Supplement of Atmos. Meas. Tech., 18, 5393–5414, 2025
https://doi.org/10.5194/amt-18-5393-2025-supplement

Supplement of

Classifying thermodynamic cloud phase using machine learning models

Lexie Goldberger et al.

Correspondence to: Damao Zhang (damao.zhang@pnnl.gov)

Table S1. Results of the CNN ablation study. We tested 15 different model architectures to determine the best result for classifying cloud phase. Column 1 is the model identifier. Columns 2 - 9 are different settings we applied: the presence of 2D spatial dropout layers; the number of convolutions in each block; the presence of batch normalization layers only, 1D dropout layers only, or both; the number of channels in the layers; the type of activation function used; and the use of class weights for the different cloud phases. Columns 10 - 13 are the training results, and columns 14 - 17 are the validation results. Training and validation used the same metrics to quantify their results: cloudy cross entropy, epoch loss, mean IOU, and sparse categorical accuracy. Model #15 was determined to be our 'best' CNN and was referred to as "CNN" in our paper, and model #5 performed 'second best' and was referred to as "CNN 2D dropout".

	Settings							
Test #	Spatial Dropout 2D	# Convolutions In Each Block	Batch Normalization only	1D Dropout Only	Batch Normalization & Dropout	# Channels in Layers	Activation Function on the Conv2D Layer(s)	Class Weights
1	TRUE	1	TRUE	FALSE	FALSE	64, 64, 64, 128, 128, 256	linear	FALSE
2	TRUE	1	FALSE	TRUE	FALSE	64, 64, 64, 128, 128, 256	linear	FALSE
3	TRUE	2	TRUE	FALSE	FALSE	64, 64, 64, 128, 128, 256	linear	FALSE
4	TRUE	2	FALSE	TRUE	FALSE	64, 64, 64, 128, 128, 256	linear	FALSE
5	**TRUE**	**2**	**FALSE**	**FALSE**	**TRUE**	**64, 64, 64, 128, 128, 256**	**linear**	**FALSE**
6	TRUE	3	TRUE	FALSE	FALSE	64, 64, 64, 128, 128, 256	linear	FALSE
7	TRUE	3	FALSE	TRUE	FALSE	64, 64, 64, 128, 128, 256	linear	FALSE
8	TRUE	3	FALSE	FALSE	TRUE	64, 64, 64, 128, 128, 256	linear	FALSE
9	TRUE	2	FALSE	FALSE	TRUE	64, 64, 64, 128, 128, 256	relu	FALSE
10	TRUE	2	FALSE	FALSE	TRUE	64, 64, 64, 128, 128, 256	leaky_relu	FALSE
11	TRUE	2	FALSE	FALSE	TRUE	64, 64, 64, 128, 128, 256	sigmoid	FALSE

12	TRUE	2	FALSE	FALSE	TRUE	32, 32, 32, 64, 64, 128	linear	FALSE
13	TRUE	2	FALSE	FALSE	TRUE	128, 128, 128, 256, 256, 512	linear	FALSE
14	TRUE	2	FALSE	FALSE	TRUE	64, 64, 64, 128, 128, 256	linear	TRUE
15	**FALSE**	**2**	**FALSE**	**FALSE**	**TRUE**	**64, 64, 64, 128, 128, 256**	**linear**	**FALSE**

Table S1 continued.

	Training Results				**Validation Results**			
Test #	Cloudy Cross-entropy	Epoch Loss	Mean IOU	Sparse Categorical Accuracy	Cloudy Cross-entropy	Epoch Loss	Mean IOU	Categorical Accuracy
1	0.297	0.0642	0.51	0.977	0.312	0.0569	0.508	0.979
2	2.3 E12	10 E 12	0.02	0.7	3.9 E12	7.01E+14	0.02	0.82
3	0.274	0.059	0.529	0.979	0.299	0.052	0.52	0.981
4	NAN	NAN	0	0.8	NAN	NAN	0	0.82
5	**0.202**	**0.044**	**0.604**	**0.984**	**0.297**	**0.054**	**0.528**	**0.982**
6	0.232	0.051	0.575	0.982	0.251	0.046	0.563	0.985
7	0.295	0.064	0.515	0.977	0.322	0.059	0.501	0.977
8	0.231	0.05	0.578	0.982	0.317	0.054	0.505	0.98
9	0.25	0.0548	0.56	0.98	0.3	0.0545	0.51	0.98
10	0.22	0.046	0.596	0.983	0.28	0.05	0.535	0.982
11	0.23	0.05	0.57	0.982	0.31	0.056	0.515	0.98
12	0.233	0.051	0.572	0.982	0.327	0.06	0.507	0.98
13	0.24	0.053	0.574	0.982	0.305	0.056	0.509	0.9805
14	0.286	0.083	0.535	0.977	0.3	0.053	0.5	0.979
15	**0.09**	**0.018**	**0.732**	**0.993**	**0.106**	**0.0192**	**0.733**	**0.993**

Table S2. Instrument Dropout Study Full Results for the CNN model.

CNN Model Results		**Intersection over Union (IOU) Score**								
Model	Missing datastream/ Instrument	Drizzle	Ice	Liquid Drizzle	Liquid	Mixed	Rain	Snow	Mean IOU	Total Accuracy %
CNN	**Control**	**0.709**	**0.958**	**0.636**	**0.788**	**0.768**	**0.883**	**0.932**	**0.811**	**95.7**
CNN	Micropulse Lidar, all datastreams	0.661	0.899	0.618	0.437	0.567	0.877	0.939	0.714	91.0
CNN	Micropulse Lidar, backscatter	0.675	0.902	0.593	0.445	0.598	0.886	0.931	0.718	91.3
CNN	Micropulse Lidar, linear depolarization ratio	0.662	0.917	0.633	0.587	0.613	0.894	0.937	0.749	92.4
CNN	Microwave Radiometer	0.706	0.954	0.637	0.779	0.754	0.885	0.932	0.807	95.4
CNN	Radar, all datastreams	0.204	0.797	0.040	0.151	0.201	0.000	0.000	0.199	75.6
CNN	Radar, linear depolarization ratio	0.709	0.957	0.649	0.781	0.762	0.883	0.932	0.810	95.6
CNN	Radar, mean doppler velocity	0.261	0.955	0.078	0.714	0.732	0.661	0.928	0.619	94.0
CNN	Radar, reflectivity	0.357	0.846	0.365	0.224	0.434	0.753	0.000	0.426	81.1
CNN	Radar, spectral width	0.693	0.891	0.619	0.654	0.329	0.868	0.929	0.712	90.3
CNN	Radiosonde Temperature	0.000	0.935	0.002	0.258	0.492	0.063	0.811	0.366	88.1

Table S3. Same as Table S2, but for the MLP model.

MLP Model Results		**Intersection over Union (IOU) Score**								
Model	Missing datastream/ Instrument	Drizzle	Ice	Liquid Drizzle	Liquid	Mixed	Rain	Snow	Mean IOU	Total Accuracy %
MLP	**Control**	**0.677**	**0.810**	**0.571**	**0.499**	**0.489**	**0.928**	**0.893**	**0.695**	**84.6**
MLP	Micropulse Lidar, all datastreams	0.658	0.834	0.509	0.490	0.511	0.925	0.877	0.686	86.1
MLP	Micropulse Lidar, backscatter	0.660	0.812	0.502	0.482	0.482	0.927	0.891	0.680	84.6
MLP	Micropulse Lidar, linear depolarization ratio	0.629	0.792	0.490	0.437	0.483	0.921	0.877	0.661	83.1
MLP	Microwave Radiometer	0.548	0.774	0.689	0.441	0.461	0.747	0.892	0.650	81.8
MLP	Radar, all datastreams	0.194	0.770	0.041	0.000	0.216	0.000	0.000	0.174	71.7
MLP	Radar, linear depolarization ratio	0.660	0.778	0.546	0.452	0.459	0.926	0.899	0.674	82.3
MLP	Radar, mean doppler velocity	0.318	0.802	0.116	0.462	0.454	0.754	0.893	0.543	82.7
MLP	Radar, reflectivity	0.290	0.771	0.259	0.047	0.296	0.745	0.000	0.344	71.2
MLP	Radar, spectral width	0.661	0.829	0.566	0.472	0.327	0.833	0.887	0.653	85.2
MLP	Radiosonde Temperature	0.000	0.778	0.000	0.086	0.280	0.000	0.782	0.275	74.3

Table S4. Same as Table S2, but for the RF model.

RF Model Results		**Intersection over Union (IOU) Score**								
Model	Missing datastream/ Instrument	Drizzle	Ice	Liquid Drizzle	Liquid	Mixed	Rain	Snow	Mean IOU	Total Accuracy %
RF	**Control**	**0.714**	**0.824**	**0.603**	**0.510**	**0.512**	**0.940**	**0.901**	**0.715**	**85.8**
RF	Micropulse Lidar, all datastreams	0.688	0.840	0.515	0.487	0.529	0.938	0.899	0.699	86.6
RF	Micropulse Lidar, backscatter	0.700	0.834	0.524	0.496	0.517	0.939	0.900	0.701	86.3
RF	Micropulse Lidar, linear depolarization ratio	0.698	0.825	0.559	0.485	0.527	0.939	0.900	0.705	85.8
RF	Microwave Radiometer	0.715	0.786	0.691	0.464	0.471	0.917	0.902	0.706	83.1
RF	Radar, all datastreams	0.204	0.772	0.026	0.000	0.204	0.000	0.000	0.172	72.8
RF	Radar, linear depolarization ratio	0.708	0.815	0.579	0.496	0.501	0.939	0.905	0.706	85.1
RF	Radar, mean doppler velocity	0.301	0.827	0.081	0.473	0.489	0.751	0.901	0.546	84.4
RF	Radar, reflectivity	0.284	0.790	0.260	0.030	0.311	0.766	0.000	0.348	72.7
RF	Radar, spectral width	0.726	0.836	0.641	0.483	0.321	0.919	0.895	0.689	86.0
RF	Radiosonde Temperature	0.000	0.797	0.000	0.084	0.295	0.000	0.801	0.282	75.9

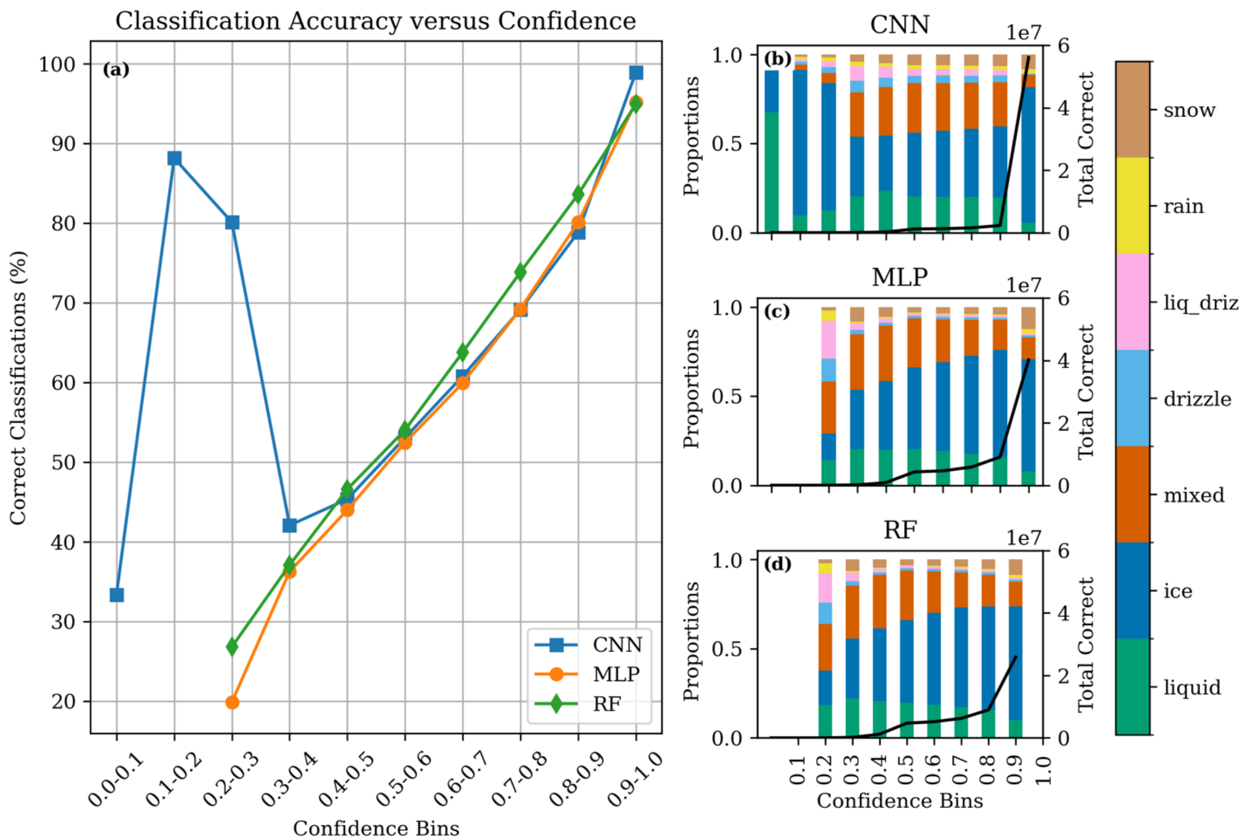

Figure S1. Distribution of accurate classifications with confidence score for test data from NSA in 2021. (a) percent of correct classifications with respect to confidence score binned by 10% intervals, (b-d) proportion of correctly classified pixels per confidence score bin, black line shows how many pixels in each bin.

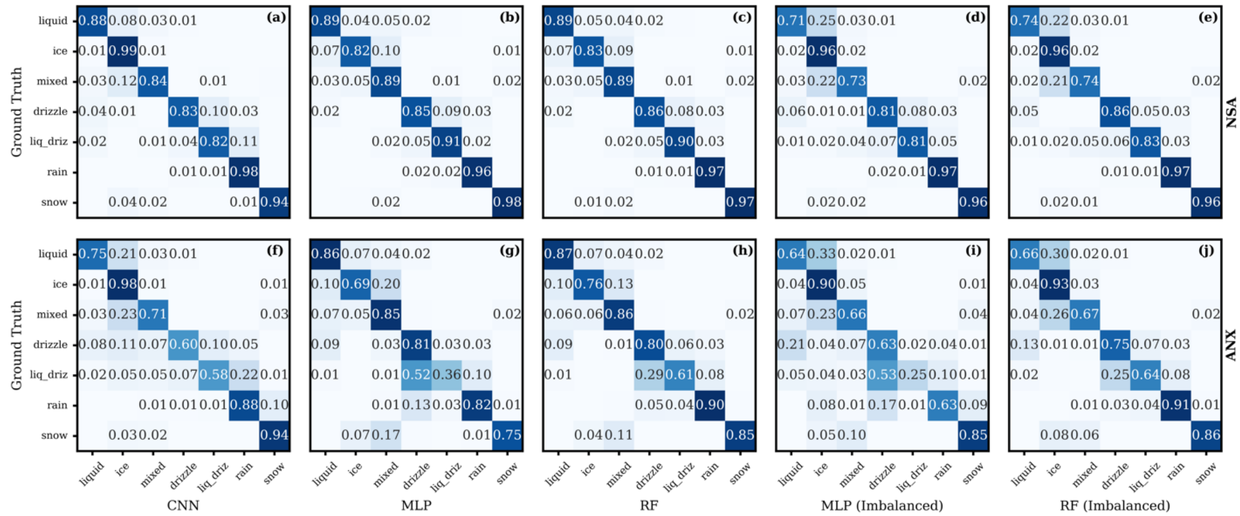

Figure S2. Confusion matrices computed on the 12 months data in 2021 at the NSA site for (a) the CNN U-net model, (b) the MLP model, (c) the RF model, (d-e) and the imbalanced MLP and RF models closest resembling the class fractions present in the CNN training data. Confusion matrices computed on the 4 months data in 2020 at the ANX site for (f) the CNN U-net model, (g) the MLP model, (h) the RF model, (i) the imbalanced MLP model, and (j) the imbalanced RF model. The values are normalized by row, with the main diagonal showing true positive predictions. Within a given row, values off the main diagonal represent false negatives.

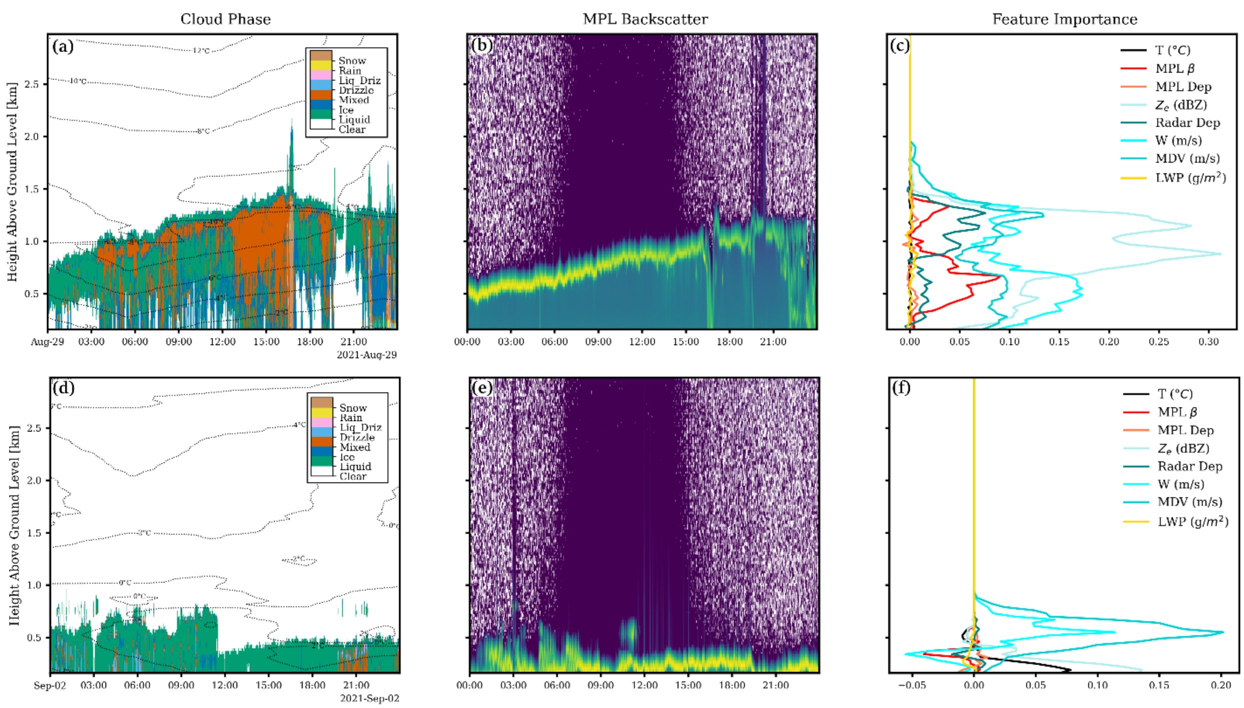

Figure S3. Permutation feature importance calculated for two sample days with single-layer, low-level liquid clouds to examine the lidar measurement's lower feature importance. Panels are: (a,d) time series of thermodynamic cloud phase from the VAP, (b,e) time series of the MPL backscatter (color bar in Figure S4a), and (c,f) vertically resolved feature importance score calculated from the CNN model. A negative feature importance score indicates the input is contributing negatively to the importance, instead contributing noise to the signal. On August 29th, 2021 (a,b,c) MPL backscatter importance peaks at cloud base and has a secondary peak at cloud top. On September 2nd, 2021 (d,e,f) likely demonstrates a fog layer where the lidar signal attenuates near surface and provides minimal contribution to feature importance. In both cases radar measurements (in blue) provide higher contributions to importance relative to lidar measurements.

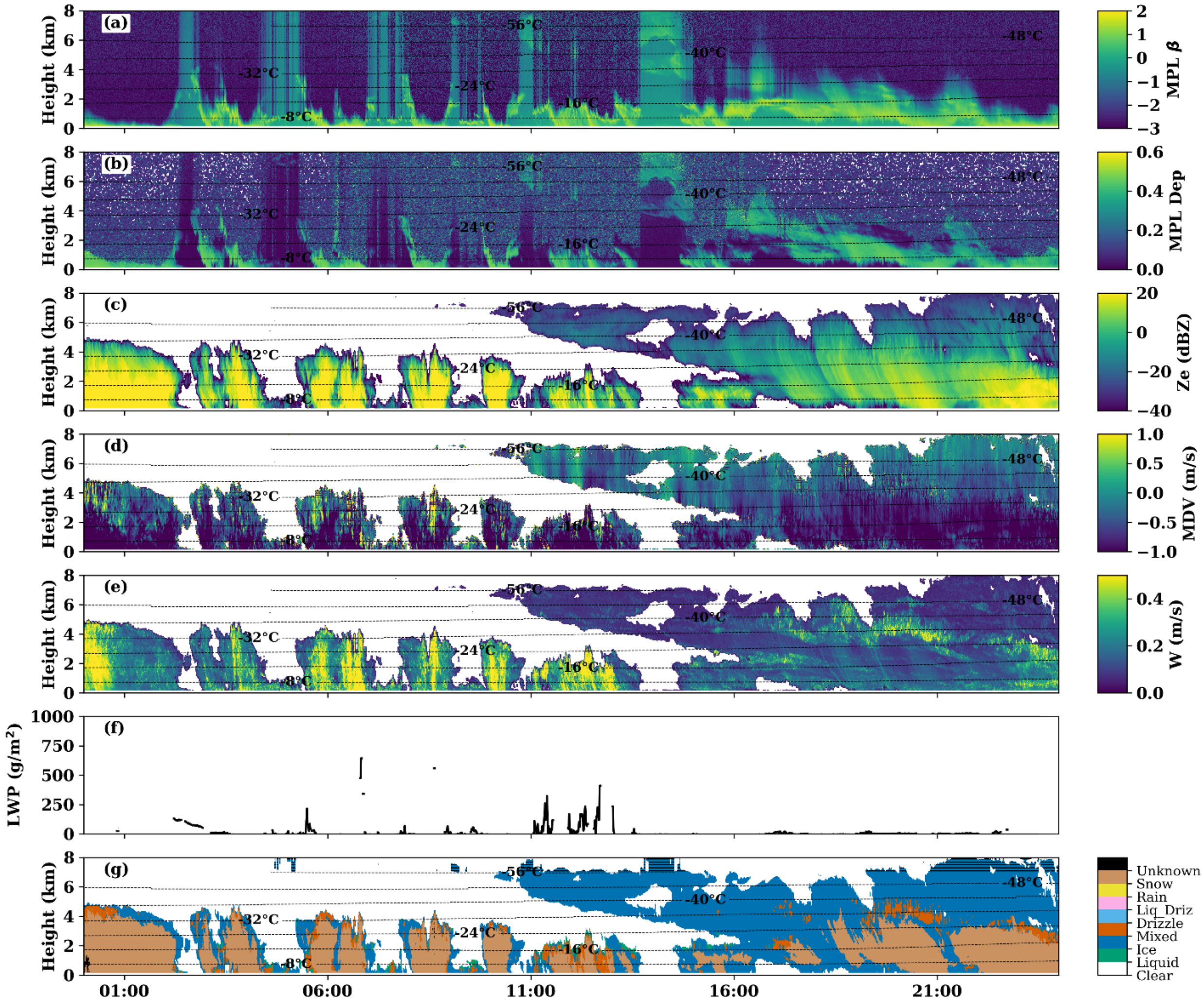

Figure S4. Multi-sensor remote sensing measurements of clouds and the thermodynamic cloud phase classification from the THERMOCLDPHASE VAP on February 25, 2020, at the ARM ANX site. Panels from top to bottom are: a) MPL attenuated backscatter (MPL β); b) MPL linear depolarization ratio (MPL LDR); c) Ka-band ARM zenith radar (KAZR) radar equivalent reflectivity factor (Z_e); d) KAZR mean Doppler velocity (MDV); e) KAZR Doppler spectral width (W), f) liquid water path (LWP) from the MWRRET VAP; and g) the thermodynamic cloud phase classification from the THERMOCLDPHASE VAP. Unknown refers to pixels the THERMOCLDPHASE algorithm was unable to resolve. The dashed lines in g) are isothermal lines based on the ARM Interpolated Sonde (INTERPSONDE) VAP.

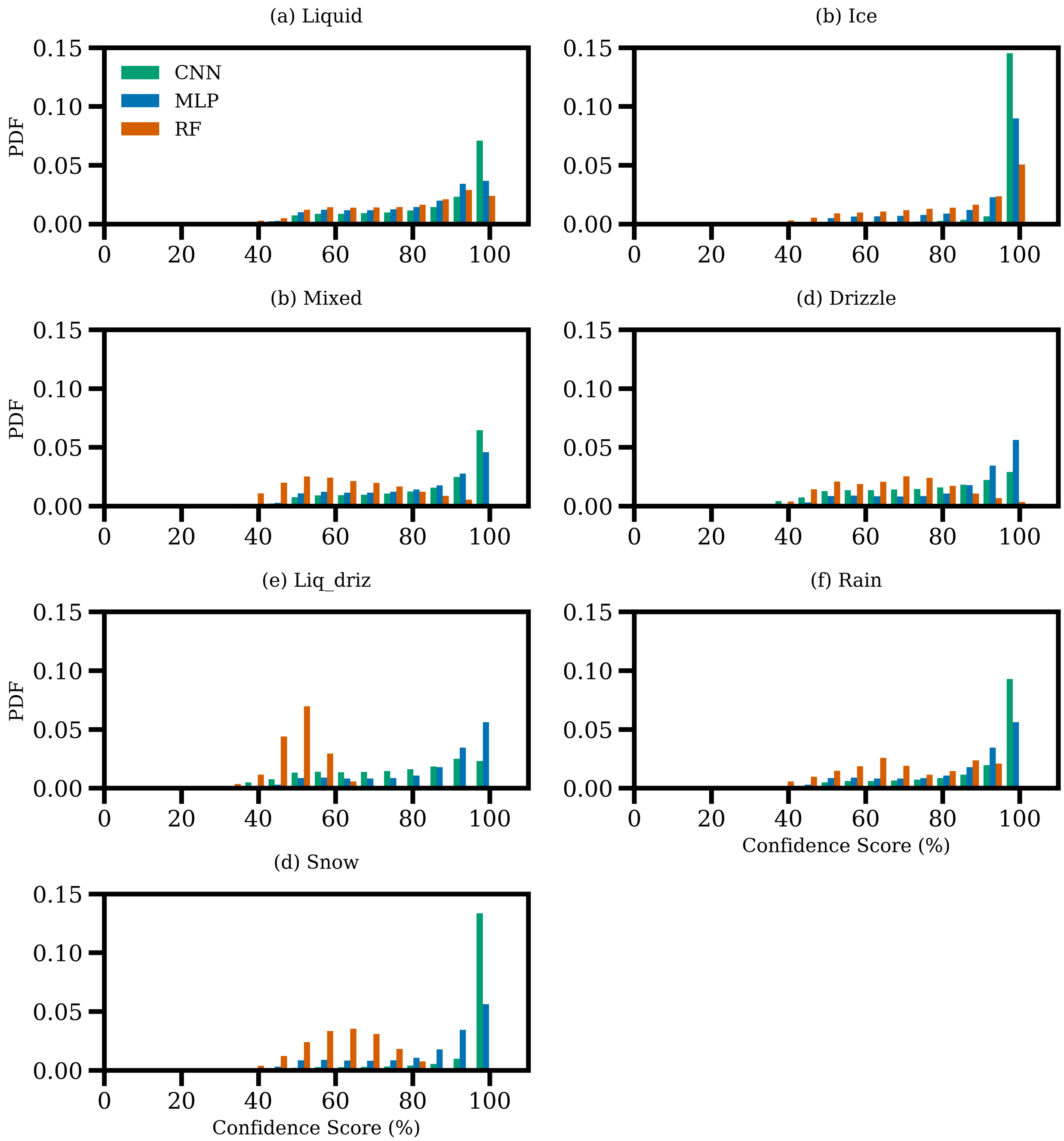

Figure S5. Probability Density Functions (PDFs) of confidence scores for thermodynamic cloud phase predictions from the three ML models using approximately 4 months of data in 2020 at the ANX site.

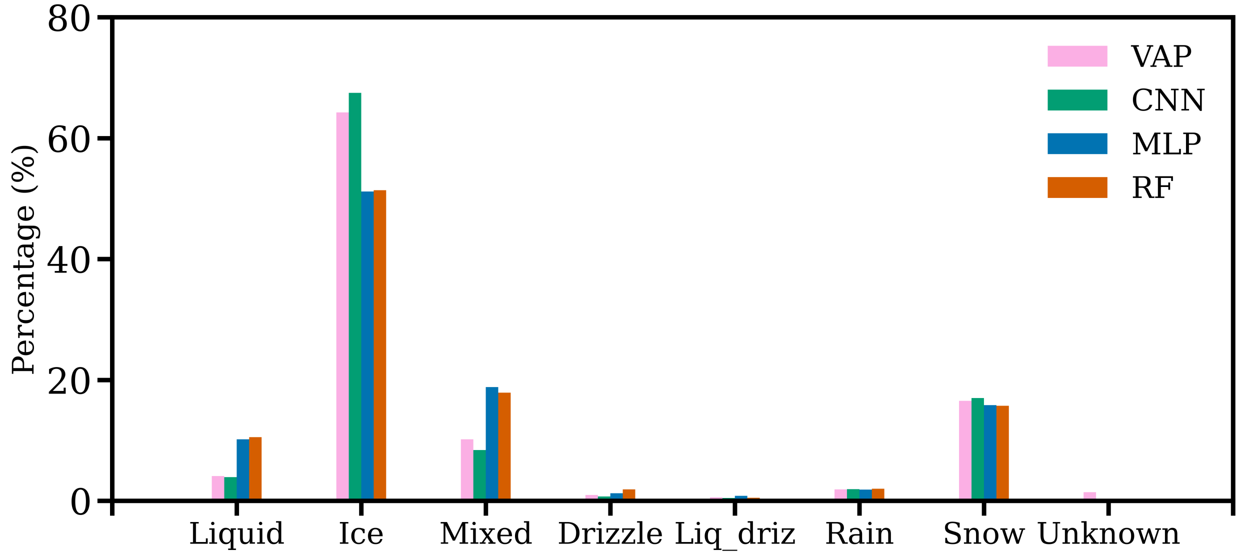

Figure S6. Percentage distributions of thermodynamic cloud phases from the THERMOCLDPHASE VAP (labeled as 'VAP') and predictions from the three ML models, based on approximately 4 months data in 2020 at the ANX site.

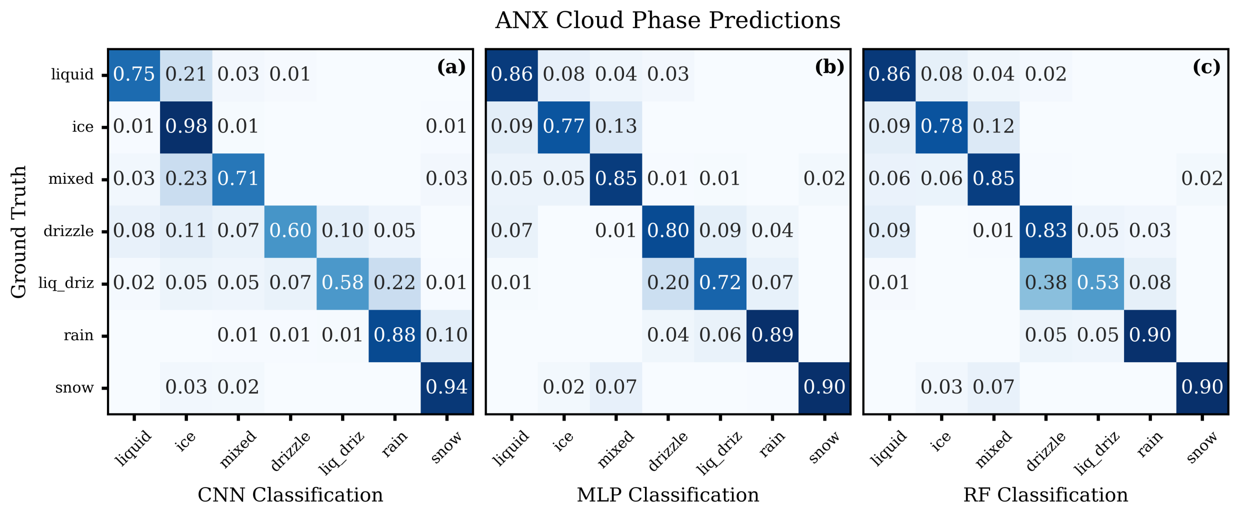

Figure S7. Confusion matrices computed on the 4 months data in 2020 at the ANX site for (a) the CNN U-net model, (b) the MLP model, and (c) the RF model. The values are normalized by row, with the main diagonal showing true positive predictions. Within a given row, values off the main diagonal represent false negatives.

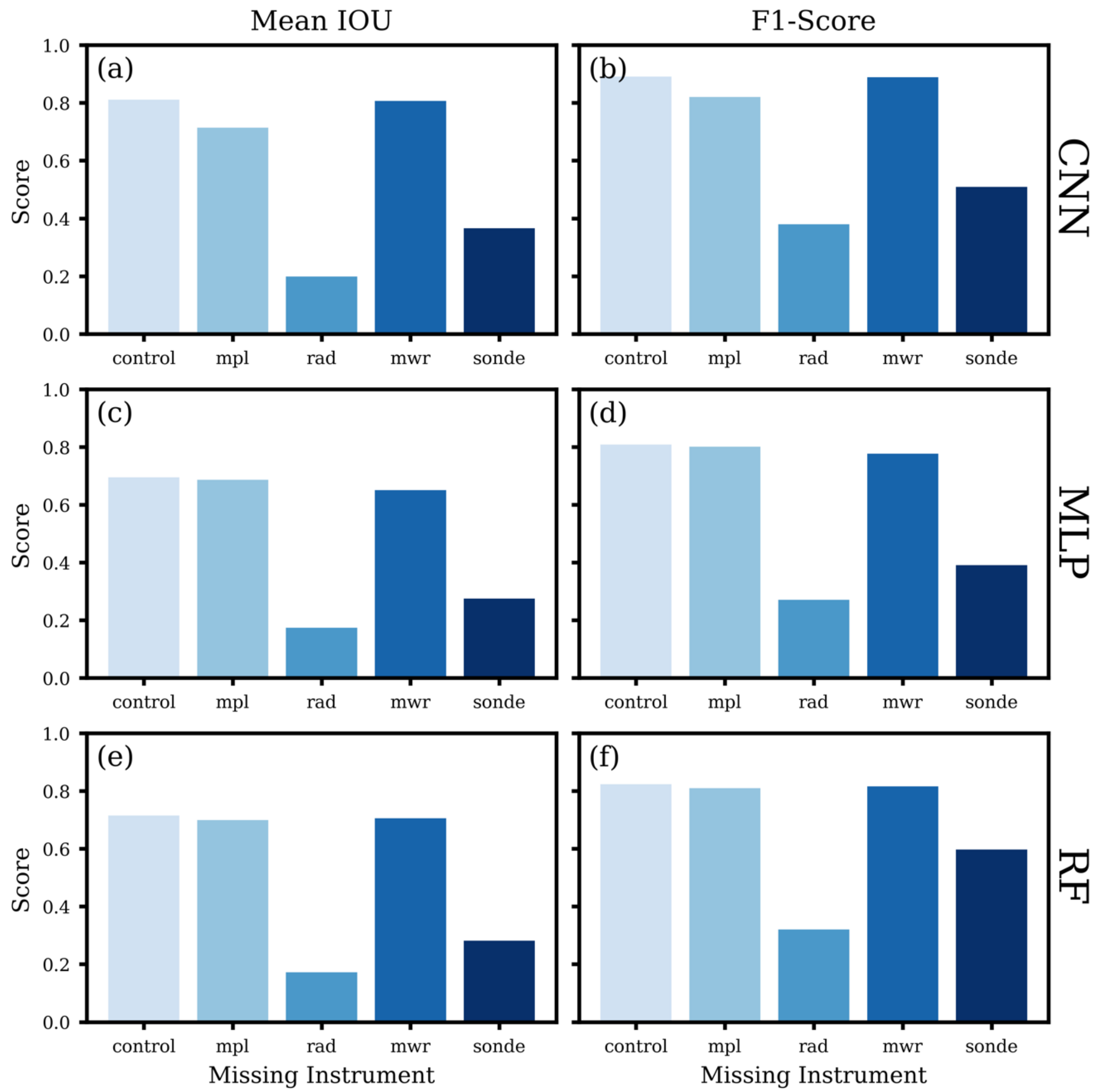

Figure S8. Instrument dropout results showing model scores resulting from removing all inputs from each of the MLP, Radar, MWR, and Radiosonde inputs. Control shows the baseline model score, with no information removed. Plots (a) and (b) show Mean IOU and F1-scores for the CNN model. Plots (c) and (d) show Mean IOU and F1-scores for the MLP model. Plots (e) and (f) show Mean IOU and F1-scores for the RF model.

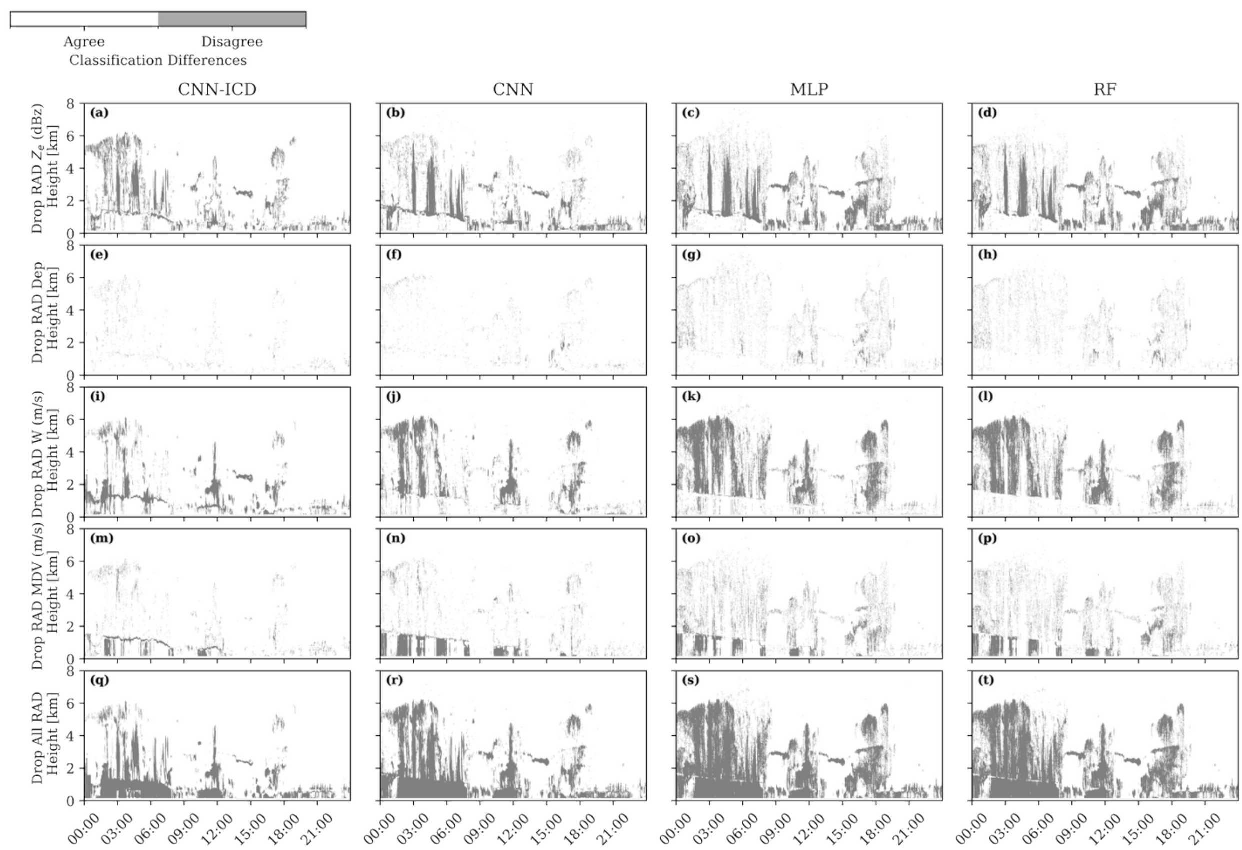

Figure S9. Differences in thermodynamic cloud phase classifications between model predictions with and without the corresponding dropout variable data for the four ML models on August 15, 2021, at the NSA site. (a-d) dropping out Radar Ze; (e-h) dropping our Radar Dep; (i-l) dropping out radar W; (m-p) dropping out radar MDV; (q-t) dropping out all Radar variables.

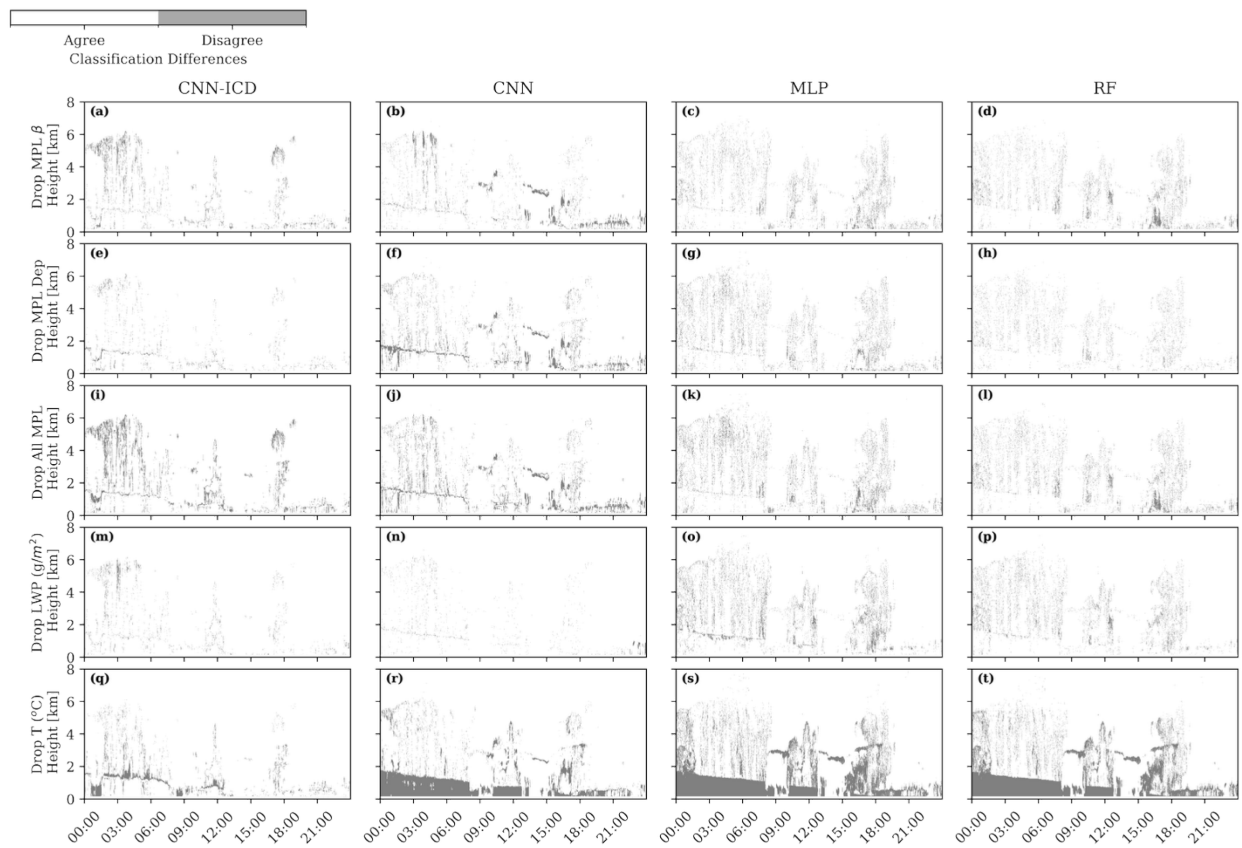

Figure S10. Similar to Figure S9, but for dropping out different variable data. (a-d) dropping out MPL Beta; (e-h) dropping our MPL Dep; (i-l) dropping out all MPL variables; (m-p) dropping out LWP; (q-t) dropping out T.