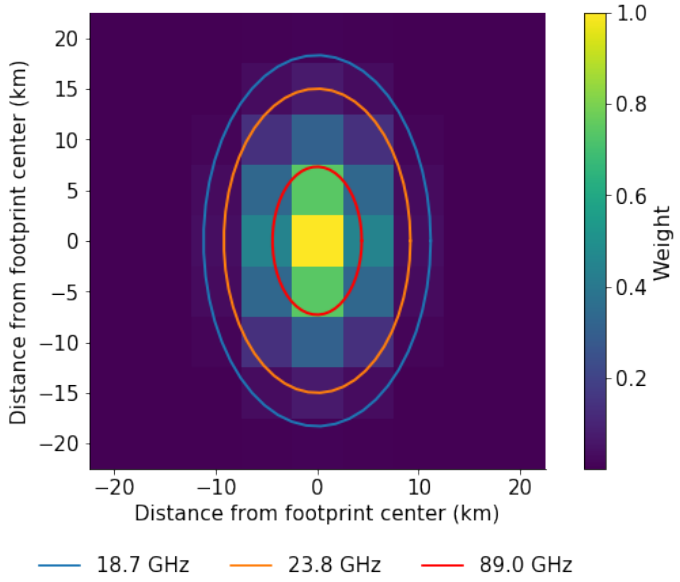

Figure 4: Weights of the averaging filter that is applied to DPR rain rates, plotted along with the footprint sizes of the 18.7, 23.8 and 89.0 GHz channels of the GMI

Changing the resolution results in more smeared out pictures. An example is shown in fig. 5 where the rain rates at DPR resolution are plotted along with rain rates at the new resolution. As can be seen, the averaging smears out the image and lowers the maximum rain rate while adding more patches of low intensity rain.

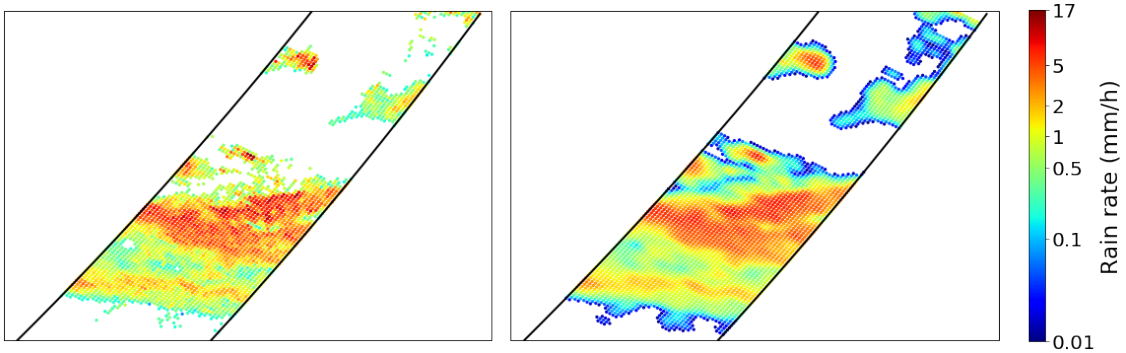

Figure 5: The DPR rain rates at original 5×5 km resolution (left panel) and at new resolution after filtering (right panel).

3.1.3 Sampling data

The data used for building the training database are taken from 2015. In order for the training data to represent the a priori distribution properly, data are taken from the first of each month throughout the year. This ensures that seasonal variations are included. For each day a number of orbits are chosen at random. Given an orbit, the data over oceans are picked out using the surface type variable of the DPR. To make sure that no footprints cover land, the longest connected sequence of ocean data are chosen and trimmed at the start and end. These footprints constitute the candidates for being included in the training database. Since the orbit patterns change over time, the chosen sequences are not always taken from the same area, but are restricted to larger oceans. An example is shown in fig. 6 where only the centre footprints of two orbits are plotted. The orbits are taken from the same day, but because of the different paths, one sequence is picked over the Pacific ocean and one over the Atlantic ocean.

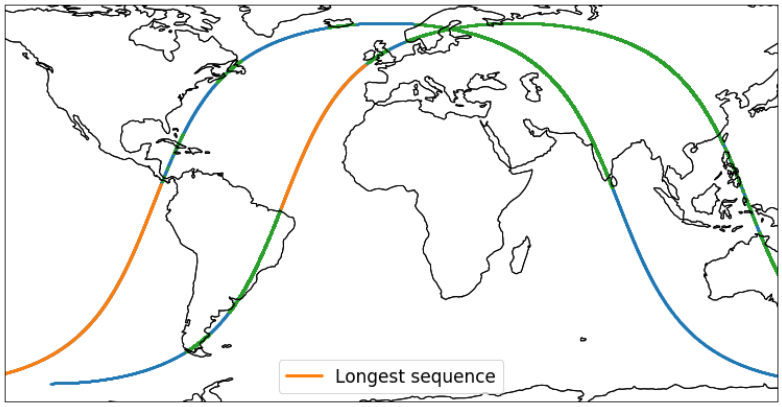

Figure 6: Centre footprints of two different orbits with the longest sequence chosen highlighted in orange. The lines do not represent the actual swath width.

3.1.4 Creating training input-output pairs

Once such a sequence is chosen, the next step is to find matching DPR and GMI footprints, that is, to each of the DPR footprints, find the closest GMI footprint. This process is fairly simple, but there are some complications. One is that while the DPR scans the surface directly below the satellite, the GMI scans in front of it with an incidence angle of about 52°. This is accounted for by adding a constant shift in the indices of the data files. In addition, an orbit of the DPR contains more across path scans than the GMI. To ensure that the closest pair of footprints is found, a search is performed in a small neighbourhood of approximately 3×15 footprints around the candidate GMI footprint. Because of the less frequent across track scans of the GMI, the closest footprint is sometimes far away, up to 7 or 8 km, compared to some of the footprint sizes. Therefore a threshold is introduced, ensuring that only matchings closer than 4 km are added to the database. An example of such a matching is shown in fig. 7, where only some matchings are closer than the threshold. The positions of the matched footprints vary along the DPR swath, because of the exact geometry that varies over the orbit.

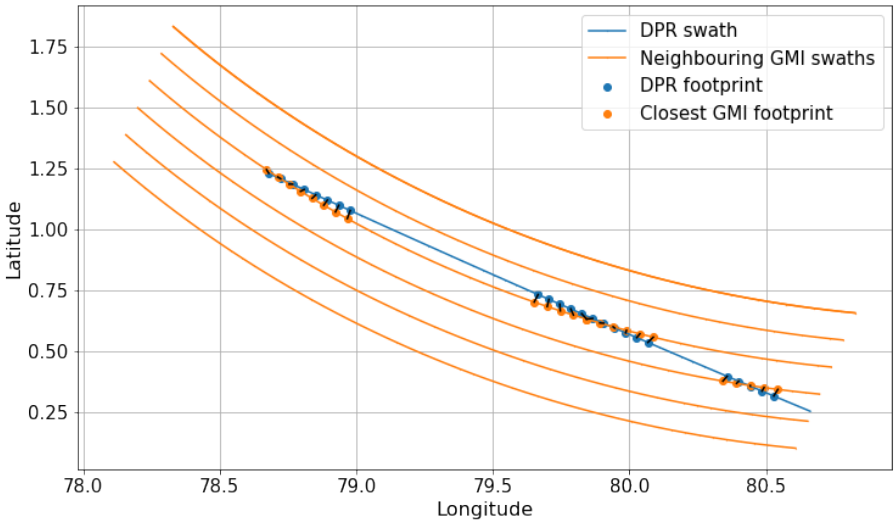

Figure 7: A number of footprints on a swath of the DPR are matched with neighbouring GMI footprints. Only matches closer than 4 km are kept.

Each pair of matched DPR and GMI footprints forms the basis of a training pair. The training label, y_i, is the DPR rain rate at the given position, taken from the filtered DPR data. The training input, $\mathbf{x}_i$, is the vector of brightness temperatures from the 9 GMI channels at the matching footprint. To include extra spatial information, the 9 brightness temperatures of 12 neighbouring GMI footprints are added as well. From the 12 neighbouring footprints, 4, 8 or 12 are chosen to train different networks. The way they are chosen relative to the closest matching footprint is shown in fig. 8. In addition, to each $\mathbf{x}_i$ a latitude variable is inserted. This variable takes a value between 0 and 5, indicating in which out of 6 equally sized latitude bins the footprint is located. The input dimension, given a number of neighbours n, is then $9 + 1 + 9 \times n$.

When an entire orbit file is processed, the training pairs $\{\mathbf{x}_i, y_i\}$ are appended to a file saved on the disk, along with the respective coordinates of the matched footprints. From this database appropriate training sets are chosen for the different neural networks to train on.

3.1.5 Testing database

The process of constructing a testing database is similar to that of the training database, but differs in a few ways. First, the data are taken only from 2017 and 2018, ensuring that there is no overlap between training and test data. Second, the testing database also includes GPROF predictions. The GPROF files have the same dimensions as the GMI files, so for every matched footprint the GPROF's predicted rain rate is simply added to the database along with the predicted Probability of Precipitation (PoP). Additionally, for each test sample the two tertials predicted by GPROF are added.

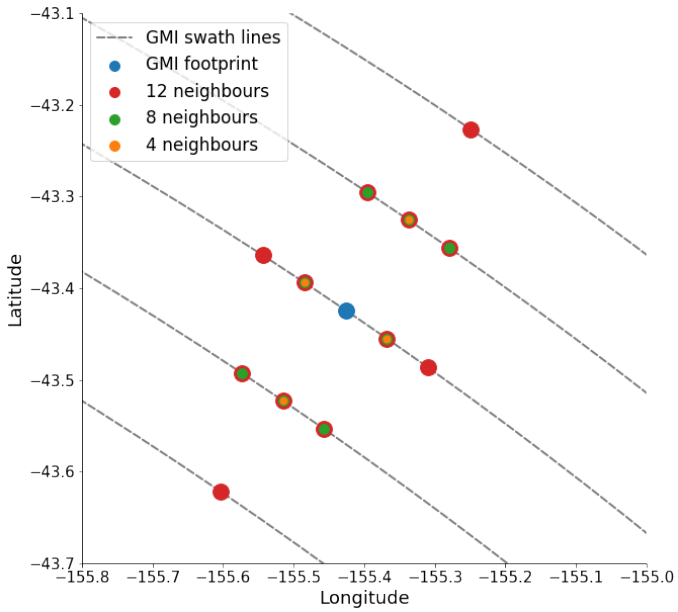

Figure 8: The configurations of different neighbourhoods given a centre GMI footprint plotted along with GMI swath lines.

3.1.6 Databases for land

To develop QRNNs to predict precipitation over land, training and testing databases are constructed in a similar way as over oceans. However, in contrast to the ocean databases, only a single area is considered - the U.S Great Plains. This particular area is chosen due to the homogeneous surface type consisting of mostly steppes and grassland. The area that measurements are taken from is shown in fig. 9, enclosed by the dashed lines. Because of the limited area, data are taken from the first and 14th of each month throughout 2015 and 2016 in order to get reasonable sizes of the training and test databases.

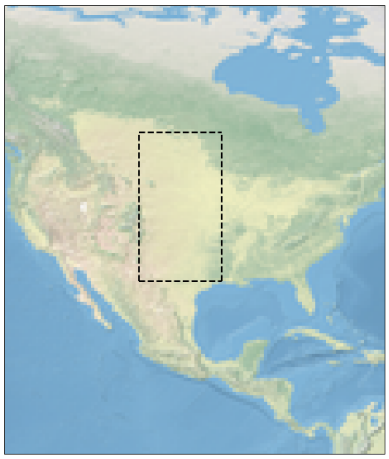

Figure 9: The U.S Great Plains with dashed box to mark area of retrievals.

3.2 Neural networks

The first step in training any neural network is choosing appropriate hyper-parameters. These can be divided into two groups; parameters regarding the general structure of the network and parameters regarding the training process. How and which hyper-parameters are chosen is dealt with here. The QRNNs are implemented using the typhon package (The typhon authors, 2019), an extension of the Python Keras package (Chollet et al., 2019).

3.2.1 Training metrics

When selecting hyper-parameters for a neural network, essentially two metrics are used and monitored between training epochs. The first one is the training loss, which measures how well the network fits the training data. The second one is the validation loss, that measures the predictive performance on a small fraction of the data that is not used in the actual training for the given epoch. The loss function evaluated is the quantile loss function defined in eq. (13), summed over all τs used. The four different sizes of inputs described in section 3.1.4, mean that four separate networks need to be trained, mainly because of the different number of neurons in the input layer. Each of the networks are also optimised with regard to certain hyper-parameters separately, in order to get as fair a comparison as possible.

3.2.2 Structural parameters

Choosing structural parameters, i.e. the depth and width of a network, is generally not straight forward. A larger network is often better at fitting the training data while, on the other hand, too large network can suffer from over-fitting, that is, adapting too much to the noise in the data and thus generalising poorly. In search of the best configuration of structural parameters for each network, a rough search is performed to find a grid of parameters that seem suitable. The search is then focused on the grid, which is however rather coarse, to find the best combination of width and depth, primarily with respect to the validation loss. The grid considered is $\{6, 8, 10\} \times \{256, 350, 400\}$ of depths and widths respectively for each network. The type of activation function for each layer in the network is also a hyper-parameter that can be optimised. To limit the search space, the rectified linear unit (ReLU) (Li and Yuan, 2017), which is usually considered as the standard, is used for all layers and networks throughout this study.

3.2.3 Training parameters

For minimising the loss function during training, stochastic batch gradient descent is used with a decreasing learning rate. When the validation loss hasn't decreased for a given number of epochs, the learning rate is decreased down to a fixed value at which the training ends. These four hyper-parameters - the initial learning rate, learning rate decay factor, minimum learning rate and the number of non-decreasing validation loss epochs - are kept fixed at 0.001, 1.5, 0.0001, and 10 respectively throughout this study. This is done in order to reduce the parameter space. The same goes for the batch size, which is kept at 128. Furthermore, it has been shown

that the training parameter's overall effects on the performance of the QRNNs are quite small (Pfreundschuh et al., 2018). Another feature implemented in the typhon package is the option to use adversarial training, which is a technique to augment the training set. Additional artificial training samples are generated, such that they are hard for the network to predict. Existing training samples are slightly perturbed to maximise the change in the loss function. The strength of the perturbations, δ_{adv}, is an additional hyper-parameter of the network which is however kept fixed at 0.2.

3.3 Evaluation metrics

When comparing different predictors, there's not a trivial single metric of their performance. For a more conclusive assessment of the predictors' performances, a collection of evaluation metrics are used. Some of them are concerned with assessing the predictor's ability to estimate point values whereas other metrics deal with its capacity to estimate the uncertainty of the predictions. Throughout the evaluation, the median of the posterior distribution predicted by the QRNNs are used as a point estimates. All of the evaluations are performed on a test set with N samples. Let y be the label, or ground truth value, and y_{pred} the prediction.

3.3.1 Point estimate metrics

One of the most commonly used metrics is the Mean Squared Error (MSE), defined by

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_{\text{pred},i} - y_i)^2. \tag{14}$$

The MSE gives a single value, representing the predictor's squared error compared to the true value, averaged over the entire test set. A similar metric is the Mean Absolute Error (MAE). As the name implies, it deals with absolute errors rather than squared, and is defined by

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|y_i - y_{\text{pred},i}|. \tag{15}$$

Here, the MAE is only used on test samples where $y_i > 0$. This is due to the fact that a large majority of the test set is comprised of non-raining footprints where the errors are usually very small.

In addition the distributions of the prediction errors, $y_{\text{pred},i} - y_i$, are shown for a more visual assessment of the predictors.

3.3.2 Rain-no rain classification metrics

Another metric used is the confusion matrix. While MSE and MAE deal with scalar value predictions, the confusion matrix is a way to assess classifications. In this case, predictions can be turned into binary classifications using a rain-no rain threshold, that is, some value R such that if $y_{\text{pred}} < R$ it's classified as no rain and vice versa. The confusion matrix for a binary classification has four fields as shown in table 2.

The numbers a and b represent the number of test cases where the correct label is rain and the predictor gets it right and wrong respectively. The second row similarly shows the number of correct and wrong classifications when the correct label is no rain.

Table 2: Confusion matrix for a binary rain-no rain classifier. Here labels refer to the correct classes of the test set. Diagonal elements a and d correspond to the number of correct classifications.

		Predictions	
		Rain	No rain
Labels	Rain	a	b
	No rain	c	d

The confusion matrix gives a measure of how careful or confident a predictor is. What is preferred generally depends on the application. As a and b as well as c and d are each others complements, only the fractions $a/(a+b)$ and $d/(c+d)$ are presented, referred to as fractions true positives and true negatives respectively. Another classification metric, based on the confusion matrix, is the Heidke Skill Score (HSS), that is defined as

$$\text{HSS} = \frac{2(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)}. \tag{16}$$

The HSS compresses all the information from the confusion matrix to a single number that can take values from $-\infty$ to 1. All values above 0 indicate that the predictor is better than completely random classification and the higher the better.

3.3.3 Error estimation metrics

For assessing the error estimation, two graphical methods are used. The first one is the calibration plot. For a given quantile y_τ, the number of rain rates with $y_i < y_\tau$ in the test set are counted. The fraction of such samples of the test set is then plotted on the y-axis against the corresponding τ on the x-axis. This is done for a number of quantiles, and the plot shows how well the predicted quantiles correspond to the actual fractions of the test set. The closer the curves are to the $y = x$ line, the better calibrated the QRNNs are. An example of a calibration plot is shown in fig. 10. Two different QRNNs are trained to predict outputs $y = x^3$ with some perturbations. The calibration plot shows that the first QRNN (blue) predicts too small quantiles. For each quantile, the observed fraction of test samples below it is smaller than the predicted one. For the second QRNN, on the other hand, all observed fractions are above the predicted fractions, indicating that they're too large.

Another way to evaluate the predicted error distribution is to sample the predicted posterior distributions. For each sample of the test set, the quantiles are predicted. The posterior distribution is then approximated from the quantiles and a sample is drawn from it. The observed error distribution, $(y_{\text{pred},i} - y_i)$, is then compared to the predicted error distribution, $(y_{\text{sampled},i} - y_{\text{pred},i})$, for $i = 1, ..., N$. This gives another

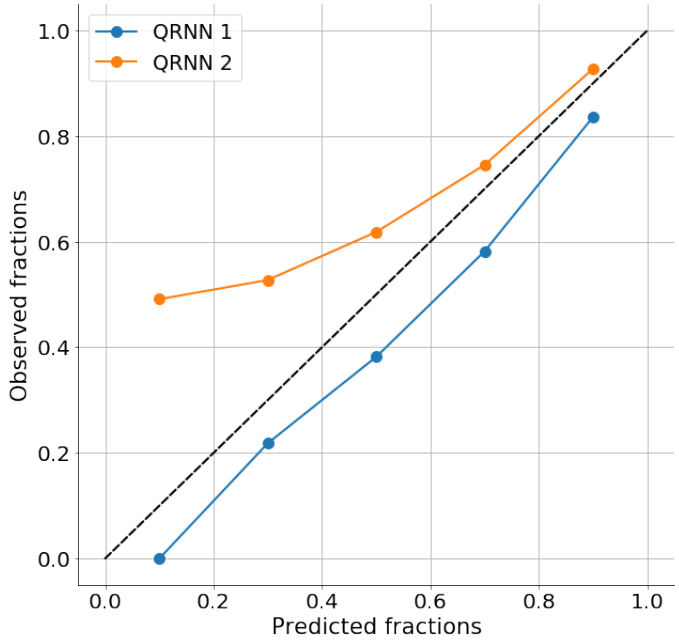

Figure 10: Example of a calibration plot for an artificial test case. The first QRNN (blue) is underestimating its quantiles while the second (orange) is overestimating them.

graphical metric to assess how well the predicted posterior captures the observed errors.

One of the benefits of the quantiles is that confidence intervals can be easily constructed. For a given retrieval, two quantiles are chosen, y_{τ_1} and y_{τ_2}. This means that there's a $\tau_2 - \tau_1$ chance of the correct value being between the two quantiles. For example, 80% confidence intervals can be constructed by taking the 0.1th and 0.9th quantiles. These confidence intervals are very similar to the ones considered in the calibration plot, but with both a lower bounds and upper bound. The length and accuracy, that is whether it contains the true value, of such confidence intervals are investigated.

4 Results

Here, all results for QRNNs trained to predict rain rates over oceans and land are presented. Both sections start with some example scenes where the predictions are shown on a map and compared to each other to give some intuition for how the predictions work and which trends might appear. These are followed by some more rigorous, quantitative results to show which trends can substantiated.

4.1 Predictions over the ocean

The QRNNs trained to predict precipitation rates over the ocean uses a training database of approximately 1.5 million matched DPR and GMI footprints. The optimal configurations for each QRNN with regard to depth and width that were found after a parameter search are shown in table 3. It turns out that with fewer neighbours, the QRNNs need to be somewhat larger. For the 0 neighbour QRNN, a larger network is necessary to learn the mapping of the training set. In general, for a fixed training set size, the smaller the network, the easier it is to avoid overfitting. The QRNNs with more neighbours are more efficient in decreasing the training loss for smaller networks and can therefore achieve a smaller validation loss as well. It's not obvious why the larger inputs have this effect, but it's plausible that more information helps the QRNNs to separate different cases without having to use larger networks.

All of the statistical results are derived using a testing database of approximately 780 thousand samples. The quantiles predicted by the QRNNs are x_τ for $\tau = 0.1$, 0.2, 0.33, 0.4, 0.5, 0.66, 0.8 and 0.9. The 0.33rd and 0.66th quantiles are included since these are the only available GPROF quantiles.

Table 3: Optimal width and depth configurations for QRNNs with different neighbours.

Number of neighbours	**Depth**	**Width**
0	10	350
4	8	400
8	8	400
12	8	256

Before moving on to the results, some comments about the databases, and their effects on the comparisons with GPROF, are necessary. The QRNNs are trained on rain rates derived directly from the DPR. These are also the rain rates that the testing database is built upon. Over ocean, the a priori database that GPROF uses have rain rates from a combined DPR and GMI precipitation product (GPM Science Team, 2018). This means that the rain rates in the testing database, regarded as the true values, favour the QRNNs in the results that are presented. Therefore, comparisons between GPROF and the QRNNs should be regarded carefully. QRNNs performing better doesn't mean that that is necessarily the case in reality, but rather points towards that the results are comparable. To give some idea of how the

combined precipitation product and DPR rain rates differ, approximately 1 million footprints are compared. The distribution of rain rates are shown in fig. 11. Overall the differences aren't very big, but the DPR has some more frequency for smaller rain rates. It turns out that both products have close to the same fractions of footprints with rain over the 0.01 mm/h threshold, 0.0614 and 0.0595 for the combined and DPR respectively.

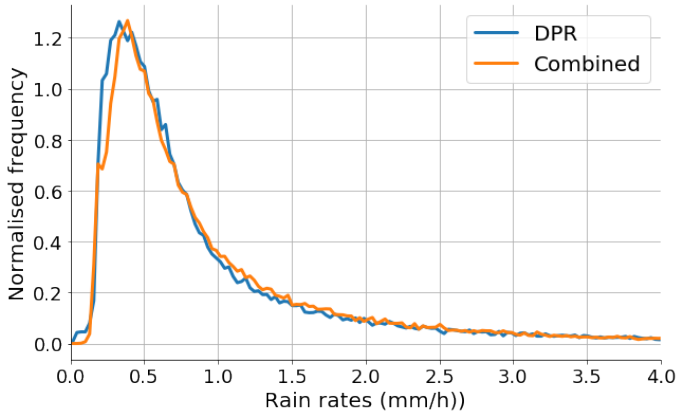

Figure 11: Rain rate distributions for the combined DPR and GMI product and the DPR product.

The distribution of differences between the product's rain rates are shown in fig. 12 for all footprints in the left panel and for rain rates above the threshold in the right panel. It appears that overall, the differences are very small, most within 0.5×10^{-4} mm/h. For footprints with rain, there are a bit more differences ranging from about -2 to 2 mm/h.

These differences between the rain rates used by GPROF and the QRNNs should be kept in mind throughout the results regarding oceans, introducing another source of uncertainty in the comparisons.

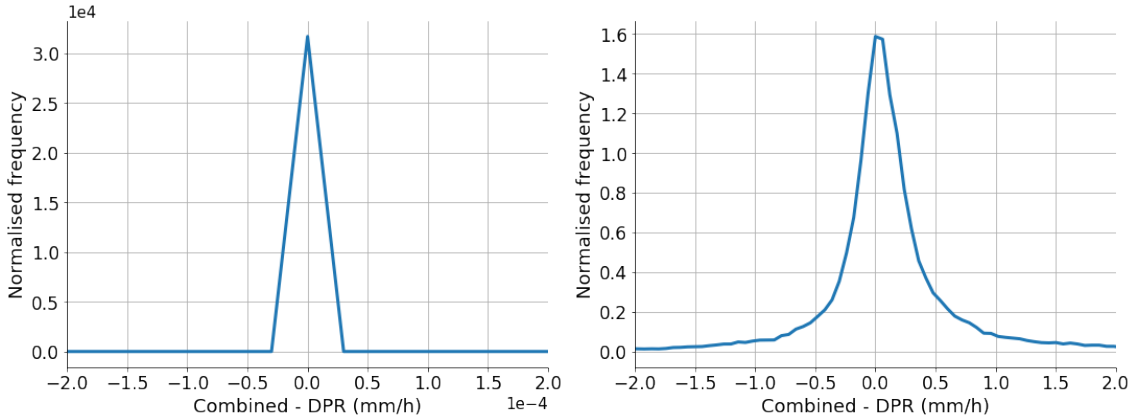

Figure 12: Distribution of differences in rain rates between the combined GMI and DPR product and the DPR product.

4.1.1 Example scenes

The first scene is from the Pacific Ocean and features a rather large area of moderate rain, where the DPR gives rain rates up to 4 mm/h, and a smaller area of light rain. In fig. 13, the predictions of the 8 neighbour QRNN and GPROF are plotted along with the DPR rain rates. Overall, both the QRNN and GPROF detect most of the areas where it's raining and manages to get the values close, although the QRNN tends to overestimate slightly in the region of most intense rain. It appears, however, that GPROF makes a lot of false positives in the sense that in areas where the DPR rain rates are below the threshold (white in the figure), GPROF predicts values way above it. At some places even rain rates up to 0.5 mm/h. This might be a consequence of the differences in databases used though. The QRNN, on the other hand, manages to correctly predict the absence of rain for a majority of the footprints. Furthermore, the quantiles work well by providing lower and upper bounds for the rain rates. Here the 0.1th and 0.9th quantiles together represent an 80% confidence interval.

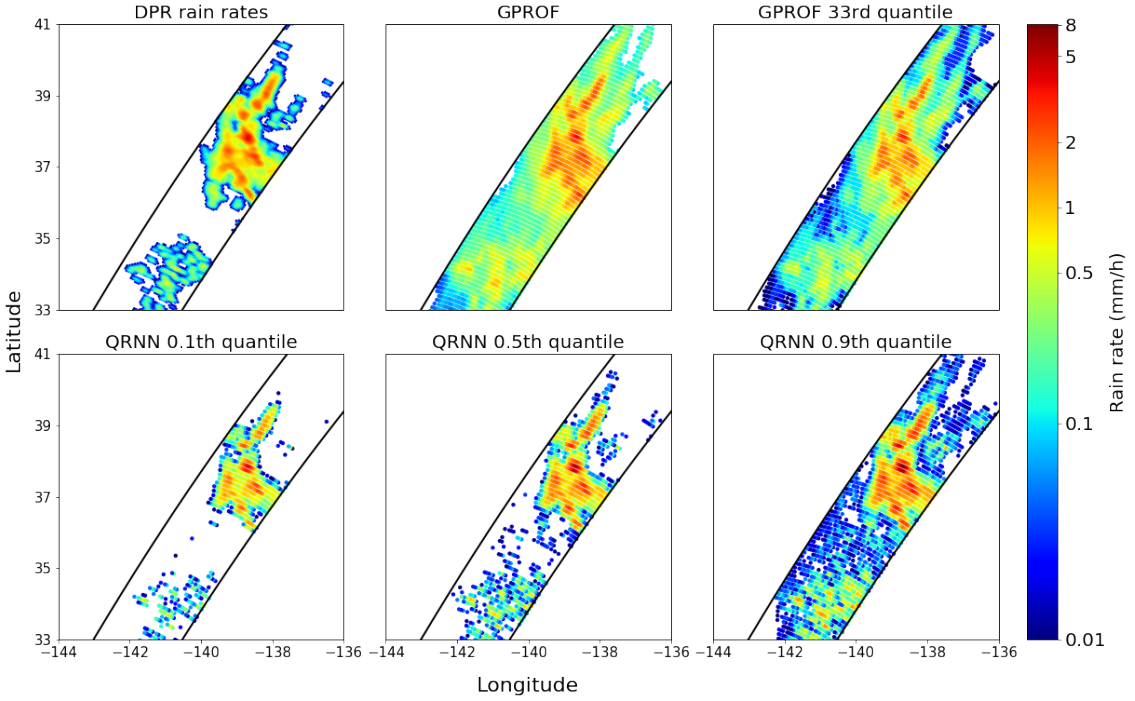

Figure 13: An example scene from the Pacific Ocean with DPR rain rates (left), GPROF point estimates (middle) and GPROF's first tercile (right) in the upper panel. In the bottom panel the 8 neighbour QRNN predictions corresponding to the 0.1th quantile (left), the 0.5th quantile (middle) and the 0.9th quantile (right) are plotted. Only rain rates above 0.01 mm/h are plotted. Black lines indicate the boundaries of the DPR swath. Data are taken from the 1st of January, 2018, 01:09.

The next scene is taken off the west coast of India. Here, the DPR rain rates are a lot higher, ranging up to about 25 mm/h. Additionally, the scene features some rain very close to and on land which should cause the QRNNs some problem. It can be seen, however, that the 8 neighbour QRNN manages to deal quite well

with the rain close to the coastline. On land it doesn't predict much rain while there appears to be some confusion about the rain on the coastline, where very light rain is predicted all along it. GPROF has some problem with the coastline as well, although doing much better and gets some of the rain on land, which is not surprising as it is developed to deal with land as well. For the rest of the scene that covers the ocean, some of the tendencies from the previous scene reappears, such as GPROF predicting rain in some areas where it shouldn't, although not as much as in fig. 13.

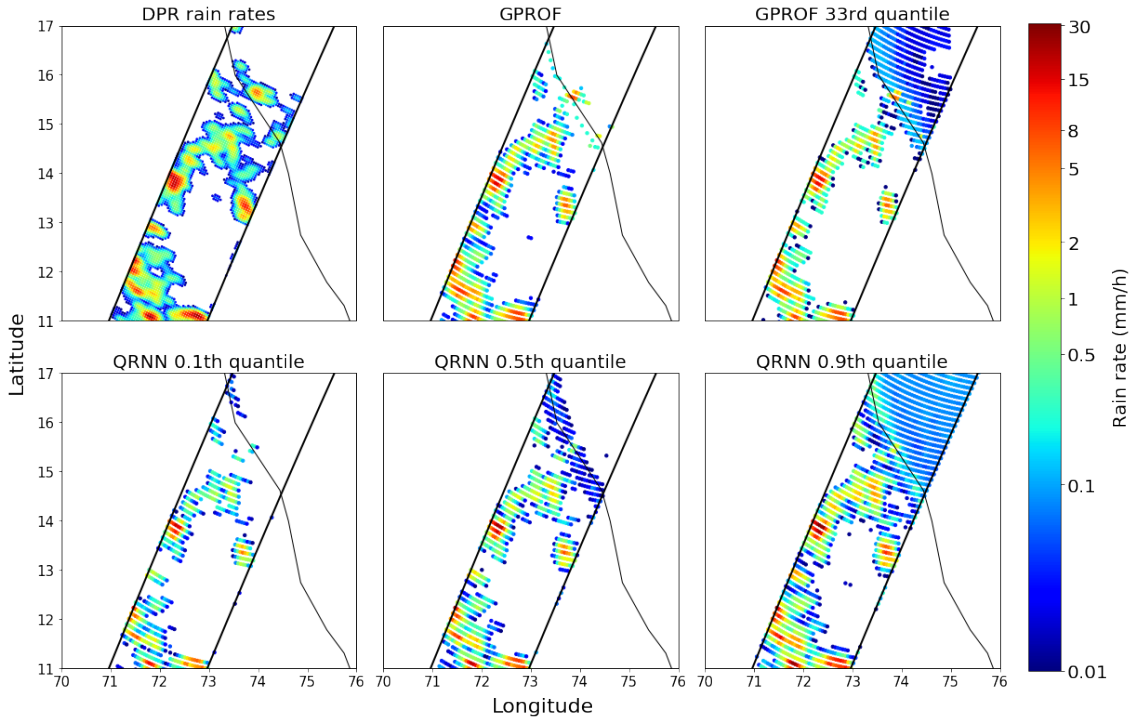

Figure 14: An example scene off the coast of India with some land included (in the upper right corner of the panels). Again, DPR rain rates and GPROF prediction are in the upper panels and different 8 neighbour QRNN quantiles in the lower panels. Data are from the 1st of July 2018, 05:13.

To investigate how the QRNNs with different number of neighbours deal with this coastal environment, predictions on the same scene are shown in fig. 15 using the 0, 4 and 12 neighbour QRNNs respectively. All of them make reasonable prediction over the ocean while it's obvious that the 0 and 4 neighbour QRNNs handle the coastline very poorly. The 12 neighbour QRNN, on the other hand, manages much better. Contrary to the 8 neighbour QRNN, it predicts a lot of light rain on land and nothing along the coastline. Observation from this scene point towards that QRNNs with more spatial information manage to deal with coastal areas better. Apart from a few other similar scenes, there are however no further quantitative results to substantiate this trend.

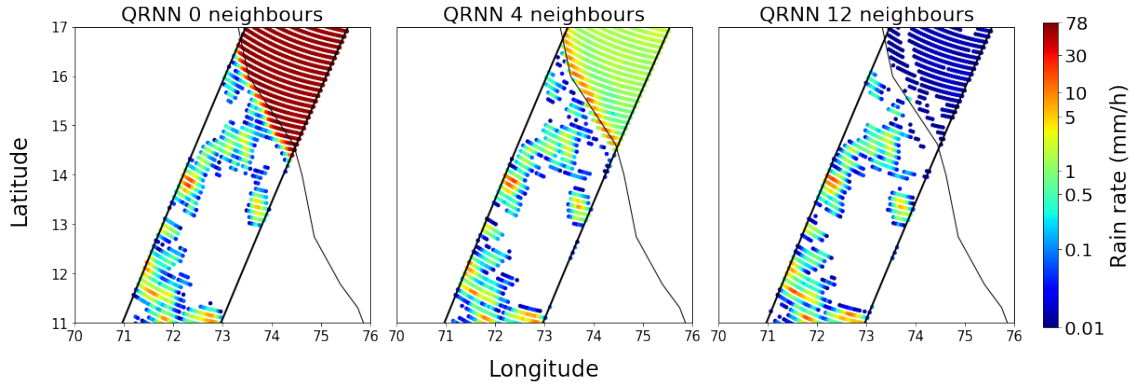

Figure 15: The same scene as in fig. 14 showing the 0.5th quantile predictions from the 0 neighbour QRNN (left), 4 neighbour QRNN (middle) and 12 neighbour QRNN (right).

4.1.2 Quantitative results

To assess the performance of the QRNNs in a statistical way, the metrics described in section 3.3 are used on the testing database. First, to find a suitable rain threshold, GPROF's Probability of Precipitation (PoP) is used. For a given percentage p, all the footprints in the test set with PoP $= p$ are picked out. Then the actual frequency of rain is calculated for this subset. A number of ps are investigated for different thresholds until one is found such that GPROF's PoP is in line with the observed frequencies. As an attempt to get a similar measure for the QRNNs, for each sample in a subset the smallest quantile x_τ above the rain threshold is found. To plot the PoP for the QRNNs, the average of these τs is taken for each p. It appears that 0.01 mm/h is the optimal threshold for GPROF which is in line with the GPROF documentation (GPM Science Team, 2018). The resulting predicted percentages for the 0.01 mm/h threshold are plotted against the observed in the left panel of fig. 16. It can be seen that both curves are close to the $y = x$ line which indicates that the predicted percentages agree with the observed. It should be noted that while GPROF's PoP only works for this fixed threshold, the QRNNs tend to make good PoP predictions for other thresholds as well, as can be seen in the right panel of fig. 16 where the threshold is 0.1 mm/h.

The Mean Squared Error (MSE) and Mean Average Error (MAE) are shown in fig. 17. With respect to the MSE, all the QRNNs perform better than GPROF except the 12 neighbour QRNN which has an MSE slightly above GPROF. Furthermore, the MSE increases with the number of neighbours. The MSE is calculated on the entire test set which is dominated by footprints without rain. To get an idea of how the predictors perform where it's raining, the MAE is calculated on a subset of footprints where the rain rates exceeds the 0.01 mm/h threshold. It appears that the QRNNs with neighbours perform better in this regard compared to both GPROF and the QRNN without neighbours. However, adding more than 4 neighbours only decreases the MAE by about 0.01 mm/h.

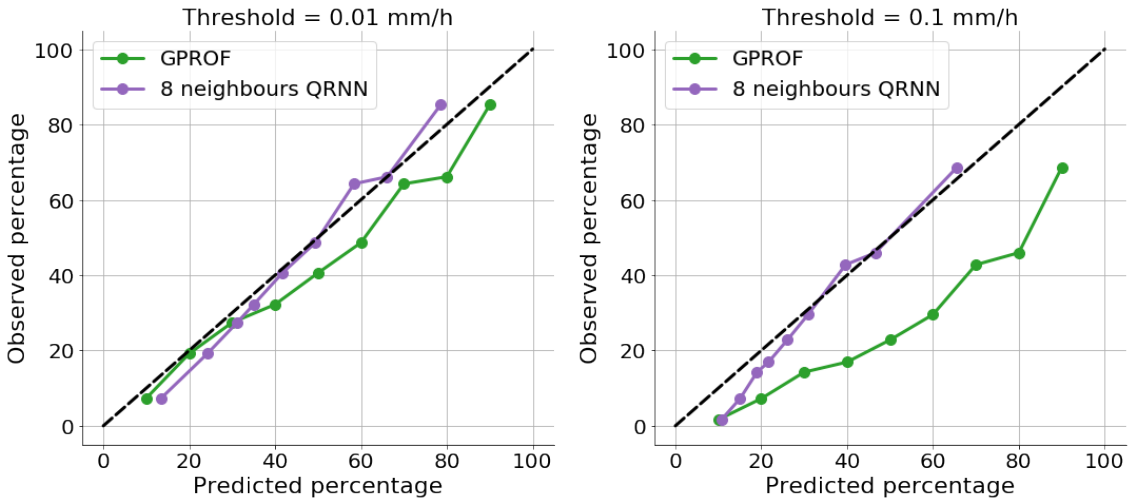

Figure 16: GPROF's Probability of Precipitation (PoP) plotted along with a similar derived measure for QRNNs against the observed frequencies of the test set for thresholds 0.01 mm/h (left panel) and 0.1 mm/h (right panel) respectively.

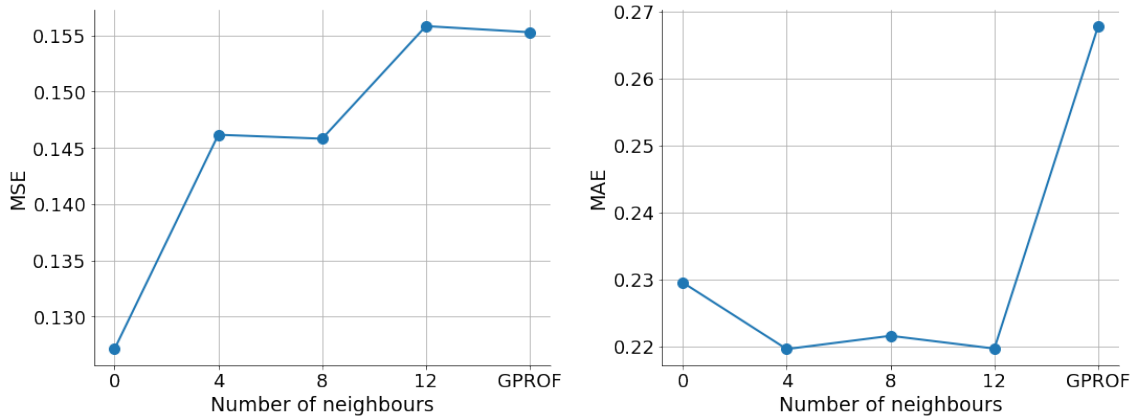

Figure 17: The Mean Squared Error (MSE) over the entire test set (left panel) and the Mean Average Error (MAE) on the subset of footprints with rain rates > 0.01 mm/h (right panel), plotted against the different QRNNs and GPROF.

As the MSE and MAE are averaged over large data sets, it's hard to say much more about the performances. To get a bit more intuition, the distribution of errors are shown in fig. 18. When making point estimates, the GPROF algorithm uses the probability of precipitation to set predictions to 0 mm/h if the probability doesn't reach a certain threshold, dependent on the location. To make the predictions more easily comparable, the QRNN point estimates are also set to 0 mm/h if the 0.9th quantile is below the rain threshold, that is if the probability of rain is less than 10%. The left panel of fig. 18 shows the error distributions on the entire test set. As mentioned, a large majority of footprints has no rain and the predictors generally handle these cases very well. Therefore the errors are tightly centred around 0 mm/h and there's not much to separate their performances. In the right panel of fig. 18, the error distributions only on footprints with rain are shown. Here, some

differences appear. All predictors have a peak to the left of 0 mm/h, meaning that they underestimate the rain rates slightly. The QRNNs, however, have a higher frequency at 0 mm/h than GPROF and the ones with neighbours even more than the QRNN without. It can be seen that for GPROF, this is caused by the frequency of positive errors being slightly higher, that is, it has a tendency of overestimating. This might again be an effect of the differences in databases shown in fig. 11, where the GPROF database has a peak at slightly higher rain rates.

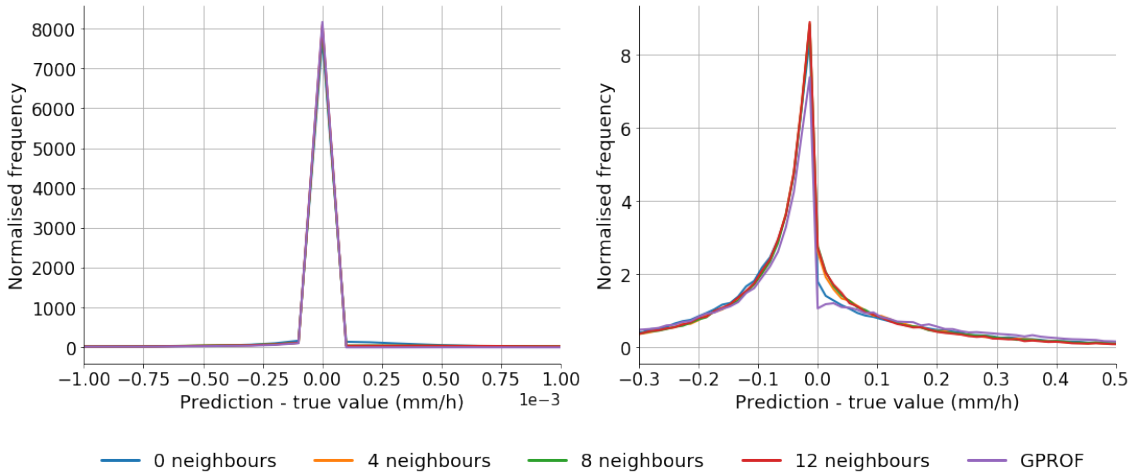

Figure 18: The error distributions for all QRNNs and GPROF on the entire test set (left panel) and only on rain footprints (right panel).

To get an idea of how the rain rates are distributed globally, the test set is divided into 1.5° sized latitude bins and the average rain rate is calculated for each bin. The results for the 0 and 12 neighbour QRNNs and GPROF are shown in fig. 19 with the DPR rain rates as reference. It can be seen that, in general, the QRNNs and GPROF show similar rain rates as the DPR for most latitudes. One exception is the peak around 10°, where the DPR shows higher rain rates than the QRNNs and GPROF. Another is the latitude bins above 50° where GPROF predicts much higher rain rates in some of the bins.

The rain-no rain classification results are shown in fig. 20. For the derived threshold of 0.01 mm/h, the benefits of adding neighbours to the input of the QRNNs are clearly visible. Most notable are the effects on the fraction of true positives, where the QRNNs with neighbours do a lot better, with fractions up to 0.73, than the QRNN without as well as GPROF having fractions around 0.6. With respect to the true negatives, QRNNs with neighbours only perform slightly better than without, having fractions of 0.96 and 0.97 respectively. These are, however, much better than GPROFs fraction around 0.91. The improvements are also visible in the Heidke Skill Score (HSS), and follow the same pattern, which is expected as the HSS is derived from the other two metrics. It can be seen however, that overall there's not as clear an improvement of adding more than 4 neighbours as the increase in HSS from 4 to 12 neighbours is roughly 0.01. Changing the rain threshold to 0.1 mm/h pushes all curves up but doesn't change the relation between the performances all that much, except for the true positives where the differences become slightly

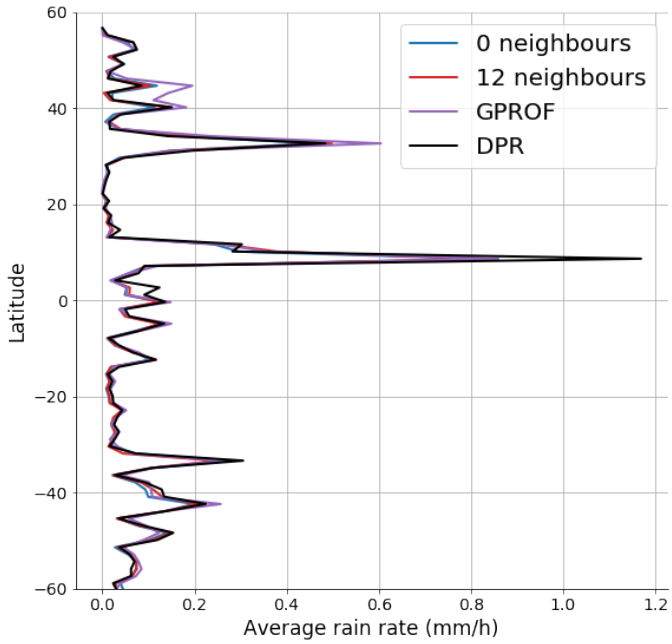

Figure 19: The average rain rates produced by the 0 and 12 neighbour QRNNs and GPROF for 1.5° sized latitude bins ranging from −60° to 60°. The DPR rain rates (black) are included as well.

smaller.

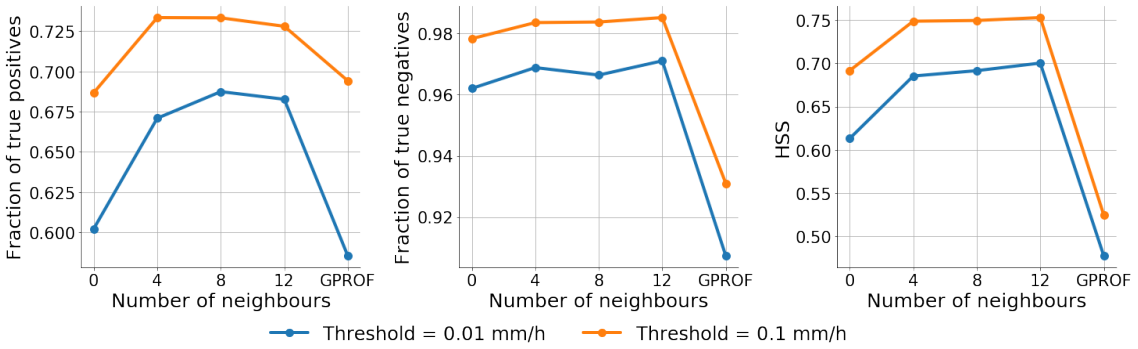

Figure 20: The results with respect to some classification metrics are plotted against the different QRNNs and GPROF - fraction of true positives (left), fraction of true negatives (middle) and HSS (right) for thresholds of 0.01 mmh/h and 0.1 mm/h respectively.

The rest of the results deal with the uncertainties of predictions and the predictor's abilities to account for them. For more clear plots, the results are presented only for GPROF and two different QRNNs, with 0 and 8 neighbours respectively. Generally, the differences are quite small between the 4, 8 and 12 neighbour QRNNs.

Figure 21 shows calibration plots for the two QRNNs and GPROF. In the left panel, the entire test set is used. It shows that both QRNNs are fairly well calibrated as the curves roughly follow the $y = x$ line, except for the 0.66th quantile of the QRNN without neighbours which is a little off. For GPROF, only the 0.33rd and

0.66th quantiles are available. The calibration plot shows that these are very poorly calibrated. Both points are below the $y = x$ line which indicates that the quantiles are too small in the sense that not a big enough fraction of the test samples are below the predicted quantiles. The right panel of fig. 21 shows the calibration on footprints with rain rates above 0.01 mm/h. The two cases are displayed to illustrate how the predicted quantiles of the QRNNs are dependent on the a priori distribution of rain rates. As expected, the calibration in the right panel is poor. When taking a subset of the test set, such as when introducing a threshold, the test set distribution deviates from the training set a priori distribution and the calibrations of both the QRNNs and GPROF are therefore notably worse.

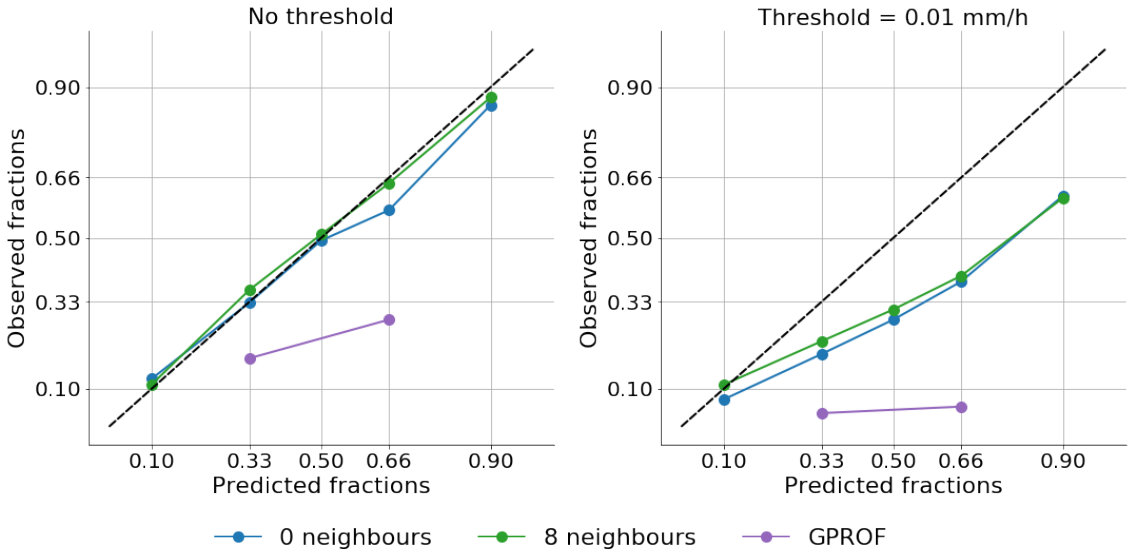

Figure 21: Calibration plot including the 0 and 8 neighbour QRNNs as well as GPROF for the entire test set (left panel) and on footprints with rain rates above 0.01 mm/h (right panel) respectively. The $y = x$ line is plotted as reference.

As mentioned before, a way to use the predicted quantiles is to construct confidence intervals. To compare with GPROF, the 0.33rd and 0.66th quantiles are used to construct 33% confidence intervals. The smaller the confidence interval, the more certain the predictor is about the estimate. To get an idea of how the predictors construct their confidence intervals, the distribution of interval lengths are show in the left panel of fig. 22. These are calculated on the subset of the test set where rain rates exceed the 0.01 mm/h threshold. The threshold is used here even though the calibration is bad for such cases, as shown in fig. 21. The reason is primarily that when using the entire test set, the non-raining footprints dominate heavily and the plots become hard to read.

In itself, the confidence interval lengths don't tell the entire story. To get more information, the interval lengths are divided into 0.05 mm/h sized bins. For each bin, the number of cases where the confidence intervals are correct are counted. Here, being correct means that the DPR rain rate is observed to be within its corresponding predicted confidence interval. The fraction of correct cases are plotted against the confidence interval lengths in the right panel of fig. 22.

All the predictors appear to have a high frequency of lengths close to 0 mm/h, indicating that a lot of the time they are very certain. Looking at the fraction of correct intervals in the first bin, it seems that the QRNNs have a good reason for it as they're correct about $30 - 35\%$ of the time, which corresponds well with the 33% confidence level. Optimally, the fractions of correct intervals would then keep a straight line around 0.33 as the interval lengths increase. It seems, however, that the QRNNs fail a bit to account for the uncertainties properly as the curve goes down to values between 0.2 and 0.25, indicating that the confidence intervals are too small. The 0 neighbours QRNN does a bit better than the 8 neighbours QRNN, with its peak at larger intervals. Overall, though, the confidence intervals work fairly well for the QRNNs, especially considering that these results are for raining footprints only, where the calibration isn't supposed to be perfect.

For GPROF, the confidence intervals are much less effective, which isn't strange considering the poor calibration shown in fig. 21. The fractions of correct intervals are very small for the small intervals. It's not until the confidence intervals are around 0.3 mm/h that GPROF reaches accuracies close to the QRNNs. These cases, however, account for a very small part of the test set. It's worth mentioning though, that GPROF's quantiles aren't as tightly linked with the point estimates as for the QRNNs. GPROF point estimates are weighted mean values of the selected subset of the a priori database, which doesn't mean that the point estimates need to be within the confidence intervals. In fact, it turns out that only about 30% of the test samples have the point estimate between the 0.33rd and 0.66th quantiles. This explains why the point estimates can show good results while the quantiles don't.

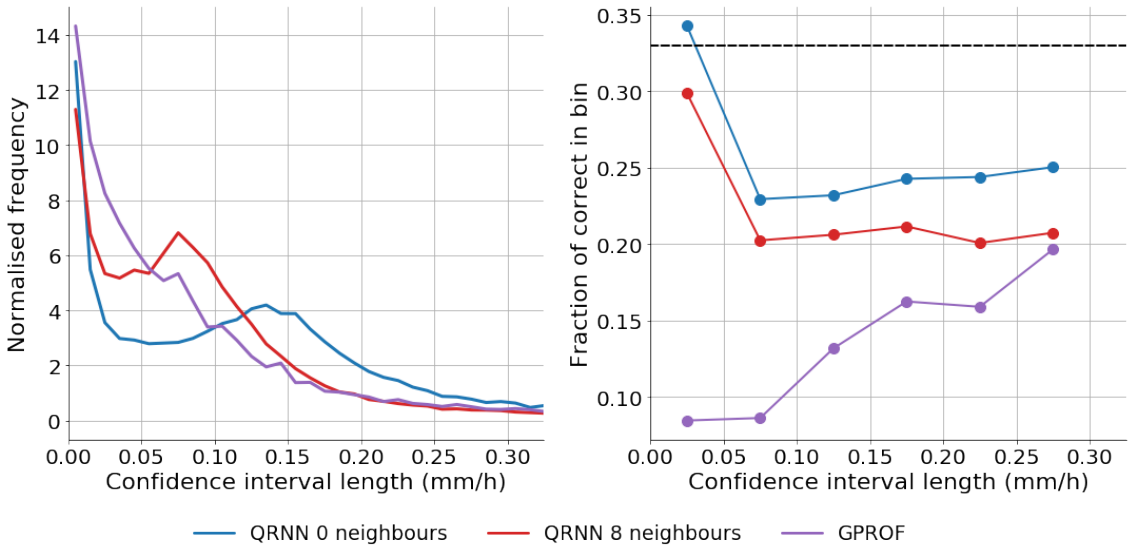

Figure 22: The distribution of confidence interval lengths (left panel) and the fractions of intervals with the DPR rain rate contained per 0.05 mm/h bin (right panel). The 33% confidence intervals are constructed using the 0.33rd and 0.66th quantiles. Dashed line shows the expected 0.33 fraction. Only footprints with rain rates above 0.01 mm/h are considered.

To conclude the results regarding uncertainties and oceans in general, the predicted

error distributions are compared to the observed. The predicted error distribution is obtained by sampling the posterior for each test case and comparing it with the point estimate, i.e. the median. The results for the two QRNNs are shown in fig. 23. Here, point estimates are not cut off using the PoP as in fig. 18 which explains the more smooth distributions. The plots show that both the QRNN's sampled errors resemble the observed errors quite well, although predicting a bit more errors around 0 mm/h. It should be noted though that these errors are very small, in the order of 10^{-3}, well below the rain threshold.

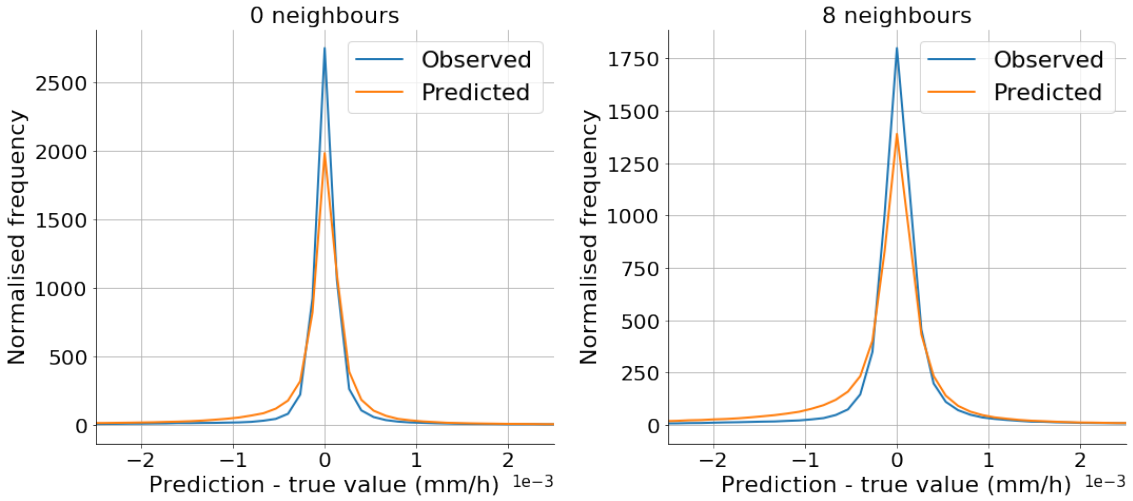

Figure 23: The observed error distribution along with predicted error distribution obtained by sampling the posteriors for the 0 neighbours QRNN (left panel) and the 8 neighbours QRNN (right). Contrary to the distributions shown in fig. 18, the rain rates are not cut off using the PoP.

4.2 Predictions over the Great Plains

The results for land precipitation retrievals follow the same outline as those for oceans with regards to example scenes and metrics. However, some results are skipped that aren't specific for land, such as the PoP and the calibration plot of subsets on the test set. Because of the much smaller area, building a large enough training and testing databases was more time consuming. As a consequence the training database is comprised of approximately 480 thousand matched GMI and DPR footprints and the testing database of 130 thousand. Overall, the optimal network architectures favours smaller depths and widths compared to over oceans, likely because of the smaller training database. The optimal configurations are shown in table 4. Similarly to the QRNNs developed for ocean retrievals, the QRNNs with more neighbours tend to favour somewhat smaller networks.

While the GPROF database over oceans uses the combined DPR and GMI product for its a priori rain rates, the database over land uses DPR Ku-Radar rain rates, i.e. the same as for the QRNNs. What complicates things slightly is that when the surface is snow-covered, GPROF uses the ground based radars, MRMS, instead of GPM data for its rain rates. Since parts of the Great Plains are sometimes snow-covered, this affects the results here as well in the QRNN's favour. No further

Table 4: Optimal width and depth configurations for QRNNs with different neighbours over land.

Number of neighbours	Depth	Width
0	6	400
4	6	350
8	6	256
12	6	256

analysis of the differences in rain rates used by GPROF and QRNNs are done here, but it should be noted that it adds some extra uncertainty to it.

4.2.1 Example scenes

Two scenes are picked out from the test set over the Great Plains. The first one, shown in fig. 24, features mostly light to moderate rain with an area of heavy rain. The QRNN with 8 neighbours is once again used, with its 0.1th, 0.5th and 0.9th quantiles plotted.

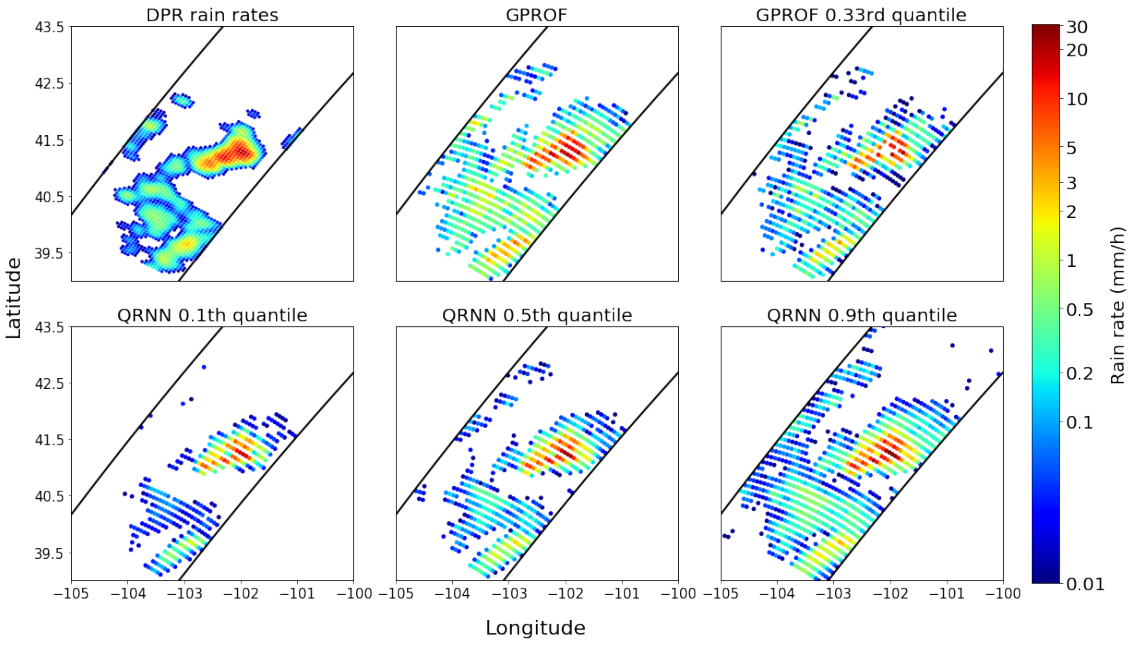

Figure 24: An example scene over the Great Plains with some moderate and heavy rain. The upper panel displays DPR rain rates (left), GPROF point estimates (middle) and GPROF 0.33rd quantile (right). In the lower panel the 0.1th quantiles (left), 0.5th quantiles (middle) and 0.9th quantiles of the 8 neighbour QRNN are shown. Data are from 31st of August, 2018, 23:54.

Similar to over oceans, GPROF tends to predict some rain where DPR doesn't show any while the QRNNs are a little bit more cautious. In this scene both the QRNN and GPROF tend to overestimate the rain rates with errors of about $4 - 5$

mm/h where the rain rates are maximal. Furthermore, the poor calibration of GPROF's quantiles is visible, especially looking at the centre of the heaviest rain where the 0.33rd quantiles are below 0.01 mm/h for some footprints.

The second example scene, shown in fig. 25, exhibits some extreme rain rates up to 126 mm/h in some areas. This scene is chosen to illustrate how such extreme cases are hard to capture properly for both GPROF and the QRNNs, predicting at most 52 and 43 mm/h respectively. In some areas, such as the centre of the most rain intensive region, neither of the predictors are close to the high rain rates. This is however not very strange considering that such extreme rain events are rare and doesn't appear very frequently in the training database. Once again, GPROF is overestimating the light rain footprints a bit more than the QRNN.

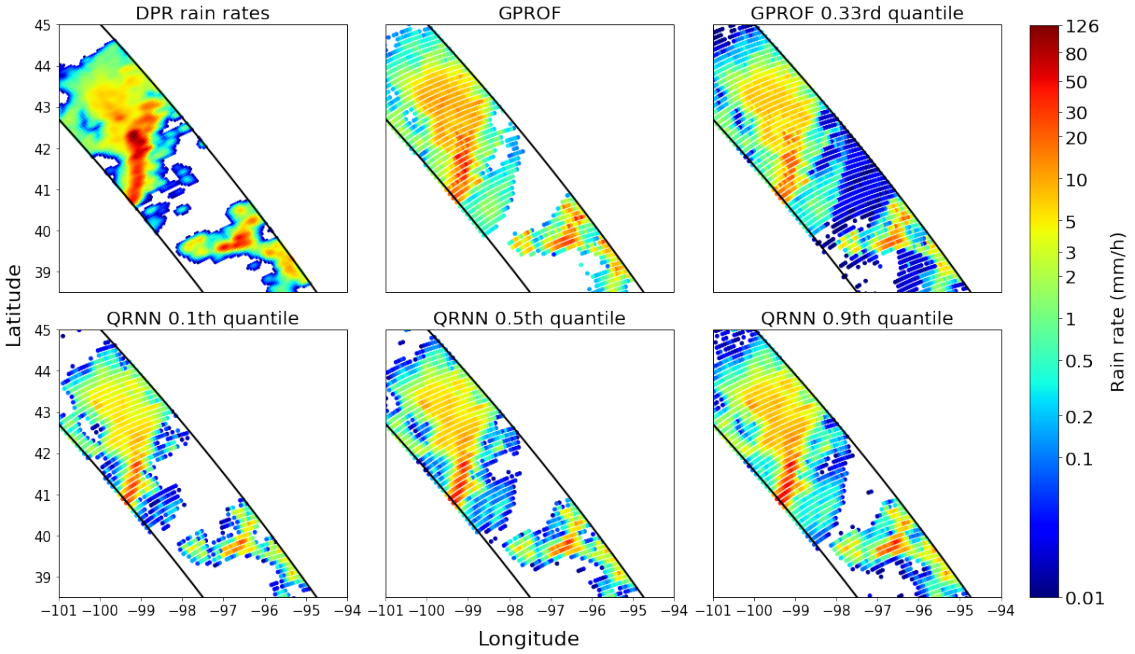

Figure 25: Another example scene from the Great Plains, similar to fig. 24, featuring some extreme rain rates. Data are from the 1st of July 2018, 02:08.

4.2.2 Quantitative results

Similar to section 4.1.2, the QRNNs and GPROF are assessed with regard to the metrics described in section 3.3. To start, fig. 26 shows the resulting MSE and MAE, with the latter on footprints above the 0.01 mm/h threshold. Contrary to over oceans, here GPROF is significantly better with respect to the MSE. Looking at the MAE, however, the QRNNs do better than GPROF. Furthermore, adding neighbouring footprints to the QRNNs doesn't improve the MSE or MAE.

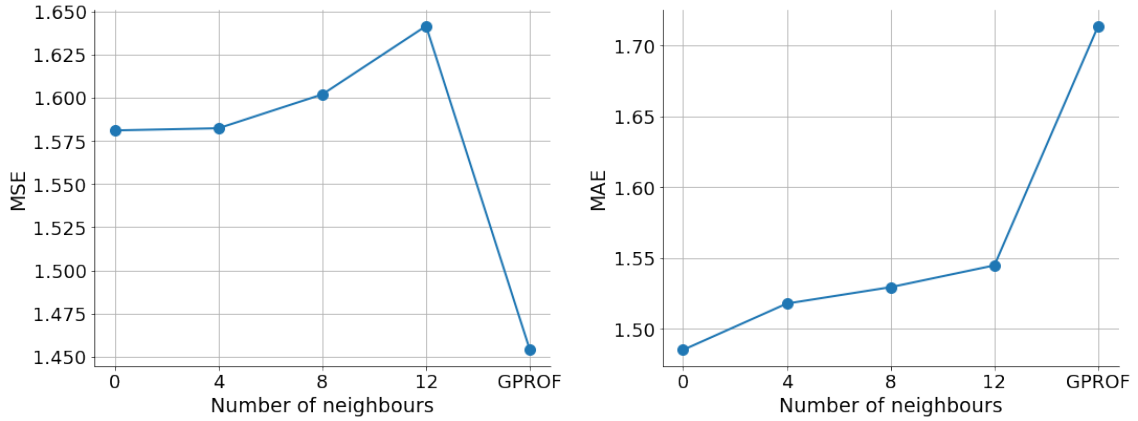

Figure 26: Mean Squared Error (left) and Mean Average Error on rain footprints (right) for the different QRNNs and GPROF.

Looking at the distribution of errors in fig. 27, some additional information about the errors is obtained. Once again, the distribution over the entire test set is very similar for all QRNNs and GPROF. The distribution over the subset of rain footprints shows some more significant differences. The peaks of all predictors are slightly below 0 mm/h, indicating tendencies of underestimating, however not by much. The QRNNs show a tendency to underestimate a lot more than GPROF though, with more frequency of negative values down to -0.5 mm/h. GPROF, on the other hand, shows much more frequent positive errors up to 0.5 mm/h.

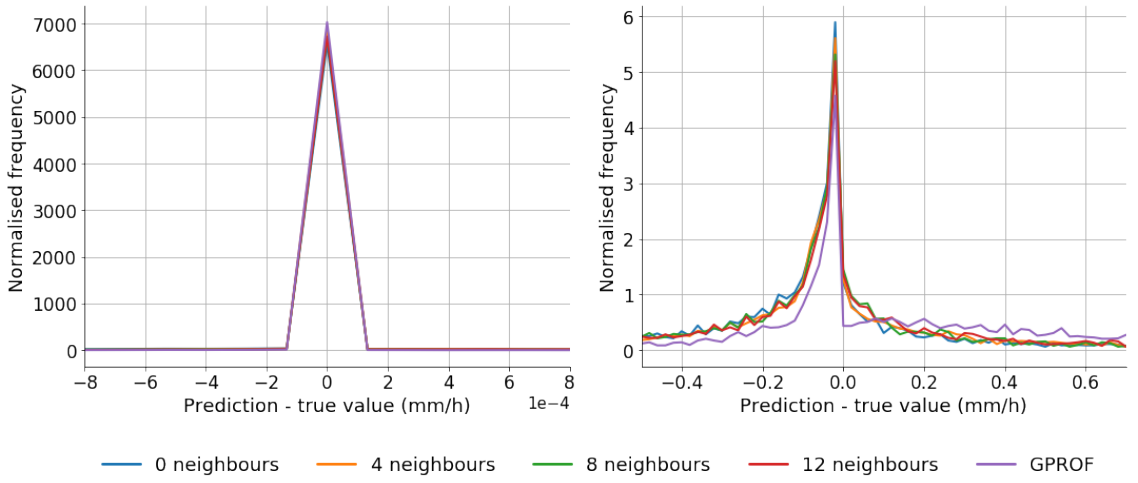

Figure 27: Error distribution, $y_{\text{pred},i} - y_i$, for GPROF and the QRNNs on the entire test set (left panel) and the subset of footprints with rain rates above 0.01 mm/h (right panel).

Turning to the rain-no rain classification results in fig. 28, it can be seen that GPROF is a lot better than the QRNNs with respect to the true positives with a fraction over 0.85 compared to the best QRNNs that is slightly below 0.65. Here, however, there's a clear benefit of adding neighbours to the QRNNs increasing the fraction by 0.05. This benefit isn't as significant, except for the 4 neighbour QRNN,

in the case of true negatives though, but all the QRNNs are somewhat better than GPROF. This is also reflected in the HSS, where the QRNNs with neighbours show the best performance. It should be noted that the much larger number of non-raining footprints favours the networks with highest true negatives score with respect to the HSS. Here a higher threshold of 0.1 mm/h seems to improve the performance of GPROF on the true positives while leaving the QRNNs almost unchanged. This might indicate that 0.01 mm/h isn't the threshold used by GPROF here. However, the calibration of Probability of Precipitation to find a suitable threshold fails over land because of the too small testing set, so this remains a hypothesis. On the true negatives the differences between them are almost the same with the entire curves pushed up.

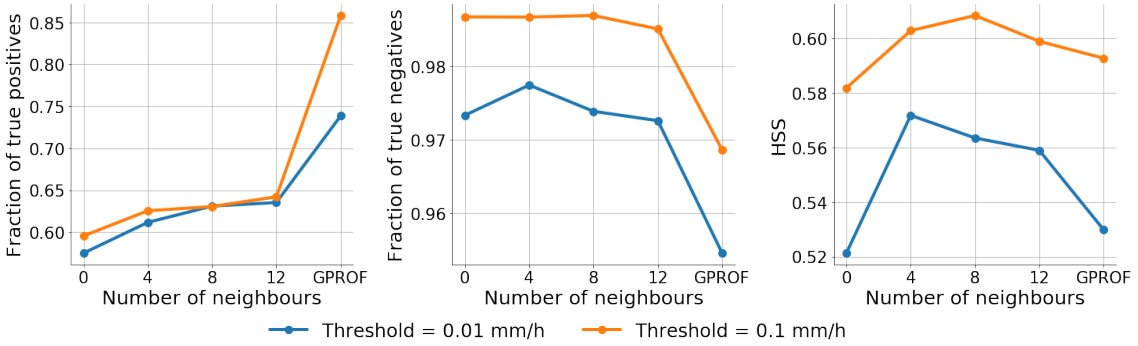

Figure 28: Classification results over land for GPROF and the QRNNs - fraction of true positives (left), true negatives (middle) and HSS (right) for thresholds of 0.01 mm/h and 0.1 mm/h respectively.

Similar to the results regarding uncertainty metrics over oceans, only the 0 and 8 neighbour QRNNs, along with GPROF, are show here. First is the calibration plot in fig. 29 which shows that the QRNN quantiles are well calibrated while the two quantiles of GPROF generally are too small.

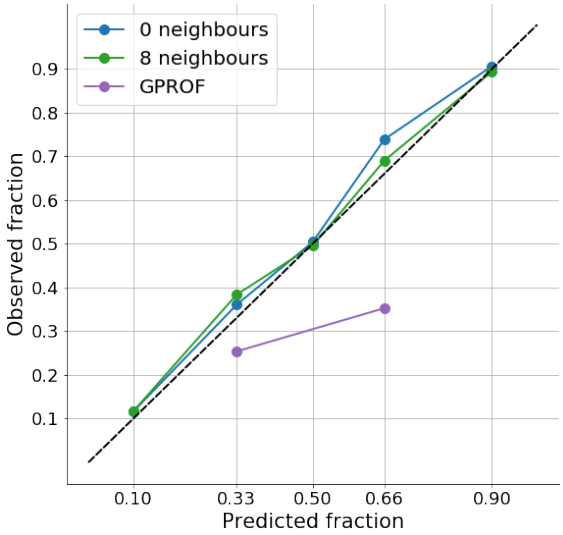

Figure 29: Calibration plot for the 0 and 4 neighbour QRNNs and GPROF over land.

Looking at how the predicted errors, $y_{\text{sampled},i} - y_{\text{pred},i}$ compare to the observed ones, $y_{\text{pred},i} - y_i$, in fig. 30, it can be seen that again the predicted errors resemble the observed ones rather well. The predicted errors tend to be a little less frequent at 0 mm/h with some more smaller errors. This confirms the good calibration seen in fig. 29.

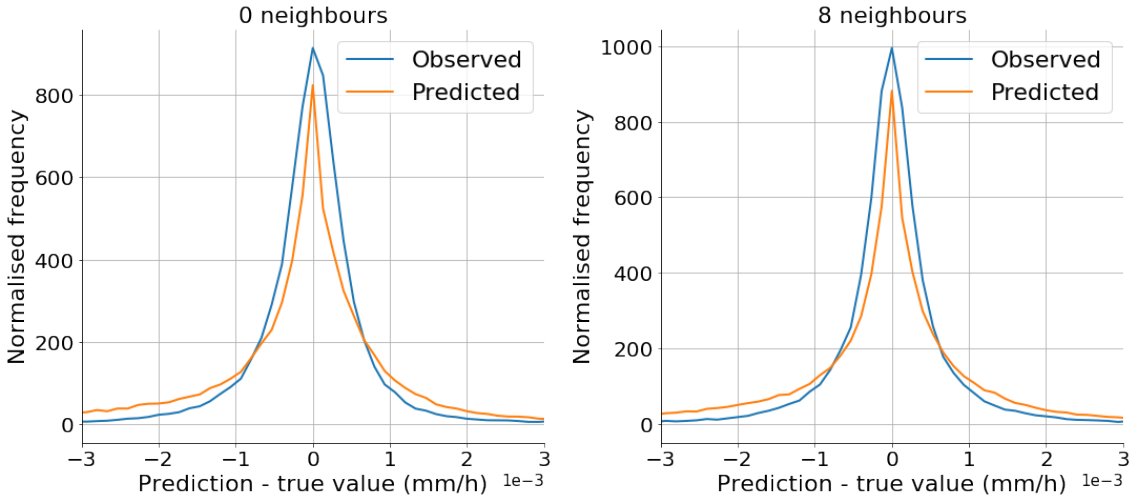

Figure 30: The distributions of observed errors, $y_{\text{pred},i} - y_i$, and predicted errors $y_{\text{sampled},i} - y_{\text{pred},i}$ for the 0 neighbours QRNN (left) and 8 neighbours QRNN (right).

Finally, the confidence intervals are assessed. It can be seen in the left panel of fig. 31, that the distributions of confidence interval lengths are quite similar for the QRNNs. GPROF produces confidence intervals that are very small more frequently. Similarly to the QRNNs trained on ocean retrievals, the QRNNs for land are correct in an appropriate fraction of the cases when they're the most certain, i.e. for the confidence intervals in the first bin, as can be seen in the right panel of fig. 31 where

the fraction of the 8 neighbour QRNN is around 0.33. As the interval lengths increase, the accuracy decreases a lot, down below 0.2, indicating that the uncertainties aren't properly accounted for by increasing the interval lengths enough. GPROF seems to suffer from the same errors as over oceans in the sense that the confidence intervals are too small overall, resulting in too small fractions of correctly constructed intervals. Most notable is the very small fraction, below 0.05, for the smallest lengths which are the most frequent. This indicates that GPROF is usually very certain but doesn't manage to construct intervals that contain the observed rain rates. But just as over oceans, it turns out that the GPROF point estimates are not always between the quantiles. This happens only in about 15% of the cases over land. So while the confidence intervals of GPROF show poor performance, its point estimates are often accurate anyway.

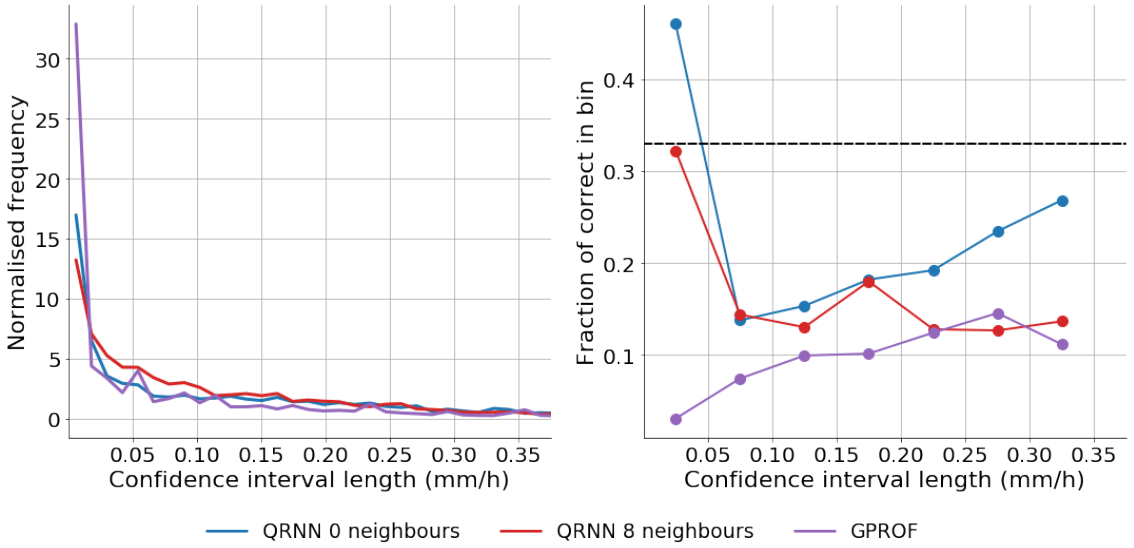

Figure 31: Distribution of confidence interval lengths (left panel) and fractions of confidence intervals containing the observed DPR rain rate per 0.05 mm/h bin (right panel). Dashed line indicates the optimal 0.33 fraction.

5 Conclusions

5.1 Results

Overall, the performance of the QRNNs are very good, especially over oceans. The results over the Great Plains are obtained using much less data, but are still promising. The size of the training database is most likely too small to see the full potential of the QRNNs and the even smaller testing database make the results less significant than those over oceans.

The trends that were hinted towards in the ocean example scenes, where the QRNNs seemed to accurately represent the rain rates of the DPR, are substantiated in the quantitative results. First in the sense that errors computed for point estimates across the entire testing database are quite small as shown by the MSE and the error distributions in fig. 17 and fig. 18. These include all the test cases where there's no rain, which account for the majority of the entire test set, which indicates that the QRNNs handle cases without rain well. This is further substantiated by the high fractions of true negatives shown in fig. 20, above 0.96 for all the QRNNs. The cases where rain is present seem to be a bit harder to predict, but the QRNNs perform well here too, although underestimating a bit.

Furthermore, the quantiles work out very well too as the calibration plot in fig. 21 shows, which was one of the main goals of the study to show. This is seen in the example scenes as well, where the 0.1th and 0.9th quantiles give appropriate bounds for the rain rates. The calibration of the Probability of Precipitation in fig. 16 is another example of the benefits of the quantiles. It shows that the QRNNs can deal with new rain threshold after retrievals are made. These results are more easily assessed in their own, in contrast to the actual rain rate predictions in the sense that it's hard to say exactly what constitutes small enough errors. For this reason GPROF was introduced as sort of a baseline for rain rate predictions. With GPROF being a state of the art passive microwave retrieval algorithm, having comparable results give some weight to the QRNN results.

As was mentioned before though, the comparisons with GPROF should be taken with a grain of salt. While the results often show the QRNNs performing better than GPROF, especially over oceans, it doesn't mean that it is necessarily the case. Such conclusions cannot be drawn from the results because of the uncertainties introduced by the different databases, and it's not the intention to do so either. With this said, though, the results overall point towards that there's reason to believe that the QRNNs are at least comparable to GPROF with regards to making rain rate point estimations. What's worth emphasising, though, is that the QRNNs seem to give better uncertainty estimates through their quantiles than GPROF does. These results are significant enough to suggest that this is the case, even considering the comparison uncertainties. Another benefit of the QRNNs is that retrievals are easily run on a standard desktop computer taking less than 0.1 milliseconds per footprint.

Turning to the comparisons between the QRNNs with different number of neighbours as input, there's not a single QRNN that outperforms the others with respect to all metrics. The QRNN without neighbours turns out to give the smallest MSE across the test set, but does worse considering the MAE on rain footprints and the error distributions don't separate them all that much. With respect to all of the

classification metrics, the QRNNs with neighbours do significantly better, especially for the true positives. While the calibration plot shows similar performance, there are differences in the confidence intervals. The QRNN without neighbours tends to make larger confidence intervals, showing more uncertainty, but gets them right more often than the other QRNNs. The example scene in fig. 15 points towards that the QRNNs with 8 or 12 neighbours are less sensitive to disturbances that could be caused by for example small island or proximity to the coast. All in all, it's hard to pick a clear winner, but the results seem to favour the QRNNs with neighbours slightly. Perhaps more training and testing data is needed to conclusively decide which works best.

5.2 Outlook

There are quite a few ways the QRNN retrievals could be improved. First, using the combined DPR and GMI precipitation product over oceans instead of only the DPR would make the comparison with GPROF more fair and would thus increase the relevance. Second, because the GMI channels of group 1 and 2 are not collocated, only the brightness temperatures from group 1 were used for retrievals. However, there is a GPM product where the group 1 and 2 footprints are matched. Using these footprints instead would give each footprints four additional brightness temperatures as input, which could positively affect the quality of predictions, especially considering that the 166 GHz channel is optimised specifically for detecting light precipitation in the tropics (Hou et al., 2014). A way to take care of both of the issues is to use GPROF's database for training the QRNNs. This could further benefit the training of QRNNs by making the ancillary data that GPROF uses, such as surface temperature and TCWV, available as additional inputs. It's reasonable to believe that with enough training data, more information should help the QRNN to make better predictions. To conclude the discussion about improvements, using more training data in general is probably the most obvious way to improve the performance of the QRNNs, especially over land. As extreme rain events are pretty rare, extending the training database would help the QRNNs to deal with them better as well.

Implementing these improvements would be natural next steps in the work. Perhaps it would be beneficial to add information from even larger neighbourhoods too. Another interesting thing would be to try different network architectures, such as Convolutional Neural Networks (CNNs), to do swath-to-swath predictions building upon the success of CNNs in image classification and pixel segmentation. The QRNNs considered in this study are limited to a certain type of surface in each case. A very interesting next step would be to train QRNNs to predict precipitation over mixed surface areas, such as Sweden where there are lots of lakes and snow covered areas. If the QRNNs can learn predict rain rates well in such areas, the ultimate goal would be to train a single QRNN to make predictions on a global scale.

References

Chollet, F. et al. (2019). Keras. `https://github.com/keras-team/keras`. Accessed: 2019-05-13.

Germann, U., Galli, G., Boscacci, M., and Bolliger, M. (2006). Radar precipitation measurement in a mountainous region. *Quarterly Journal of the Royal Meteorological Society*, 132(618):1669–1692.

GPM Science Team (2018). Algorithm Theoretical Basis Document (ATBD): GPROF 2017 Version 1 and Version 2 (used in GPM V5 processing). `https://pmm.nasa.gov/sites/default/files/document_files/ATBD_GPM_GPROF_June1_2017.pdf`. Accessed: 2019-05-23.

Håkansson, N., Adok, C., Thoss, A., Scheirer, R., and Hörnquist, S. (2018). Neural network cloud top pressure and height for modis. *Atmospheric Measurement Techniques*, 11(5):3177–3196.

Hou, A. Y., Kakar, R. K., Neeck, S., Azarbarzin, A. A., Kummerow, C. D., Kojima, M., Oki, R., Nakamura, K., and Iguchi, T. (2014). The global precipitation measurement mission. *Bulletin of the American Meteorological Society*, 95(5):701–722.

Huffman, G. J., Bolvin, D. T., Nelkin, E. J., Wolff, D. B., Adler, R. F., Gu, G., Hong, Y., Bowman, K. P., and Stocker, E. F. (2007). The trmm multisatellite precipitation analysis (tmpa): Quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. *Journal of hydrometeorology*, 8(1):38–55.

Islam, T., Hu, Y., Kokhanovsky, A. A., and Wang, J. (2017). *Remote Sensing of Aerosols, Clouds, and Precipitation*. Elsevier.

Islam, T., Rico-Ramirez, M. A., Han, D., Srivastava, P. K., and Ishak, A. M. (2012). Performance evaluation of the trmm precipitation estimation using ground-based radars from the gpm validation network. *Journal of Atmospheric and Solar-Terrestrial Physics*, 77:194–208.

Kidd, C. and Huffman, G. (2011). Global precipitation measurement. *Meteorological Applications*, 18(3):334–353.

Kidd, C., Tan, J., Kirstetter, P.-E., and Petersen, W. A. (2018). Validation of the version 05 level 2 precipitation products from the gpm core observatory and constellation satellite sensors. *Quarterly Journal of the Royal Meteorological Society*, 144:313–328.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of economic perspectives*, 15(4):143–156.

Kummerow, C. D., Randel, D. L., Kulie, M., Wang, N.-Y., Ferraro, R., Joseph Munchak, S., and Petkovic, V. (2015). The evolution of the goddard profiling algorithm to a fully parametric scheme. *Journal of Atmospheric and Oceanic Technology*, 32(12):2265–2280.

Li, Y. and Yuan, Y. (2017). Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pages 597–607.

NASA (2019a). An Overview of EOSDIS. `https://earthdata.nasa.gov/about`. Accessed: 2019-03-28.

NASA (2019b). Global precipitation measurement. `https://www.nasa.gov/mission_pages/GPM/main/index.html`. Accessed: 2019-03-27.

Pfreundschuh, S., Eriksson, P., Duncan, D., Rydberg, B., Håkansson, N., and Thoss, A. (2018). A neural network approach to estimating a posteriori distributions of bayesian retrieval problems. *Atmospheric Measurement Techniques*, 11(8):4627–4643.

Rico-Ramirez, M. A. and Cluckie, I. D. (2008). Classification of ground clutter and anomalous propagation using dual-polarization weather radar. *IEEE Transactions on Geoscience and Remote Sensing*, 46(7):1892–1904.

Rodgers, C. D. (2000). *Inverse methods for atmospheric sounding: theory and practice*, volume 2. World scientific.

Sanò, P., Panegrossi, G., Casella, D., Marra, A., D'Adderio, L., Rysman, J., and Dietrich, S. (2018). The passive microwave neural network precipitation retrieval (pnpr) algorithm for the conical scanning global microwave imager (gmi) radiometer. *Remote Sensing*, 10(7):1122.

Tang, H., Tan, K. C., and Yi, Z. (2007). *Neural networks: computational models and applications*, volume 53. Springer Science & Business Media.

Tapiador, F. J., Turk, F. J., Petersen, W., Hou, A. Y., García-Ortega, E., Machado, L. A., Angelis, C. F., Salio, P., Kidd, C., Huffman, G. J., et al. (2012). Global precipitation measurement: Methods, datasets and applications. *Atmospheric Research*, 104:70–97.

The typhon authors (2019). typhon – Tools for atmospheric research. `https://github.com/atmtools/typhon`. Accessed: 2019-05-13.