DeepPrecip: A deep neural network for precipitation retrievals

Fraser King[1], George Duffy[2,3], Lisa Milani[4,5], Christopher G. Fletcher[1], Claire Pettersen[6], and Kerstin Ebell[7]

[1]Dept. of Geography & Environmental Management, University of Waterloo, 200 University Ave W, Waterloo, Ontario, Canada
[2]NASA, Jet Propulsion Laboratory, 4800 Oak Grove Dr, Pasadena, 91109, California, USA
[3]Earth and Environmental Sciences, University of Syracuse, 900 South Crouse Ave, Syracuse, New York, USA
[4]NASA, Goddard Space Flight Center, 8800 Greenbelt Rd, Greenbelt, Maryland, USA
[5]Earth System Science Interdisciplinary Center, University of Maryland, 5825 University Research Ct suite 4001, College Park, Maryland, USA
[6]Climate and Space Sciences and Engineering, University of Michigan, Space Research Building, Climate &, 2455 Hayward St, Ann Arbor, Michigan, USA
[7]Institute for Geophysics and Meteorology, University of Cologne, Albertus-Magnus-Platz, Cologne, Germany

Correspondence: Fraser King (fdmking@uwaterloo.ca)

Abstract. Remotely-sensed precipitation retrievals are critical for advancing our understanding of global energy and hydrologic cycles in remote regions. Radar reflectivity profiles of the lower atmosphere are commonly linked to precipitation through empirical power laws, but these relationships are tightly coupled to particle microphysical assumptions that do not generalize well to different regional climates. Here, we develop a robust, highly generalized precipitation retrieval from a deep convolutional neural network (DeepPrecip) to estimate 20-minute average surface precipitation accumulation using near-surface radar data inputs. DeepPrecip displays high retrieval skill and can accurately model total precipitation accumulation, with a mean square error (MSE) 160% lower, on average, than current methods. DeepPrecip also outperforms a less complex machine learning retrieval algorithm, demonstrating the value of deep learning when applied to precipitation retrievals. Predictor importance analyses suggest that a combination of both near-surface (below 1 km) and higher-altitude (1.5 - 2 km) radar measurements are the primary features contributing to retrieval accuracy. Further, DeepPrecip closely captures total precipitation accumulation magnitudes and variability across nine distinct locations without requiring any explicit descriptions of particle microphysics or geospatial covariates. This research reveals the important role for deep learning in extracting relevant information about precipitation from atmospheric radar retrievals.

1 Introduction

Accurate estimates of surface precipitation are highly sought-after as they inform flood forecasting operations, water resource management practices and energy planning (Buttle et al., 2016; Gergel et al., 2017). Due to the sparse nature of in situ precipitation measurement networks, remote sensing has become a prominent alternative source of observations for deriving surface precipitation estimates (Liu, 2008). Ground-based scanning radars are valuable resources as they provide estimates of precipitation over a wider area and at a higher temporal resolution compared to traditional in situ gauges (Lemonnier et al., 2019).

Additionally, the size and availability of both vertically pointing and space-borne remote sensing datasets have expanded greatly in recent decades as a result of technological instrument improvements and new satellite missions (Quirita et al., 2017).

Remotely-sensed radar observations used in empirical, power-law relationships can relate radar reflectivity (RFL) estimates (Z_e) to surface snowfall (S) or rainfall (R) rates (Eq. 1) (Matrosov et al., 2008; Kulie and Bennartz, 2009; Schoger et al., 2021).

$$Z = a \times (S/R)^b \tag{1}$$

These radar-based retrievals are powerful tools for filling current observational gaps and have been applied to great effect in previous literature (Levizzani et al., 2011; Hiley et al., 2010). However, these relationships demonstrate an inability to generalize well to unseen validation data as a consequence of the microphysical particle assumptions (e.g. shape, diameter, particle size distribution (PSD), terminal fall velocity and mass) used in each relationship's unique derivation (Jameson and Kostinski, 2002).

Recent machine learning (ML) approaches have demonstrated improvements in estimating surface precipitation from remotely-sensed data compared to traditional nowcasting methods (Shi et al., 2017; Kim and Bae, 2017). Deep learning models have benefited greatly from the increased observational sample provided by remote sensing missions and have shown skill in learning complex spatiotemporal characteristics of the underlying datasets (Chen et al., 2020b). However, a deep learning convolutional surface precipitation retrieval using vertical column radar data with no spatiotemporal covariates has yet to be developed to our knowledge. Previous ML studies have typically focused on passive microwave and infrared datasets which lack a detailed analysis of the vertical column structure, or suffer from a limited sample for model training across multiple, distinct regional climates (Xiao et al., 1998; Adhikari et al., 2020; Ehsani et al., 2021).

In this work, we evaluate the abilities of a novel deep learning precipitation retrieval algorithm trained on vertically pointing radar (up to 3 km above the surface). The regression model we present (DeepPrecip) is a hybrid deep learning neural network consisting of a feature extraction convolutional neural network (CNN) front-end and a regression feedforward multilayer perceptron (MLP) back-end. The combination of these two architectures allows DeepPrecip to recognize and learn the nonlinear relationships between different layers in the vertical column of radar observations and produce an accurate surface precipitation estimate. Through an analysis of feature input combinations, DeepPrecip performance is examined to identify regions within the vertical column that contain the most important contributions to retrieval accuracy (Lundberg and Lee, 2017). The relationships that exist between different layers of the vertical profile (and each atmospheric covariate) can be used to help inform current and future active radar retrievals of surface precipitation.

2 Data

2.1 Study Sites

In situ data was collected from 9 study sites (Fig. 1.a) from 2012-2020 (Table 1). Colored markers in Fig. 1.b indicate periods where non-zero surface precipitation was recorded. Study sites were selected based on the required presence of a micro rain

radar (MRR) and collocated Pluvio2 weighted precipitation gauge. Rain, snow and mixed-phase precipitation were recorded, with each site's precipitation phase and intensity distribution of observations differing based on the regional climate. For instance, Marquette experienced strong lake-effect snowfall while Cold Lake received mostly light, shallow snowfall. Further, due to the warmer temperatures recorded at OLYMPEx, these sites were classified as primarily experiencing liquid precipitation, while ICE-POP received only solid precipitation.

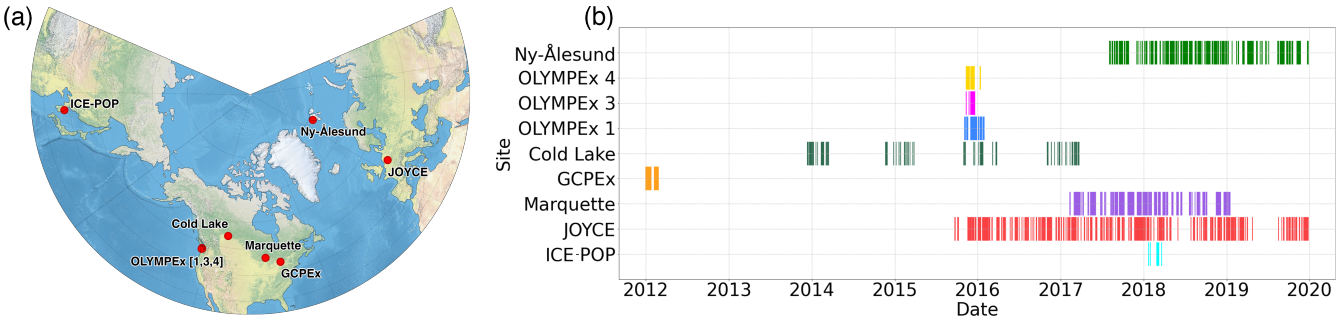

Figure 1. Observational input data locations and temporal coverage periods. (a), Geographic study site locations. **(b),** timeline of observational coverage (periods of active precipitation) for each site from 2012 to 2020.

2.2 Pluvio2 precipitation weighing gauge

Reference surface precipitation observations were collected by OTT Pluvio2 weighted gauges at each site. The Pluvio2 gauge records the precipitation accumulation from falling hydrometeors with a minimum time resolution of 1 minute (Colli et al., 2014). It includes a 200 cm^2 heated surface orifice (400 cm^2 at Ny-Ålesund) to prevent snow and ice buildup, along with site-specific wind shielding implemented as described in Table 1. These fence setups include a Double Fence Intercomparison Reference (DFIR) shield which is a large, double fenced wooden structure which helps significantly reduce the impact of wind on surface precipitation measurements (Rasmussen et al., 2012; Kochendorfer et al., 2022). The Alter shield system consists of multiple freely hanging, spaced metal slats around the gauge top opening which also helps mitigate undercatch issues during strong winds (Colli et al., 2014). Sensitivity analyses of different rolling temporal windows indicated an optimal temporal resolution of 20-minute non-real time accumulation (measurement results 5 minutes after precipitation accumulation), with minimum observational thresholds of at least 0.2 mm over the course of an hour from the Pluvio2 gauge.

2.3 Micro rain radar

Vertical pointing MRRs (developed by METEK) were located nearby the Pluvio2 gauges at each site to record complementary atmospheric observations. The MRR is a K-band (24 GHz) continuous wave Doppler radar which provides information related to hydrometeor particle activity up to 3.1 km above the surface (or 1 km for Ny-Ålesund) as a function of spectral power backscatter intensity. The MRR provides 29 vertical bins (of size 100 m) spanning 300 m to 3100 m above the surface as shown for each site in Fig. 2.a. Raw radar measurements were preprocessed using Maahn's improved MRR processing tool

Table 1. Summary of in situ study site locations, identifiers, and observational details.

Site	ID	Lat	Lon	Elev.	Sample (N)	Shielding	Source
Ny-Ålesund	0	78.92	11.92	11	19068	Alter	(Schoger et al., 2021)
ICE-POP	1	37.67	128.7	789	1705	DFIR	(Kim et al., 2021; Munchak et al., 2022)
GCPEx	2	44.23	-79.78	252	2314	DFIR	(Skofronick-Jackson et al., 2015)
Marquette	3	46.53	-87.55	430	8369	Alter	(Pettersen et al., 2020; Kulie et al., 2021)
OLYMPEx 4	4	47.39	-123.87	2155	6444	None	(Houze et al., 2017)
OLYMPEx 1	5	47.5	-123.58	3340	9114	None	(Houze et al., 2017)
OLYMPEx 3	6	47.68	-123.38	2100	5727	None	(Houze et al., 2017)
JOYCE	7	50.9	6.4	95	43579	Alter	(Lahnert et al., 2015)
Cold Lake	8	54.4	-110.26	541	1692	Alter	(Boudala et al., 2021)

(IMProToo) for noise removal, dealiasing and for extending the minimum detectable dBZ to -14 which allows for improved measurements of solid precipitation. This data was then temporally averaged to align to the same 20-minute windows generated for the Pluvio2 observations and used as a model input (Maahn and Kollias, 2012).

2.4 ERA5

European Centre for Medium-Range Weather Forecasts Reanalysis version 5 (ERA5) hourly temperature (TMP) and vertical wind velocity (WVL) on pressure levels from 0 to 3 km were also included as additional input covariates to DeepPrecip (Hersbach et al., 2020). These inputs allow the model to more accurately recognize different precipitation event structures, large-scale atmospheric dynamics and hydrometeor phases during training. Note that WVL units (Pa/s) are defined using the ECMWF Integrated Forecasting System (IFS) which adopts a pressure based vertical co-ordinate system (i.e. negative values indicate upwards air motion, since pressure decreases with height). Each of these variables were linearly interpolated to align with the MRR data over 20 minute intervals and at 100 m vertical resolution.

2.5 Surface meteorology

Collocated surface temperature (degrees Celsius ($^\circ$ C)) and 10-meter wind speed (m/s) meteorologic observations were also collected from instruments installed at each site and temporally aligned to the Pluvio2 and MRR datasets. Surface wind data acts as an additional observational constraint for mitigating the effects of undercatch on unshielded measurement gauges (Rasmussen et al., 2012). Undercatch occurs when precipitation falling in the presence of wind can cause hydrometeors to pass over the gauge top orifice. This effect has been shown to bias reported precipitation quantities by up to 10% (Ehsani and Behrangi, 2022). We therefore limit the available training dataset to periods when surface wind speeds are < 5 m/s, as this restricts the analysis to low-medium wind speed events at each location to maintain a high gauge-catch efficiency (Yang, 2014).

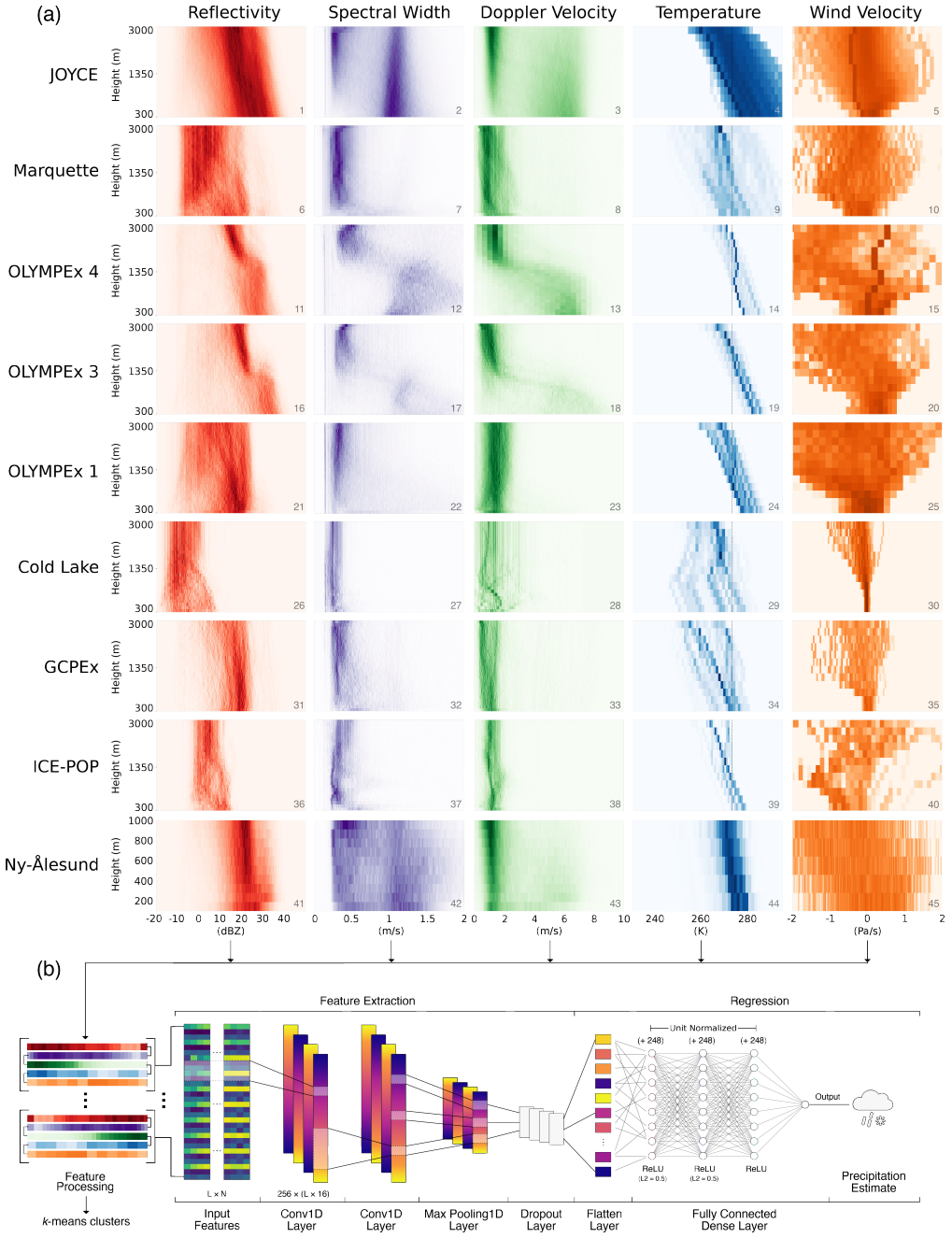

Figure 2. DeepPrecip input covariates, feature processing pipeline and model architecture. (a), Site-predictor matrix of normalized Micro-Rain Radar (MRR) and ERA5 observational frequency histograms used in model training and testing. Note that darker colors in the 2D heatmaps indicate a higher frequency of observations. **(b),** DeepPrecip convolutional neural network diagram for L inputs with N predictors.

This preprocessing step reduces the average size of our total observational pool by 16% across all stations, however we note that maximum intensity precipitation events are not removed using this technique.

Surface meteorologic station temperature data is used for precipitation-phase partitioning at 5° C to allow for $Z_e - S/R$ comparisons with DeepPrecip. Additional dry surface air temperature thresholds of 0°, 1° and 2° C were also examined, but $Z_e - S/R$ performance for both rain and snow appeared optimal when classified using a 5° C threshold (where temperatures $< 5°$ C are considered as solid precipitation and temperatures $>= 5°$ C are considered as rainfall). This simple temperature threshold is an additional source of uncertainty in our comparisons with the $Z_e - S/R$ relationships due to the influence of mixed-phase precipitation on power law accuracy, along with uncertainties in the location of the active melting layer (Jennings et al., 2018). A more sophisticated phase partitioning system (e.g. using wet-bulb temperature as described in Sims and Liu (2015)) could also be linked to DeepPrecip as an additional predictor to further improve classification of mixed-phase precipitation in future work.

3 Methods

3.1 Radar-precipitation power laws

Relating radar reflectivity observations to surface accumulation has been done extensively in past surface and spaceborne radar missions through $Z_e - S/R$ power law relationships (Skofronick-Jackson et al., 2017; Liu, 2008). These power law relationships are empirically defined by relating reflectivity values in a near surface bin to observed surface accumulation under a set of assumed particle microphysics (e.g. size, shape, density and fallspeed) (Matrosov et al., 2008). While these techniques have been used to great success in previous studies Schoger et al. (2021); Levizzani et al. (2011), the assumptions about snowfall and rainfall particle microphysics makes the generalization of these power laws less robust, which contributes to high uncertainty when applied across large areas with unique regional climates (Jameson and Kostinski, 2002).

We examine an ensemble of 12 Ka- and K-band $Z_e - S/R$ relationships in this work to compare with model output from DeepPrecip (Table 2). As a consequence of the short temporal period (20 minutes) used in this analysis, MSE values are typically small (< 0.1 mm^2). Each $Z_e - S/R$ relationship was applied to a near-surface bin in the reflectivity profile (bin 5 for DP_{full} and DP_{near}, and bin 11 for DP_{far}) to derive a corresponding surface precipitation estimate. These bins were selected based on a sensitivity analysis where we examined the performance of multiple near-surface high-importance regions of the vertical column (not shown). The best performing regions were identified as the above bins (5 and 11) based on the respective region of the vertical column being considered (near or far). More information regarding the derivation of each $Z_e - S/R$ relationship can be found in Table 2.

To further evaluate the performance of DeepPrecip, we also include model comparisons to a set of six site-derived $Z_e - P$ (reflectivity precipitation) power law relations. Each $Z_e - P$ relationship is empirically derived from the collocated MRR and Pluvio data at each each observational site examined in this work (excluding Cold Lake and Ny-Ålesund due to the limited available sample and vertical extent of each site, respectively). Each $Z_e - P$ relation is fit via a non-linear least-squares approach for finding optimal a and b coefficients in Eq. 1 using SciPy's $curve\_fit$ optimization algorithm (Virtanen et al.,

Table 2. Details for each multi-phase precipitation power law relationship.

Phase	Name	Source	Power Law	Reference
Solid	AVE_K	K	$Z_e = 77.61 \times S^{1.22}$	(Schoger et al., 2021)
	KB09sp	Ka	$Z_e = 19.66 \times S^{1.47}$	(Kulie and Bennartz, 2009)
	KB09ag	Ka	$Z_e = 313.29 \times S^{1.85}$	(Kulie and Bennartz, 2009)
	KB09br	Ka	$Z_e = 24.04 \times S^{1.51}$	(Kulie and Bennartz, 2009)
	M07	Ka	$Z_e = 56.00 \times S^{1.20}$	(Matrosov, 2007)
	S17	K	$Z_e = 18.00 \times S^{1.10}$	(Souverijns et al., 2017)
Liquid	BP09h	K	$Z_e = 32.00 \times R^{3.30}$	(Van Baelen et al., 2009)
	BP09m	K	$Z_e = 324.00 \times R^{2.40}$	(Van Baelen et al., 2009)
	MP48	–	$Z_e = 200.00 \times R^{1.60}$	(Marshall and Palmer, 1948)
	J19bb	K	$Z_e = 367.00 \times R^{1.37}$	(Jash et al., 2019)
	J19nbb	K	$Z_e = 211.00 \times R^{1.44}$	(Jash et al., 2019)
	J19hr	K	$Z_e = 168.00 \times R^{1.40}$	(Jash et al., 2019)

2020). Each $Z_e - P$ relationship was then applied to bin 5 reflectivities at each site (i.e. the same process as is used for $Z_e - S/R$ relationships) and compared with in situ observations to assess their general accuracy.

3.2 Neural network architecture

DeepPrecip is a feedforward convolutional neural network that takes as input a vector of 115 atmospheric covariates (Table 3), performs a feature extraction of the vertical column and outputs a single surface precipitation estimate using a fully connected multilayer perceptron. While the structure of this final version of DeepPrecip is complex, the retrieval evolved from a much simpler initial state based on a multiple linear regression (MLR) model. Due to clear nonlinearities between observed reflectivity data and surface precipitation accumulation, the MLR model was unable to capture in situ variability and provided estimates near the mean accumulation value. Similar radar-based precipitation retrieval studies by Chen et al. (2020a) and Choubin et al. (2016) have demonstrated much better performance using an ML-based approach which led to the development of an RF model, an MLP and finally the CNN.

The 1D convolutional layers perform a feature extraction of the vertical column of inputs to reduce the total number of parameters being fed into DeepPrecip's fully connected dense layers. This 1D-CNN structure can identify relationships within the vertical column, save on memory and lower computational training time requirements. To perform a 1D feature extraction, the forward propagation step between the previous convolutional layer $(l-1)$ to the input neurons of the current layer (l) are expressed in Eq. 2 (Abdeljaber et al., 2017).

Table 3. Summary of DeepPrecip full vertical column model input covariates.

Predictor	Abbreviation	Count	Units	Source	Type
Reflectivity	RFL	29	dBZ	MRR	float64
Doppler velocity	DOV	29	m/s	MRR	float64
Spectral width	SPW	29	m/s	MRR	float64
Temperature	TMP	12	K	ERA5	float64
Wind velocity	WVL	12	P_a/s	ERA5	float64
Profile group	PG	4	Indicator	K-mean	Boolean

$$x_k^l = f(b_k^l + \sum_{i=1}^{N_{l-1}} Conv1d\left(w_{ik}^{l-1}, s_i^{l-1}\right)) \tag{2}$$

Where k and l refer to the k^{th} neuron for layer l with x as the resulting input and b as the scalar bias. s and w terms represent the neuron output and kernel weight matrix respectively, from the i^{th} neuron of layer $l-1$ (and to the k^{th} neuron of layer l for w). The function '$f()$' represents the activation function used to transform the weighted sum into an output to be used in the following network layer.

The RF model tested in this study was based on previous work from King et al. (2022) where a RF was used to retrieve surface snow accumulation from a collocated X-band and Pluvio2 instrument at a single experiment site (GCPEx). The RF developed in said study demonstrated good skill in estimating surface accumulation, and so we incorporate the same model here (retrained on the MRR and ERA5 data from this study) as a baseline comparison to other ML retrieval methods (i.e. DeepPrecip).

The final DeepPrecip model structure is outlined in Fig. 2.b. It includes two 1d-convolutional layers, a 1d max pooling layer, dropout layer, flattening layer and concludes in a dense MLP regressor with 3 hidden layers. The total number of trainable model parameters in DeepPrecip is 3,937,793. Model training and testing was performed using a 90/10 (non-shuffled) split on each site to generate training and testing datasets for each location. As an additional preprocessing step, we standardize all input covariates to remove the mean and by scaling inputs to unit variance. The non-shuffled nature of this splitting process allows for DeepPrecip estimates to be validated against unseen data and prevents overfitting from training on temporally autocorrelated vertical column inputs. Additionally, this stratified selection process guarantees that an equal percentage of data is included from each site during training.

Retrieval accuracy is primarily assessed using a mean squared error (MSE) skill metric calculated between each model's estimated surface accumulation values and the total Pluvio2 non-real-time reference accumulation observations over 20 minutes. Performance statistics are reported from the average skill of the test portion of a non-shuffled 90/10 train/test CV split (i.e. DeepPrecip trained and tested 10 times on different contiguous portions of the full available sample). Note that each

split is stratified to include 10% of each station's sample in every test split. Uncertainty estimates are calculated from running each CV split 50 times using dropout to gain additional insight into model variability (resulting in 500 total model instances). The dropout layers simulate training a large number of models with differing architectures in a highly parallelized manner by randomly deactivating (or dropping) a certain fraction of nodes within the network to provide a distribution of retrieval estimates.

3.3 Hyperparameter optimization

DeepPrecip was developed, trained and optimized on Graphcore intelligence processing units (IPUs) MK2 Classic IPU-POD4 (Louw and McIntosh-Smith, 2021), which significantly sped up the training time by a factor of 6.5 compared to a state-of-the-art NVIDIA Tesla V100 GPU. Additional training throughput comparisons are included in Table 4. Training was completed using a combination of open-source Python packages including Keras, Tensorflow and scikit-learn. An extension of stochastic gradient descent known as Adam optimization (adaptive moment estimation) is used to continually update internal network weights in the model during training to minimize a standard MSE loss function (Eq. 3) and track model learning over time.

$$L = \sum_{i=1}^{D} (x_i - y_i)^2 \tag{3}$$

Table 4. DeepPrecip model training throughput comparisons running on Tensorflow (v2.4.3) using a batch size of 128 samples on different hardware. Note that 2 IPUs were used in comparison to 1 GPU/TPU to equalize average computation costs when training DeepPrecip using each piece of hardware.

Hardware	Processors	Samples/second
Graphcore Intelligence Processing Unit (IPU)	2	500
NVIDIA Tesla V100 Tensor Core GPU	1	77
Google Tensor Processing Unit (TPU)	1	56
NVIDIA Tesla K80 GPU	1	23

Hyperparameters do not change value during training (in contrast to model parameters like internal node weights), but they play a critical role in the neural network learning process to map input features to an output. Selecting optimal hyperparameter values is an important part in constructing a model which minimizes loss, improves model efficiency and quality, and mitigates overfitting. Multiple steps were taken to address concerns of model overfitting. In addition to the use of non-shuffled training, we employ multiple regularization methods including early stopping, dropout, the application of layer weight constraints and L2 regularization (details in Table 5). L2 regularization (or ridge regression) adds an additional penalty term to the MSE loss function which helps to create less complex models when dealing with many input features to improve model generalization.

Table 5. DeepPrecip hyperparameters optimization details.

Hyperparameter	Value	Parameter Space
Activation	ReLU	['relu', 'tanh', 'sigmoid']
Batch Size	128	[64, 128, 256, 512]
Dropout Rate	0.1	[0.001, 0.01, 0.1, 0.25, 0.5, 0.75]
Early Stop Patience	8	[4, 8, 16, 32]
Epochs	512	[64, 128, 256, 512, 1024]
Filters	256	[4, 16, 64, 128, 256]
Hidden Layers	3	[1, ..., 20]
Kernel Size	16	[2, 4, 8, 16, 32]
L2 Regularization	0.5	[0.001, 0.01, 0.1, 0.5]
Learning Rate	1e-7	[0.001, 0.0001, 1e-5, 1e-7]
Loss Function	MSE	['MSE']
Neurons	256	[64, 128, 256, 512, 1024]
Optimizer	Adam	['Adam']
Pool Size	2	[2]

To select the optimal values for the aforementioned hyperparameters, and to optimize DeepPrecip's general structure, we use a form of hyperparameterization known as hyperband optimization (Li et al., 2017). Hyperband is a variation of Bayesian optimization which intelligently samples the parameter space to find hyperparameter values that minimize loss while learning from previous selections. Hyperband adds an additional component to the analysis by also slowly increasing the number of epochs run during each phase of the optimization process to sample in a more efficient manner. DeepPrecip hyperparameters were derived by running a 10-fold CV hyperband optimization continuously on a single Graphcore IPU for approximately two weeks. The final hyperparameter values (and their respective parameter search spaces) can be found in Table 5.

3.4 Unsupervised classification layer

An unsupervised k-means clustering preprocessing step is also applied using MRR reflectivity profiles as input to provide DeepPrecip with insights into distinct profile group (PG) vertical column structures (Fig. 2.b). Minimizing within-cluster sum of squares between each vertical column radar estimate results in $k = 4$ PGs being selected using the within-cluster-sum of squared errors elbow criterion method (Fig. 3). The elbow method is a clustering heuristic which allows for an optimal number of clusters to be selected as a function of diminishing returns of explained variation (i.e. finding the elbow or "knee of the curve"). K-means clustering was applied using Python's scikit-learn package on all input reflectivity data to generate four profile clusters which were included as additional input parameters to DeepPrecip. These clusters are useful for partitioning the precipitation data into groups based on different precipitation intensity-classes (trace, low, medium and high intensity) to

identify where DeepPrecip finds the most important contributors to high retrieval accuracy for each category of storm intensity. Derived cluster groups are useful for interpreting feature importances from model output (Section 4.2).

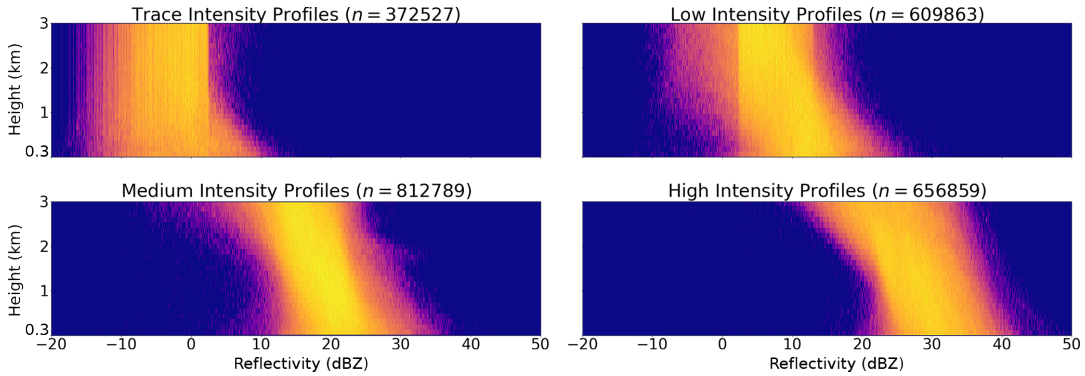

Figure 3. K-means cluster reflectivity intensity-classes of vertical profiles from the MRR instruments at all sites. A total of 2452038 vertical profiles are organized by reflectivity intensity (dBZ) into $k = 4$ precipitation intensity subsets. The four groups were selected using the within-cluster sum of square elbow method.

4 Results

4.1 DeepPrecip retrieval performance

We first examine the differences in performance between DeepPrecip and a random forest (RF) that has demonstrated good performance in our previous work (not shown) to assess the capabilities of a less-sophisticated ML-based approach over a CNN. DeepPrecip demonstrates improved skill in capturing most of the peaks and troughs in observed precipitation variability (Fig. 4.a). These differences are most clearly demonstrated in Fig. 4.a at OLYMPEx and JOYCE, where DP more accurately predicts Pluvio2 precipitation extremes compared to the RF. Both models appear to struggle in capturing accumulation intensities during periods of mixed-phase precipitation when temperatures are near zero degrees C (i.e. Marquette, JOYCE and the tail end of OLYMPEx 1) due to a lack of training data with similar climate conditions and the complex nature of such events. DP does demonstrate improved skill at capturing light intensity precipitation at the beginning of the JOYCE period (compared to the RF), however this is with some uncertainty as noted by the wider shaded region (1 standard deviation). Performance statistics (Fig. 4.b) summarize these improvements with DeepPrecip showing MSE values 21% lower and r^2 values 34% higher (significant at $\alpha < 0.05$) compared to the RF.

Total cumulative surface accumulation comparisons between DeepPrecip and each $Z_e - S/R$ relationship are then examined in Fig. 4.c for both rain and snow. To examine model skill across different precipitation phases, a simple temperature threshold is imposed where retrievals recorded during periods with temperatures below five degrees C are classified as snow and periods equal to or warmer than five degrees C as rain. DeepPrecip more accurately captures surface precipitation quantities when compared to the $Z_e - S/R$ estimates, with a total accumulation curve similar in shape to that of in situ indicating that

DeepPrecip more closely captures the observed precipitation variability and magnitude. Log-scale MSE statistics are calculated between each model and in situ records in Fig. 4.d and indicate that DeepPrecip consistently outperforms traditional $Z_e - S/R$ power-law methods by 200% on average. As a general precipitation retrieval algorithm, we do not explicitly train a DeepPrecip$_{snow}$ and DeepPrecip$_{rain}$ model for different precipitation phases with unique regional atmospheric microphysical conditions. While the $Z_e - S/R$ models shown in Fig. 4.c/d are bespoke for rain or snow, DeepPrecip is trained on all data with no a priori knowledge of the underlying physical precipitating particle state.

DeepPrecip estimates of accumulated rain display a lower MSE than that of snow (Fig. 4.d). We believe these differences to be twofold: 1) the larger sample of rainfall events in the training data (3 times that of snowfall); and 2) the more complex nature of snow particle microphysics. Unlike the uniform properties of a rain droplet, the shape, size and fallspeed of solid precipitation is much more dynamic and challenging to model (Wood et al., 2013). Continued issues with interference from wind may have also impacted the accuracy of in situ measurements of snow accumulation leading to higher uncertainty and error (further discussions on these uncertainties in Sect. 5) (Kochendorfer et al., 2017). To visualize the range in uncertainty from the CNN model estimates, we display confidence intervals showing 1 standard deviation in Fig. 4.b/d from 50 DeepPrecip model realizations using dropout. Both ML-based models exhibit the highest uncertainty during periods of mixed-phase precipitation at GCPEx and Marquette along with high intensity precipitation at OLYMPEx.

To further evaluate DeepPrecip's retrieval skill over traditional methods, we compare model performance to a set of six custom $Z_e - P$ site-derived power laws (derivation details in Sect. 3). While $Z_e - P$ relationships typically perform well in the regional climate under which they were derived, they do not generalize well outside of said climate. This lack of robustness is visible in the differences between in situ and $Z_e - P$ estimates of accumulation in Fig. 5.a, where each $Z_e - P$ (light gray line) displays consistent positive or negative biases and no single power law captures the high variability in accumulation across multiple sites. For instance, OLYMPEx 1 and OLYMPEx 3-derived relationships produce a strong positive bias at JOYCE, and the JOYCE-derived $Z_e - P$ power law is quite negatively biased when applied at OLYMPEx. The mean of all six custom power laws is shown in bold gray, and while it closely captures total mean accumulation across all sites, it is unable to model the high variability in precipitation intensity.

The resulting MSE from the application of each custom $Z_e - P$ relationship to each site (along with DeepPrecip) further demonstrates DeepPrecip's improved robustness (Fig. 5.b). In all other cases, DeepPrecip either outperforms all $Z_e - P$ power laws or is only slightly worse than the power law derived for the site in which it is being tested. On average, DeepPrecip retrievals result in 160% lower MSE values than all $Z_e - P$ site-derived power laws estimates when applied to the testing data across the full spatiotemporal domain (Table 6). Figure 5.b also displays a model intercomparison of each $Z_e - P$ relation, where we can clearly see how $Z_e - P$ relations like those derived at OLYMPEx 1 and 3 are clearly unable to capture the vastly different snowfall regimes at sites like ICE-POP, GCPEx and JOYCE with their much larger MSE values for these sites.

The robustness of DeepPrecip was further evaluated using a leave-one-out cross validation (CV) for each site of training observations. This approach tests the skill of DeepPrecip at predicting precipitation for a location that was not included in the training data, which is a strong indicator of the generalizability of the model. Log-scale MSE results of this test for each site are shown in Fig. 6 for each precipitation-phase subset, along with the corresponding average $Z_e - P/S/R$ estimate when

Table 6. MSE values (in e^{-3} mm^2) for all vertical extent experiments across all models for both solid and liquid precipitation.

Phase	Model	Mean Squared Error (e^{-3} mm^2)		
		Full Column	< 1 km	$1-3$ km
All	DeepPrecip	0.7	0.94	1.2
	RF	1.1	0.92	1.5
	$\overline{Z_e - P}$	20.3	20.3	21.4
Solid	DeepPrecip	1.2	1.5	2.2
	RF	2.9	1.5	4.2
	$\overline{Z_e - S}$	31	31	85
Liquid	DeepPrecip	0.43	0.47	0.85
	RF	0.5	0.53	0.6
	$\overline{Z_e - R}$	16.9	16.9	19.7

applied at that site. These findings demonstrate similar performance to the baseline DeepPrecip model skill, which continues to outperform all traditional power law techniques on average. The large range in skill in the power law relationships at most sites (wide error bars) further demonstrates the relative lack of generalizabiltiy of $Z_e - P/S/R$ relationships to different regional climates. Further, the site-derived power law fits (gray dots) perform worse on average than DeepPrecip for locations that are close in proximity (i.e. the OLYMPEx sites).

Predictably, DeepPrecip performance degrades compared to the baseline model when the testing site is left out since the model is no longer trained using data representing the regional climate of the site being tested. This difference in performance is most notable at the set of OLYMPEx sites, and while DeepPrecip performance is still improved over the $Z_e - S/R$ relationships, we note a substantial percentage increase in MSE (375% on average) at these locations. OLYMPEx measurements were the only observational datasets without any gauge shielding and which is a likely source of uncertainty further contributing to this increase in error when the site is removed from the training set (Kochendorfer et al., 2022).

4.2 Quantifying sources of retrieval accuracy

Identifying regions within the vertical column that are the most important contributors towards retrieval accuracy is critical for informing future satellite-based radar precipitation retrievals. The ground-based radar instruments used in this work do not suffer from the same ground clutter contamination issues typical of satellite-based radar observations and we are therefore able to quantify the contributions to model skill arising from the included boundary layer reflectivity measurements in DeepPrecip. Separating the training data into three subsets based on vertical extent and generating new models with this data, allows us to examine changes in performance as a function of information availability. These subsets include: DP_{full} (all 29 vertical bins, i.e. the baseline model), DP_{near} (the lowest 1 km; 8 bins), and DP_{far} (1-3 km; 21 bins). DeepPrecip MSE results (Table 6)

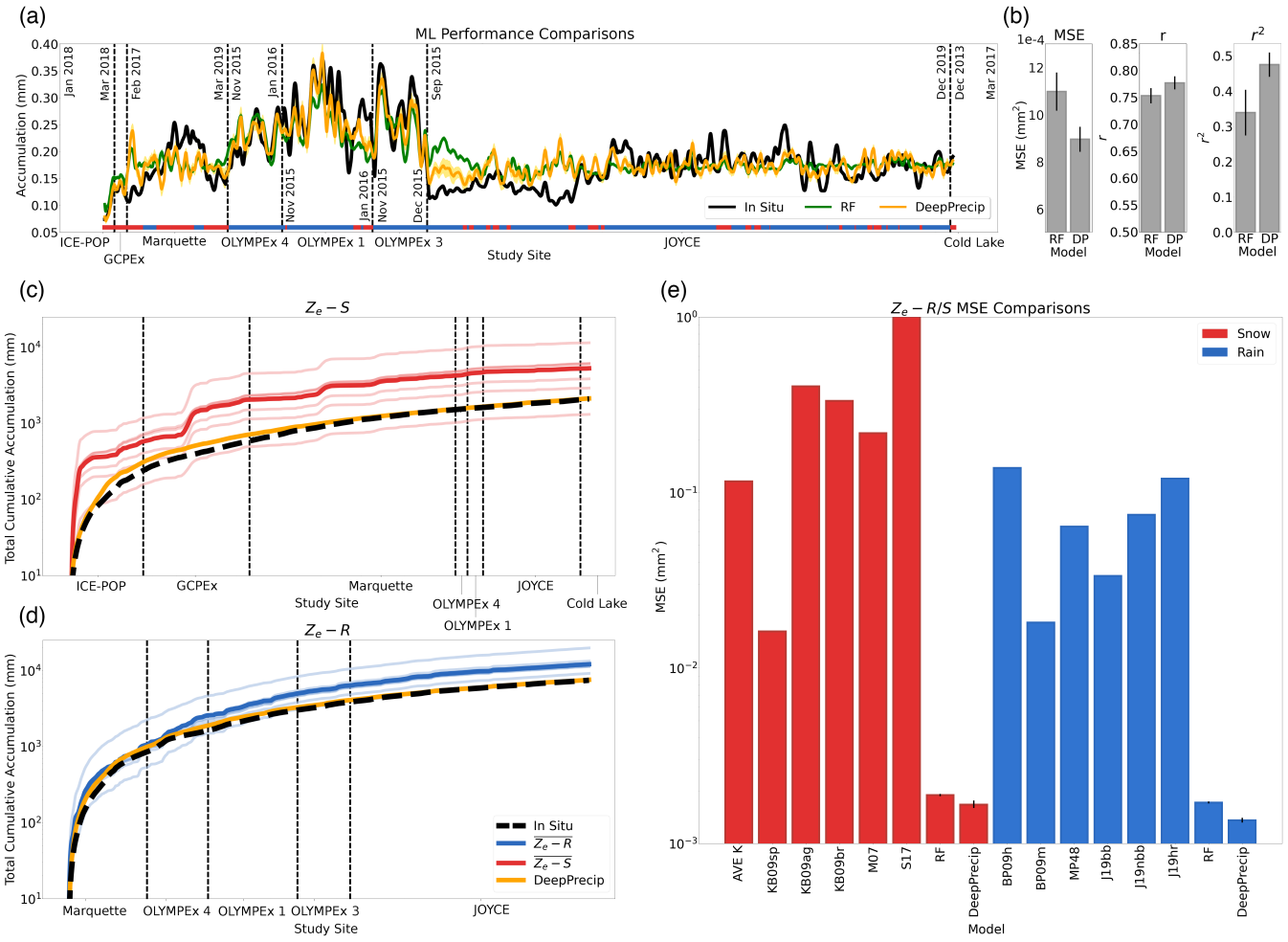

Figure 4. Performance comparisons between DeepPrecip (DP), a random forest and an ensemble of power law-derived retrievals of surface precipitation. (a), Running mean (window size 500 time steps) of accumulation for all sites with Pluvio2 measurements in black, RF estimates in green and DeepPrecip in yellow. Data is sorted by station and then time, with each station separated by a dashed vertical line. 1 standard deviation from 50 dropout runs per cross-validated instance is shown in the shaded regions. **(b),** performance statistics for RF/DeepPrecip accuracy including MSE, Pearson correlation (r) and r^2 with error bars showing 1 standard deviation. **(c),** Timeseries of total accumulation estimates over the full observation period for all $Z_e - S$ relationships and DeepPrecip. The mean of the $Z_e - S$ relationships is shown in bold. **(d),** The same as in (c) but for $Z_e - R$. **(e),** Phase-partitioned log-scale MSE values between each model and in situ observations from 50 model realizations. Note that S17 MSE values extend beyond the top of the graph to 10^1 mm^2.

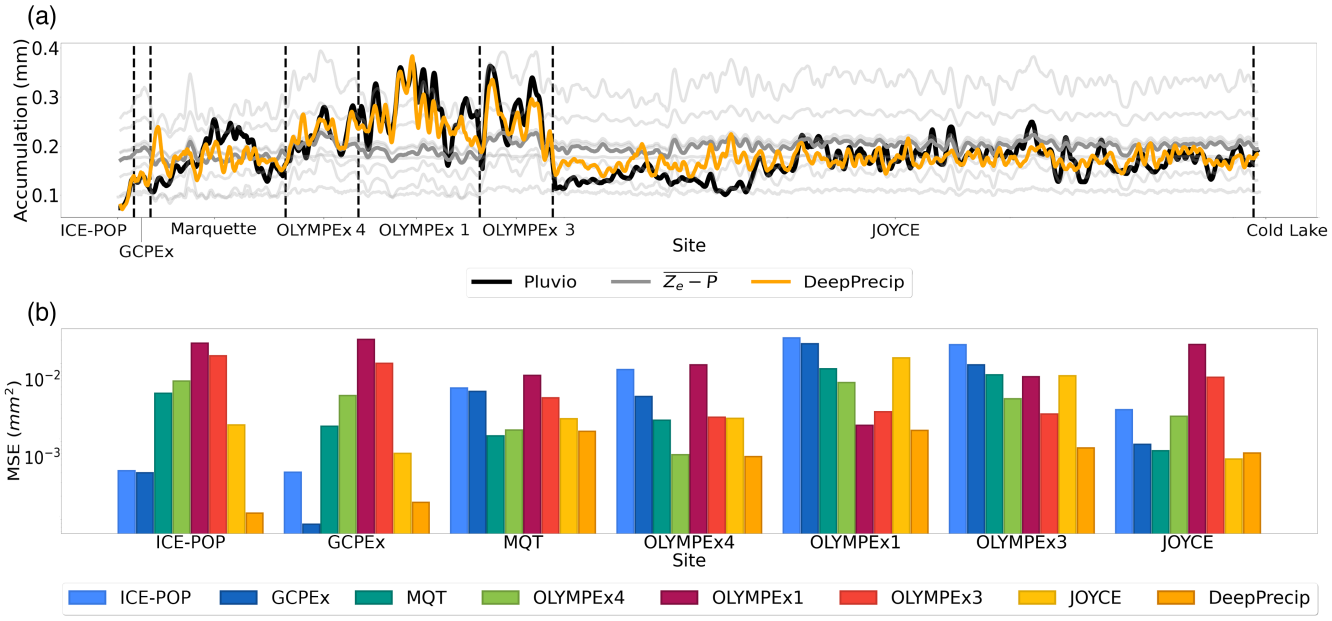

Figure 5. Site-derived empirical $Z_e - P$ power law performance comparisons. (a), The same as Fig. 4.a, except now using $Z_e - P$ relationships derived at each study site. **(b),** MSE values for DeepPrecip and each $Z_e - P$ relationship when tested on each site.

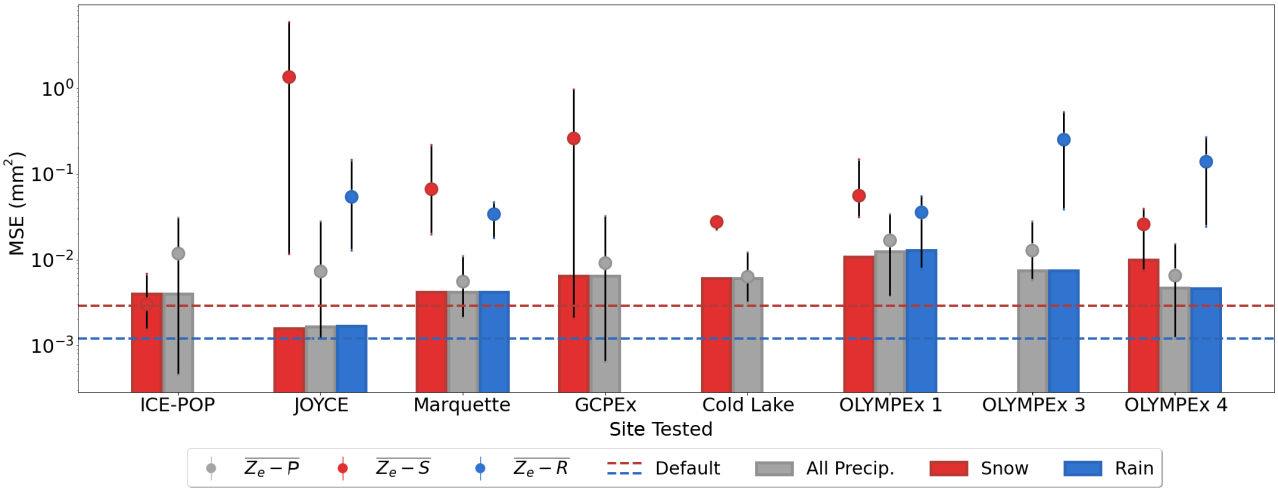

Figure 6. Leave-site-out full column DeepPrecip performance robustness analysis. Each bar represents a DeepPrecip full column log-scale MSE value when trained on all precipitation data excluding the noted site, and then validated against said excluded site (dashed line is the default DeepPrecip model with all sites). Each red and blue dot represents the average $Z_e - S/R$ relationship estimate tested in the same manner (error bars represent the min and max ensemble values). Gray dots represent the mean, min and max ensemble values from all site-derived $Z_e - P$ relationships (excluding the relationship derived from site being tested), when applied to each site.

for each subset suggest that the information provided by a combination of both near-surface and far-profile data results in the highest accuracy.

Since Ny-Ålesund MRR observations were recorded with a maximal vertical extent of 1 km, they are only included in DP_{near}. Model skill when including/excluding Ny-Ålesund training data ($19,000$ samples) was examined to determine whether it was confounding comparisons between the aforementioned vertical profile subset models. The results of these tests suggested that the impact on overall performance is negligible across both precipitation phases when Ny-Ålesund is included or excluded in the training set.

Distributions of surface precipitation anomalies appear distinct for rain and snow (Fig. 7), with the full column model more closely capturing accumulation recorded by in situ gauges. Anomaly frequencies are derived by removing the mean accumulation estimate for each phase at each site. We attribute the structural differences between the anomaly distributions of of snow and rain to the more complex particle size distributions (PSDs) of snowfall coupled with the more variable particle water content of snow compared to that of rain (Yu et al., 2020). Additional uncertainties in the surface Pluvio2 measurement gauge observational records of snowfall due to gauge undercatch is another likely contributor of increased error (Kochendorfer et al., 2022). In Fig. 7.a, both DP_{far} and DP_{near} exhibit higher anomaly values with a flattened curve top and heavy tails. Using a combination of information from both near and far bins reduce these biases and tightens each accumulation anomaly distribution around zero. A similar trend is also present for rain in 7.b, where we again most closely capture the in situ anomaly distribution using DP_{full}.

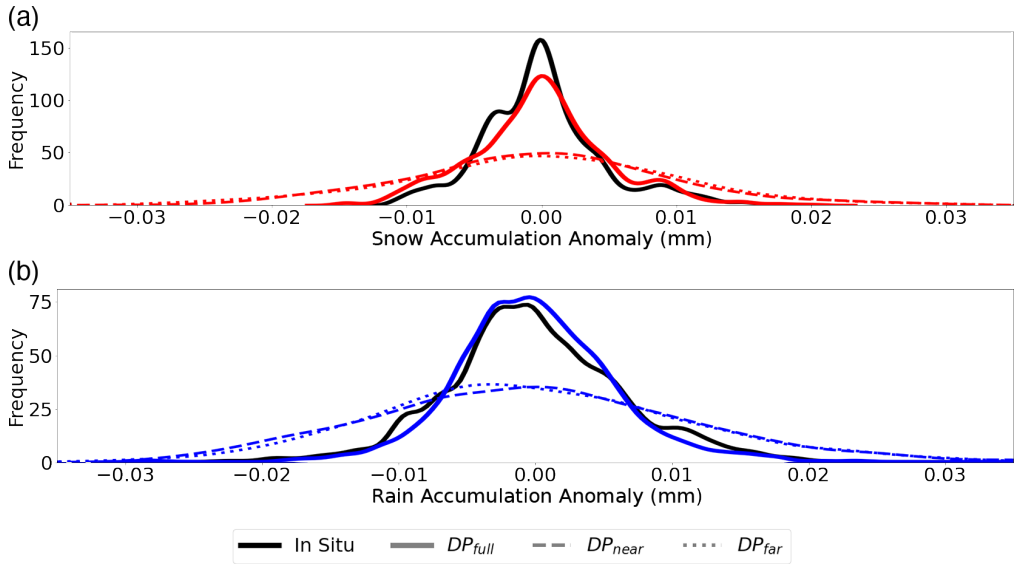

Figure 7. Phase-partitioned surface precipitation accumulation anomaly frequency distributions. DeepPrecip is trained and tested on three subsets of bins from the vertical column: DP_{near} (< 1 km), DP_{far} ($1-3$ km) and DP_{full} (the entire vertical column) for **(a),** solid and **(b),** liquid precipitation.

A major challenge in deep learning is interpreting model output. SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017), is a game theory approach to artificial intelligence model interpretability based on Shapley values that has previously been used to great effect in the Geosciences (Maxwell and Shobe, 2022; Li et al., 2022). Shapley values quantify the contributions from all permutations of input features on retrieval accuracy to identify which are the most meaningful. While computationally expensive (with exponential time complexity), this process provides local interpretability within the model by examining how each possible combination of all input features impacts model accuracy (Jia et al., 2020). Here, the calculated Shapley values give insight into the regions of the vertical column that are contributing the most useful radar information in the precipitation retrieval.

Shapley values for the entire dataset used in DP_{full} indicate that the most important model predictors comprise a combination of both near-surface and far profile bins (Fig. 8). Reanalysis variable model inputs are generally the least influential, except for the trace precipitation case where low-mid level TMP and WVL bins appear highly important (Fig. 8). In all cases, TMP and WVL decrease in importance as a function of height above the surface. DeepPrecip typically considers MRR-derived bins in the 1.5-2.5 km range as the most important predictors. In non-trace intensity profiles, it is the 2 km region Doppler velocity (DOV) observations which are the dominant contributing predictor. When we consider all profiles, reflectivity (the input to $Z_e - S/R$ relationships) is not necessarily the dominant feature, and it is a combination of 1.5-2 km profile information from reflectivity, Doppler velocity and spectral width (SPW) that results in the highest model skill. Combinations of these regions within the vertical column appear to allow DeepPrecip to better understand precipitation events with complex cloud structures which would not necessarily be recognized by conventional $Z_e - S/R$ relations that primarily rely on information from a small subset of near-surface bins.

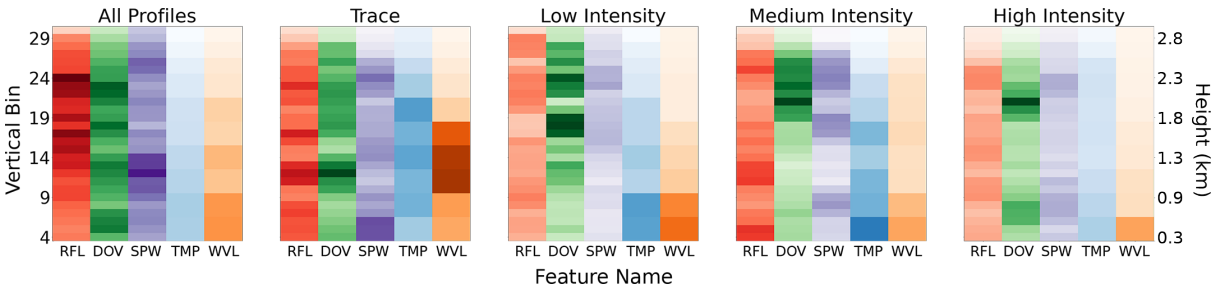

Figure 8. Normalized vertical column Shapley global feature importance values (i.e. $\overline{|SHAP_{DP}|}$). Shapley output values are calculated for different subsets of vertical column reflectivities separated into all profiles, trace intensity, low intensity, medium intensity, and high intensity precipitation events based on a k-means clustering of input data (more in Sect. 3.2). Areas of dark color indicate a high feature importance at that location within the vertical column.

5 Discussion and Conclusions

DeepPrecip not only demonstrates considerable retrieval accuracy without the need for physical assumptions about hydrometeors or spatio-temporal information, but also provides insight into the regions of the vertical column which are most important for improving predictive accuracy. The results from Sect. 4.2 suggest that while the exact altitudes providing predictive information from the vertical column may shift up or down under different precipitation intensities, there exists a consistent combination of both near-surface and far profile bins that always appear as highly important contributors to model skill. Furthermore, while RFL is typically considered as the most important predictor in radar-based precipitation retrievals (Stephens et al., 2008; Skofronick-Jackson et al., 2015), we find that contributions from RFL, DOV and SPW provide a near-equal level of importance, with respective average percent contributions to model output of 30%, 31% and 28%, while ERA5 TMP and WVL variables have a total combined importance of 10%.

The combined insights from DeepPrecip's multi-model vertical extent evaluations and feature importance analyses demonstrate a potential to influence current and future remote sensing precipitation retrievals using deep learning. Instruments like CloudSat's Cloud Profiling Radar (CPR), or the Global Precipitation Measurement (GPM) mission's Dual-frequency Precipitation Radar (DPR) also use active radar systems to perform similar, radar-based precipitation retrievals based on data from vertical column reflectivities (Stephens et al., 2008). While CPR and GPM-derived products use a more sophisticated Bayesian retrieval to the $Z_e - S/R$ relationships evaluated here, the resulting precipitation estimates are still tightly coupled to a priori physical assumptions of particle shape, size and fallspeed which is a substantial source of uncertainty (Hiley et al., 2010; Wood et al., 2013). Additionally, the results of this study further support prior inference regarding the existence of regions of high importance in the < 1 km (near-surface) region of the vertical column relating to shallow-cumuliform precipitation strongly influencing retrieval accuracy. This is an area that is typically masked in satellite-based products (i.e. the radar "blind-zone") due to surface clutter contamination, and has been shown in previous work to likely be a major source of underestimation from missing shallow cumuliform precipitation (Maahn et al., 2014; Bennartz et al., 2019). This work motivates the importance of continued research towards obtaining high-quality, non-cluttered near surface radar data to use as additional model inputs in future space-based retrievals of precipitation.

DeepPrecip is not without uncertainty and error which will reduce its accuracy when tested against new data. Uncertainties present in the training data (stemming from the MRR, ERA5 or Pluvio2 observations), will propagate through the model and bias the output estimates (Kochendorfer et al., 2022; Jakubovitz et al., 2019). We have taken steps to mitigate the impact of these uncertainties through multiple data alignment and preprocessing decisions (details in Sect. 3), however precipitation gauge undercatch, wind shielding configurations, MRR attenuation and differences in site-specific vertical extent cannot be eliminated as contributors of retrieval error. While 60% of the power laws examined in this work were MRR-derived K-band relationships, the remaining 40% where either Ka-band or the Marshall-Palmer (MP) Rayleigh relationship. While K and Ka are similar radar frequencies, the differences between the two can bias the resulting precipitation estimate when a Ka-derived power law is applied to K-band data (especially during periods of intense precipitation). Furthermore, while the collection of data from multiple sites provides us with a robust training set under multiple regional climates, due to the unique experimental

setups at each site, calibration biases between study locations may further reduce DeepPrecip's skill when applied to new data. As the MRR instrument has a limited 3 km maximum vertical range, we also miss possible precipitation events occurring outside of this region, which may contribute to further surface precipitation underestimation. Internal CNN model uncertainty is likely driven, in part, by a combination of the high variability that is typical of precipitation and the limited sample from nine measurement sites over 8 years, which does not fully capture all different forms of possible precipitation structure and occurrence.

Code and data availability. DeepPrecip example code is fully open-source and available for download and use on the project's public GitHub repository (https://github.com/frasertheking/DeepPrecip). In situ data is freely accessible for download on Zenodo (https://doi.org/10.5281/zenodo.5976046). ERA5 hourly atmospheric data can be downloaded for free from the Copernicus Climate Change Service (C3S) Climate Data Store. The MRR and Pluvio data used in this study at OLYMPEx, GCPEx and ICE-POP were provided by NASA's Global Precipitation Measurement (GPM) Ground Validation program. POC: David B. Wolff, David.B.Wolff@nasa.gov. MRR and Pluvio data for the Cold Lake site was provided from ECCC observation sites. POC: Robert Crawford, Robert.Crawford@ec.gc.ca. JOYCE MRR and Pluvio data was provided by the University of Cologne. POC: Stefan Kneifel, skneifel@meteo.uni-koeln.de. Ny-Ålesund MRR and Pluvio data were provided by the German Alfred Wegener Institute for Polar and Marine Research and the French Polar Institute Paul Emile Victor. POC: Kerstin Ebell, kebell@meteo.uni-koeln.de. Marquette MRR and Pluvio data is provided by the Climate and Space Sciences and Engineering group at the University of Michigan. POC: Claire Pettersen, pettersc@umich.edu.

Author contributions. Project concept by FK and GD, data provided by FK, CP, KE and LM, methods developed by FK, GD and CF, model design by FK, data processing and experiments performed by FK, manuscript writing from FK, editing from FK, CF, GD, LM, CP and KE.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This study was supported by a grant from the Canadian Space Agency Earth System Science: Data Analyses fund. Further support was also provided by the Natural Sciences and Engineering Research Council of Canada. We also thank the data suppliers: Environment and Climate Change Canada (ECCC), the National Aeronautics and Space Administration (NASA), the Institute for Geophysics and Meteorology (IGM) at the University of Cologne, the Korean Meteorological Administration (KMA), the German Alfred Wegener Institute for Polar and Marine Research (AWI), the French Polar Institute Paul Emile Victor (IPEV), and the Climate and Space Sciences and Engineering group at the University of Michigan. This research would not have been possible without data contributions from David Wolff (DW), Claire Pettersen (CP) and Kerstin Ebell (KE). Finally, we would like to thank Graphcore for providing access to their high performance computing systems for training and optimizing the model.

KE appreciates the funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 268020496 - TRR172. The JOYCE data was provided by the Cloud and Precipitation Exploration Laboratory (CPEX-LAB, http://cpex-lab.de), a competence centre within the Geoverbund ABC/J.

References

Abdeljaber, O., Avci, O., Kiranyaz, S., Gabbouj, M., and Inman, D. J.: Real-time vibration-based structural damage detection using one-dimensional convolutional neural networks, Journal of Sound and Vibration, 388, 154–170, https://doi.org/10.1016/j.jsv.2016.10.043, 2017.

Adhikari, A., Ehsani, M. R., Song, Y., and Behrangi, A.: Comparative Assessment of Snowfall Retrieval From Microwave Humidity Sounders Using Machine Learning Methods, Earth and Space Science, 7, e2020EA001 357, https://doi.org/10.1029/2020EA001357, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2020EA001357, 2020.

Bennartz, R., Fell, F., Pettersen, C., Shupe, M. D., and Schuettemeyer, D.: Spatial and temporal variability of snowfall over Greenland from CloudSat observations, Atmospheric Chemistry and Physics, 19, 8101–8121, https://doi.org/10.5194/acp-19-8101-2019, publisher: Copernicus GmbH, 2019.

Boudala, F. S., Gultepe, I., and Milbrandt, J. A.: The Performance of Commonly Used Surface-Based Instruments for Measuring Visibility, Cloud Ceiling, and Humidity at Cold Lake, Alberta, Remote Sensing, 13, 5058, https://doi.org/10.3390/rs13245058, number: 24 Publisher: Multidisciplinary Digital Publishing Institute, 2021.

Buttle, J. M., Allen, D. M., Caissie, D., Davison, B., Hayashi, M., Peters, D. L., Pomeroy, J. W., Simonovic, S., St-Hilaire, A., and Whitfield, P. H.: Flood processes in Canada: Regional and special aspects, Canadian Water Resources Journal / Revue canadienne des ressources hydriques, 41, 7–30, https://doi.org/10.1080/07011784.2015.1131629, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/07011784.2015.1131629, 2016.

Chen, H., Chandrasekar, V., Cifelli, R., and Xie, P.: A Machine Learning System for Precipitation Estimation Using Satellite and Ground Radar Network Observations, IEEE Transactions on Geoscience and Remote Sensing, 58, 982–994, https://doi.org/10.1109/TGRS.2019.2942280, conference Name: IEEE Transactions on Geoscience and Remote Sensing, 2020a.

Chen, L., Cao, Y., Ma, L., and Zhang, J.: A Deep Learning-Based Methodology for Precipitation Nowcasting With Radar, Earth and Space Science, 7, e2019EA000 812, https://doi.org/10.1029/2019EA000812, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2019EA000812, 2020b.

Choubin, B., Khalighi-Sigaroodi, S., and Malekian, A.: Multiple linear regression, multi-layer perceptron network and adaptive neuro-fuzzy inference system for forecasting precipitation based on large-scale climate signals, Hydrological Sciences Journal, 61, 1001–1009, https://doi.org/10.1080/02626667.2014.966721, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/02626667.2014.966721, 2016.

Colli, M., Lanza, L. G., La Barbera, P., and Chan, P. W.: Measurement accuracy of weighing and tipping-bucket rainfall intensity gauges under dynamic laboratory testing, Atmospheric Research, 144, 186–194, https://doi.org/10.1016/j.atmosres.2013.08.007, 2014.

Ehsani, M. R. and Behrangi, A.: A comparison of correction factors for the systematic gauge-measurement errors to improve the global land precipitation estimate, Journal of Hydrology, 610, 127 884, https://doi.org/10.1016/j.jhydrol.2022.127884, 2022.

Ehsani, M. R., Behrangi, A., Adhikari, A., Song, Y., Huffman, G. J., Adler, R. F., Bolvin, D. T., and Nelkin, E. J.: Assessment of the Advanced Very High Resolution Radiometer (AVHRR) for Snowfall Retrieval in High Latitudes Using CloudSat and Machine Learning, Journal of Hydrometeorology, 22, 1591–1608, https://doi.org/10.1175/JHM-D-20-0240.1, publisher: American Meteorological Society Section: Journal of Hydrometeorology, 2021.

Gergel, D. R., Nijssen, B., Abatzoglou, J. T., Lettenmaier, D. P., and Stumbaugh, M. R.: Effects of climate change on snowpack and fire potential in the western USA, Climatic Change, 141, 287–299, https://doi.org/10.1007/s10584-017-1899-y, 2017.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horanyi, A., Muaoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Halm, E., Janiskova, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., and Thacpaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3803, 2020.

Hiley, M. J., Kulie, M. S., and Bennartz, R.: Uncertainty Analysis for CloudSat Snowfall Retrievals, Journal of Applied Meteorology and Climatology, 50, 399–418, https://doi.org/10.1175/2010JAMC2505.1, 2010.

Houze, R. A., McMurdie, L. A., Petersen, W. A., Schwaller, M. R., Baccus, W., Lundquist, J. D., Mass, C. F., Nijssen, B., Rutledge, S. A., Hudak, D. R., Tanelli, S., Mace, G. G., Poellot, M. R., Lettenmaier, D. P., Zagrodnik, J. P., Rowe, A. K., DeHart, J. C., Madaus, L. E., Barnes, H. C., and Chandrasekar, V.: The Olympic Mountains Experiment (OLYMPEX), Bulletin of the American Meteorological Society, 98, 2167–2188, https://doi.org/10.1175/BAMS-D-16-0182.1, publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, 2017.

Jakubovitz, D., Giryes, R., and Rodrigues, M. R. D.: Generalization Error in Deep Learning, in: Compressed Sensing and Its Applications: Third International MATHEON Conference 2017, edited by Boche, H., Caire, G., Calderbank, R., Kutyniok, G., Mathar, R., and Petersen, P., pp. 153–193, Springer International Publishing, Cham, 2019.

Jameson, A. R. and Kostinski, A. B.: Spurious power-law relations among rainfall and radar parameters, Quarterly Journal of the Royal Meteorological Society, 128, 2045–2058, https://doi.org/10.1256/003590002320603520, 2002.

Jash, D., Resmi, E. A., Unnikrishnan, C. K., Sumesh, R. K., Sreekanth, T. S., Sukumar, N., and Ramachandran, K. K.: Variation in rain drop size distribution and rain integral parameters during southwest monsoon over a tropical station: An inter-comparison of disdrometer and Micro Rain Radar, Atmospheric Research, 217, 24–36, https://doi.org/10.1016/j.atmosres.2018.10.014, 2019.

Jennings, K. S., Winchell, T. S., Livneh, B., and Molotch, N. P.: Spatial variation of the rain–snow temperature threshold across the Northern Hemisphere, Nature Communications, 9, 1148, https://doi.org/10.1038/s41467-018-03629-7, number: 1 Publisher: Nature Publishing Group, 2018.

Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gurel, N. M., Li, B., Zhang, C., Song, D., and Spanos, C.: Towards Efficient Data Valuation Based on the Shapley Value, arXiv:1902.10275 [cs, stat], http://arxiv.org/abs/1902.10275, arXiv: 1902.10275, 2020.

Kim, H.-U. and Bae, T.-S.: Preliminary Study of Deep Learning-based Precipitation, Journal of the Korean Society of Surveying, Geodesy, Photogrammetry and Cartography, 35, 423–430, https://doi.org/10.7848/ksgpc.2017.35.5.423, publisher: Korean Society of Surveying, Geodesy, Photogrammetry and Cartography, 2017.

Kim, K., Bang, W., Chang, E.-C., Tapiador, F. J., Tsai, C.-L., Jung, E., and Lee, G.: Impact of wind pattern and complex topography on snow microphysics during International Collaborative Experiment for PyeongChang 2018 Olympic and Paralympic winter games (ICE-POP 2018), Atmospheric Chemistry and Physics, 21, 11 955–11 978, https://doi.org/10.5194/acp-21-11955-2021, publisher: Copernicus GmbH, 2021.

King, F., Duffy, G., and Fletcher, C. G.: A Centimeter Wavelength Snowfall Retrieval Algorithm Using Machine Learning, Journal of Applied Meteorology and Climatology, -1, https://doi.org/10.1175/JAMC-D-22-0036.1, publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology, 2022.

Kochendorfer, J., Nitu, R., Wolff, M., Mekis, E., Rasmussen, R., Baker, B., Earle, M. E., Reverdin, A., Wong, K., Smith, C. D., Yang, D., Roulet, Y.-A., Buisan, S., Laine, T., Lee, G., Aceituno, J. L. C., Alastrua, J., Isaksen, K., Meyers, T., Braskkan, R., Landolt, S., Jachcik, A.,

and Poikonen, A.: Analysis of single-Alter-shielded and unshielded measurements of mixed and solid precipitation from WMO-SPICE, Hydrology and Earth System Sciences, 21, 3525–3542, https://doi.org/10.5194/hess-21-3525-2017, 2017.

Kochendorfer, J., Earle, M., Rasmussen, R., Smith, C., Yang, D., Morin, S., Mekis, E., Buisan, S., Roulet, Y.-A., Landolt, S., Wolff, M., Hoover, J., Thériault, J. M., Lee, G., Baker, B., Nitu, R., Lanza, L., Colli, M., and Meyers, T.: How Well Are We Measuring Snow Post-SPICE?, Bulletin of the American Meteorological Society, 103, E370–E388, https://doi.org/10.1175/BAMS-D-20-0228.1, publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, 2022.

Kulie, M. S. and Bennartz, R.: Utilizing Spaceborne Radars to Retrieve Dry Snowfall, Journal of Applied Meteorology and Climatology, 48, 2564–2580, https://doi.org/10.1175/2009JAMC2193.1, publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology, 2009.

Kulie, M. S., Pettersen, C., Merrelli, A. J., Wagner, T. J., Wood, N. B., Dutter, M., Beachler, D., Kluber, T., Turner, R., Mateling, M., Lenters, J., Blanken, P., Maahn, M., Spence, C., Kneifel, S., Kucera, P. A., Tokay, A., Bliven, L. F., Wolff, D. B., and Petersen, W. A.: Snowfall in the Northern Great Lakes: Lessons Learned from a Multisensor Observatory, Bulletin of the American Meteorological Society, 102, E1317–E1339, https://doi.org/10.1175/BAMS-D-19-0128.1, publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, 2021.

Lahnert, U., Schween, J. H., Acquistapace, C., Ebell, K., Maahn, M., Barrera-Verdejo, M., Hirsikko, A., Bohn, B., Knaps, A., OConnor, E., Simmer, C., Wahner, A., and Crewell, S.: JOYCE: Jaelich Observatory for Cloud Evolution, Bulletin of the American Meteorological Society, 96, 1157–1174, https://doi.org/10.1175/BAMS-D-14-00105.1, publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, 2015.

Lemonnier, F., Madeleine, J.-B., Claud, C., Genthon, C., Durán-Alarcón, C., Palerme, C., Berne, A., Souverijns, N., van Lipzig, N., Gorodetskaya, I. V., L'Ecuyer, T., and Wood, N.: Evaluation of CloudSat snowfall rate profiles by a comparison with in situ micro-rain radar observations in East Antarctica, The Cryosphere, 13, 943–954, https://doi.org/10.5194/tc-13-943-2019, publisher: Copernicus GmbH, 2019.

Levizzani, V., Laviola, S., and Cattani, E.: Detection and Measurement of Snowfall from Space, Remote Sensing, 3, 145–166, https://doi.org/10.3390/rs3010145, number: 1 Publisher: Molecular Diversity Preservation International, 2011.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A.: Hyperband: a novel bandit-based approach to hyperparameter optimization, The Journal of Machine Learning Research, 18, 6765–6816, 2017.

Li, L., Qiao, J., Yu, G., Wang, L., Li, H., Liao, C., and Zhu, Z.: Interpretable tree-based ensemble model for predicting beach water quality, Water Research, 211, 118 078, https://doi.org/10.1016/j.watres.2022.118078, 2022.

Liu, G.: Deriving snow cloud characteristics from CloudSat observations, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2007JD009766, 2008.

Louw, T. and McIntosh-Smith, S.: Using the Graphcore IPU for Traditional HPC Applications, AccML, https://easychair.org/publications/preprint/ztfj, number: 4896 Publisher: EasyChair, 2021.

Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, pp. 4768–4777, Curran Associates Inc., Red Hook, NY, USA, 2017.

Maahn, M. and Kollias, P.: Improved Micro Rain Radar snow measurements using Doppler spectra post-processing, Atmospheric Measurement Techniques, 5, 2661–2673, https://doi.org/10.5194/amt-5-2661-2012, publisher: Copernicus GmbH, 2012.

Maahn, M., Burgard, C., Crewell, S., Gorodetskaya, I. V., Kneifel, S., Lhermitte, S., Van Tricht, K., and van Lipzig, N. P. M.: How does the spaceborne radar blind zone affect derived surface snowfall statistics in polar regions?, Journal of Geophysical Research (Atmospheres), 119, 13,604–13,620, https://doi.org/10.1002/2014JD022079, aDS Bibcode: 2014JGRD..11913604M, 2014.

Marshall, J. S. and Palmer, W. M. K.: THE DISTRIBUTION OF RAINDROPS WITH SIZE, Journal of the Atmospheric Sciences, 5, 165–166, https://doi.org/10.1175/1520-0469(1948)005<0165:TDORWS>2.0.CO;2, publisher: American Meteorological Society Section: Journal of the Atmospheric Sciences, 1948.

Matrosov, S. Y.: Modeling Backscatter Properties of Snowfall at Millimeter Wavelengths, Journal of the Atmospheric Sciences, 64, 1727–1736, https://doi.org/10.1175/JAS3904.1, publisher: American Meteorological Society Section: Journal of the Atmospheric Sciences, 2007.

Matrosov, S. Y., Shupe, M. D., and Djalalova, I. V.: Snowfall Retrievals Using Millimeter-Wavelength Cloud Radars, Journal of Applied Meteorology and Climatology, 47, 769–777, https://doi.org/10.1175/2007JAMC1768.1, publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology, 2008.

Maxwell, A. and Shobe, C.: Land-surface parameters for spatial predictive mapping and modeling, Earth-Science Reviews, p. 103944, https://doi.org/10.1016/j.earscirev.2022.103944, 2022.

Munchak, S. J., Schrom, R. S., Helms, C. N., and Tokay, A.: Snow microphysical retrieval from the NASA D3R radar during ICE-POP 2018, Atmospheric Measurement Techniques, 15, 1439–1464, https://doi.org/10.5194/amt-15-1439-2022, publisher: Copernicus GmbH, 2022.

Pettersen, C., Kulie, M. S., Bliven, L. F., Merrelli, A. J., Petersen, W. A., Wagner, T. J., Wolff, D. B., and Wood, N. B.: A Composite Analysis of Snowfall Modes from Four Winter Seasons in Marquette, Michigan, Journal of Applied Meteorology and Climatology, 59, 103–124, https://doi.org/10.1175/JAMC-D-19-0099.1, publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology, 2020.

Quirita, V. A. A., da Costa, G. A. O. P., Happ, P. N., Feitosa, R. Q., Ferreira, R. d. S., Oliveira, D. A. B., and Plaza, A.: A New Cloud Computing Architecture for the Classification of Remote Sensing Data, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 10, 409–416, https://doi.org/10.1109/JSTARS.2016.2603120, conference Name: IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017.

Rasmussen, R., Baker, B., Kochendorfer, J., Meyers, T., Landolt, S., Fischer, A. P., Black, J., Thériault, J. M., Kucera, P., Gochis, D., Smith, C., Nitu, R., Hall, M., Ikeda, K., and Gutmann, E.: How Well Are We Measuring Snow: The NOAA/FAA/NCAR Winter Precipitation Test Bed, Bulletin of the American Meteorological Society, 93, 811–829, https://doi.org/10.1175/BAMS-D-11-00052.1, publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, 2012.

Schoger, S. Y., Moisseev, D., Lerber, A. v., Crewell, S., and Ebell, K.: Snowfall-Rate Retrieval for K- and W-Band Radar Measurements Designed in Hyyti, Finland, and Tested at Ny-Alesund, Svalbard, Norway, Journal of Applied Meteorology and Climatology, 60, 273–289, https://doi.org/10.1175/JAMC-D-20-0095.1, publisher: American Meteorological Society Section: Journal of Applied Meteorology and Climatology, 2021.

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.-Y., Wong, W.-k., and Woo, W.-c.: Deep Learning for Precipitation Nowcasting: A Benchmark and A New Model, NeurIPS, https://arxiv.org/abs/1706.03458v2, 2017.

Sims, E. M. and Liu, G.: A Parameterization of the Probability of Snow–Rain Transition, Journal of Hydrometeorology, 16, 1466–1477, https://doi.org/10.1175/JHM-D-14-0211.1, publisher: American Meteorological Society Section: Journal of Hydrometeorology, 2015.

Skofronick-Jackson, G., Hudak, D., Petersen, W., Nesbitt, S. W., Chandrasekar, V., Durden, S., Gleicher, K. J., Huang, G.-J., Joe, P., Kollias, P., Reed, K. A., Schwaller, M. R., Stewart, R., Tanelli, S., Tokay, A., Wang, J. R., and Wolde, M.: Global Precipitation Measurement Cold

Season Precipitation Experiment (GCPEX): For Measurement Sake Let It Snow, Bulletin of the American Meteorological Society, 96, 1719–1741, https://doi.org/10.1175/BAMS-D-13-00262.1, publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, 2015.

Skofronick-Jackson, G., Petersen, W. A., Berg, W., Kidd, C., Stocker, E. F., Kirschbaum, D. B., Kakar, R., Braun, S. A., Huffman, G. J., Iguchi, T., Kirstetter, P. E., Kummerow, C., Meneghini, R., Oki, R., Olson, W. S., Takayabu, Y. N., Furukawa, K., and Wilheit, T.: The Global Precipitation Measurement (GPM) Mission for Science and Society, Bulletin of the American Meteorological Society, 98, 1679–1695, https://doi.org/10.1175/BAMS-D-15-00306.1, publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society, 2017.

Souverijns, N., Gossart, A., Lhermitte, S., Gorodetskaya, I. V., Kneifel, S., Maahn, M., Bliven, F. L., and van Lipzig, N. P. M.: Estimating radar reflectivity - Snowfall rate relationships and their uncertainties over Antarctica by combining disdrometer and radar observations, Atmospheric Research, 196, 211–223, https://doi.org/10.1016/j.atmosres.2017.06.001, 2017.

Stephens, G. L., Vane, D. G., Tanelli, S., Im, E., Durden, S., Rokey, M., Reinke, D., Partain, P., Mace, G. G., Austin, R., L'Ecuyer, T., Haynes, J., Lebsock, M., Suzuki, K., Waliser, D., Wu, D., Kay, J., Gettelman, A., Wang, Z., and Marchand, R.: CloudSat mission: Performance and early science after the first year of operation, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2008JD009982, _eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2008JD009982, 2008.

Van Baelen, J., Tridon, F., and Pointin, Y.: Simultaneous X-band and K-band study of precipitation to derive specific ZR relationships, Atmospheric Research, 94, 596–605, https://doi.org/10.1016/j.atmosres.2009.04.003, 2009.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods, 17, 261–272, https://doi.org/10.1038/s41592-019-0686-2, 2020.

Wood, N. B., L'Ecuyer, T. S., Bliven, F. L., and Stephens, G. L.: Characterization of video disdrometer uncertainties and impacts on estimates of snowfall rate and radar reflectivity, Atmospheric Measurement Techniques, 6, 3635–3648, https://doi.org/10.5194/amt-6-3635-2013, publisher: Copernicus GmbH, 2013.

Xiao, R., Chandrasekar, V., and Liu, H.: Development of a neural network based algorithm for radar snowfall estimation, IEEE Transactions on Geoscience and Remote Sensing, 36, 716–724, https://doi.org/10.1109/36.673664, conference Name: IEEE Transactions on Geoscience and Remote Sensing, 1998.

Yang, D.: Double Fence Intercomparison Reference (DFIR) vs. Bush Gauge for "true" snowfall measurement, Journal of Hydrology, 509, 94–100, https://doi.org/10.1016/j.jhydrol.2013.08.052, 2014.

Yu, T., Chandrasekar, V., Xiao, H., and Joshil, S. S.: Characteristics of Snow Particle Size Distribution in the PyeongChang Region of South Korea, Atmosphere, 11, 1093, https://doi.org/10.3390/atmos11101093, number: 10 Publisher: Multidisciplinary Digital Publishing Institute, 2020.