Atmospheric Research 240 (2020) 104928

Contents lists available at ScienceDirect

Atmospheric Research

journal homepage: www.elsevier.com/locate/atmosres

An improved forecast of precipitation type using correlation-based feature selection and multinomial logistic regression

Seung-Hyun Moon[a], Yong-Hyuk Kim[b,*]

[a] Department of Computer Science & Engineering, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea
[b] School of Software, Kwangwoon University, 20 Kwangwoon-ro, Nowon-gu, Seoul 01897, Republic of Korea

ARTICLE INFO

Keywords:
Precipitation type
Precipitation phase
Multinomial logistic regression
Correlation-based feature selection

ABSTRACT

Accurate prediction of precipitation type is an important part of weather forecasting. But using meteorological insight to make such predictions from a small set of weather variables achieves only limited success. We use correlation-based feature selection to assemble an effective subset of the large number of weather variables available in short-range weather forecasts, and from these we obtain the coefficients for multinomial regression, which can then be used to predict precipitation type. We applied this approach to data for significant locations in South Korea, obtained from the European Centre for Medium-Range Weather Forecasts and from the Regional Data Assimilation and Prediction System, and achieved predictions which are respectively 15% and 13% more accurate than those contained in the original forecasts.

1. Introduction

Precipitation refers to all forms of the condensed atmospheric water vapor that falls on the ground. In winter, rain, snow, and sleet (which is partly melted falling snow) have different consequences. For example, rainfall percolates quickly into the ground, but snow accumulates on the surface, and melts later (Zhong et al., 2018). A covering of snow increases albedo and alters the surface energy budget (Box et al., 2012). Precipitation also affects travels: the highest risk of road accidents is associated with rain or sleet falling on a frozen road surface (Norrman et al., 2000). Thus, the accurate determination of precipitation type is important, especially when the temperature is near the freezing point (Ralph et al., 2005).

Surprisingly, precipitation type is not monitored by most meteorological stations, and these data are often unavailable (Liu et al., 2018). More attention has been put to the correct classification of precipitation when it occurs, than to the forecasting of precipitation types. The most common meteorological variables used for predicting precipitation types are temperatures, among which surface air temperature is the key predictor (Froidurot et al., 2014). Many schemes have been presented for predicting precipitation type using threshold values of air temperature (Gao et al., 2010; Hynčica and Huth, 2019; Kienzle, 2008; Lindström et al., 1997; Wigmosta et al., 1994; Yang et al., 1997) or an S-shaped curve that describes the relation between precipitation type and air temperature (Dai, 2008; Liu et al., 2018).

It has been suggested (Behrangi et al., 2018; Ding et al., 2014; Froidurot et al., 2014) that wet-bulb temperature is a better predictor of precipitation type than air temperature. However, Froidurot et al. (2014) have reported that the wet bulb temperature is rarely measured and its calculation requires a recursive algorithm which makes this model less suitable for operational purposes, and Behrangi et al. (2018) assert that a combination of dew-point temperature with surface air temperature can provide effective classification of precipitation types. However, there have also been reports (Chen et al., 2014; Zhong et al., 2018) of wet-bulb temperatures failing to provide a significant advantage over air temperatures.

There are meteorological quantities different from temperatures that can affect precipitation type. For example, relative humidity is known to influence precipitation (Ding et al., 2014), and wind speed can change the phase of precipitation (Behrangi et al., 2018). In addition, the thicknesses of various pressure layers are used to differentiate precipitation types (Keeter and Cline, 1991), and the vertical temperature lapse rate affects the type of precipitation (Sims and Liu, 2015). At high altitudes, surface pressure can also affect precipitation (Dai, 2008).

A more complicated model is needed to predict precipitation types using multiple meteorological variables rather than temperature thresholds. Such models are difficult to be constructed using meteorological insight alone, and a promising alternative is the mathematical analysis of historic weather data to synthesize a model that uses

* Corresponding author.
E-mail addresses: shmoon@soar.snu.ac.kr (S.-H. Moon), yhdfly@kw.ac.kr (Y.-H. Kim).

https://doi.org/10.1016/j.atmosres.2020.104928
Received 10 October 2019; Received in revised form 21 February 2020; Accepted 22 February 2020
Available online 24 February 2020
0169-8095/

weather variables more effectively in generating forecasts. Logistic regression is a technique that has been used to build a model from multiple meteorological variables to predict whether precipitation will occur as rain or snow. For example, Behrangi et al. (2018) used a principal component analysis (PCA) to decorrelate candidate predictors, and then predicted rain or snow using logistic regression; Froidurot et al. (2014) used logistic regression on the data from 14 Swiss weather stations for the same purpose; and Jennings et al. (2018) mapped rain-snow temperature thresholds over the Northern Hemisphere using logistic regression on air temperature, relative humidity, and atmospheric pressure. Precipitation types can also be predicted using decision trees. For instance, Lee et al. (2014) used air temperature, relative humidity, and thickness data for pressure layers between 1000 and 850 to construct a decision tree that distinguishes rain, snow, and sleet in South Korean weather; and Reeves et al. (2016) produced a decision tree to identify six types of precipitation: rain, snow, sleet, freezing rain, ice pellets, and a mix of freezing rain and ice pellet.

In this paper, we introduce a new method of predicting the precipitation types already included in short-range weather forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF) and the Regional Data Assimilation and Prediction System (RDAPS). Meteorological variables made available in these weather forecasts provide the input to our scheme, in which a feature selection technique is used to determine discriminatory subsets of the candidate predictors, and multinomial logistic regression was used for prediction. Experiments using different forecasting models were conducted on data from 22 significant locations in South Korea covering 3-hour intervals over periods from 3 to 72 h.

The rest of the paper is organized as follows: in Section 2 we introduce the datasets and meteorological variables that we use. In Section 3 we give details of the feature selection method and multinomial logistic regression. In Section 4 we present the experimental setup and results. Finally, we draw conclusions and discuss future research directions in Section 5.

2. Data

There are 780 automatic weather stations in South Korea. Every station has a binary sensor that detects the occurrence of precipitation. The types of precipitation, however, are manually recorded only at 22 significant locations every hour. Fig. 1 shows these 22 locations, which are major cities and islands, including Seoul. The latitude, longitude, and altitude of these locations, together with a representation of the range of temperatures commonly experienced over the winter (December to February) and non-winter (March to November) periods, and frequencies of the different winter precipitation types at each site are given in Appendix B.

The Korea Meteorological Administration provided our study with the precipitation-type data for the significant locations. Winter precipitation data for January 1, 2013 to December 31, 2015 at these locations are used for training and evaluating our scheme. Table 1 shows the frequency of each precipitation type over this period: it is mostly rainy particularly from March to November.

The ECMWF and RDAPS short-range weather forecasts are announced twice a day at 00:00 UTC and 12:00 UTC. Each forecast predicts weather variables at 3-hour intervals, up to 72 h ahead. The precipitation type at the start of each interval is predicted from the weather variables forecast for that time. The 93 weather variables, which are listed in Appendix A, include temperatures, wind speed, and relative humidity, as well as the forecast of precipitation type on which we aim to improve.

We have predicted precipitation types from short-range weather-forecast data, and assessed them against observed precipitation. Since we excluded the data having no precipitation, only 12.35% of winter data were used.

3. Methods

We use a machine learning approach, in which the subset of the available candidate predictors is selected by correlation-based feature selection, and multinomial logistic regression is then used to predict precipitation types for each forecasting interval. The proposed method is evaluated by accuracy, Heidke skill score, and Peirce skill score.

3.1. Nomenclature

The nomenclature used in this section is given below. See Cover and Thomas (2006) for background information theory.

- $(\mathbf{x}, y)$ is an M-tuple in which $\mathbf{x}$ contains values of the $M(=93)$ random variables X_m $(1 \le m \le M)$ which correspond to the measurable quantities listed in Table A.1, and y is the resulting precipitation forecast. That is another random variable Y, which can have one of three values, where 1 means rain, 2 means snow, and 3 means sleet.
- $\mathrm{H}(A)$ is an entropy of a discrete random variable A. The random variable A has a set of possible values α and a probability mass function $p(a) = \Pr(A = a)$, $a \in \alpha$. Then $\mathrm{H}(A)$ is defined by

$$\mathrm{H}(A) = -\sum_{a\in\alpha} p(a)\log p(a)$$

The entropy measures the uncertainty of a random variable, and the conditional entropy of two random variables A and B with a joint probability mass function $p(a,b)$ is defined as

$$\mathrm{H}(A \mid B) = -\sum_{a\in\alpha, b\in\beta} p(a,b)\log\frac{p(a,b)}{p(b)}$$

where β is a set of possible values for B, and $p(b)$ is a probability mass function of B. The conditional entropy $\mathrm{H}(A|B)$ quantifies the amount of randomness in A given the value of B.

- $\mathrm{I}(A,B)$ is a mutual information of two random variables A and B. The mutual information is calculated as

$$\begin{aligned}
\mathrm{I}(A,B) &= \sum_{a\in\alpha, b\in\beta} p(a,b)\log\frac{p(a,b)}{p(a)p(b)} \\
&= \sum_{a\in\alpha, b\in\beta} p(a,b)\log\frac{p(a\mid b)}{p(a)} \\
&= -\sum_{a\in\alpha, b\in\beta} p(a,b)\log p(a) + \sum_{a\in\alpha, b\in\beta} p(a,b)\log\frac{p(a,b)}{p(b)} \\
&= -\sum_{a\in\alpha} p(a)\log p(a) - \left(-\sum_{a\in\alpha, b\in\beta} p(a,b)\log\frac{p(a,b)}{p(b)}\right) \\
&= \mathrm{H}(A) - \mathrm{H}(A\mid B)
\end{aligned}$$

Thus, the mutual information quantifies the reduction in the uncertainty of one random variable through observing the other random variable.

- $\mathrm{U}(A,B)$ is a symmetric uncertainty of two random variables A and B

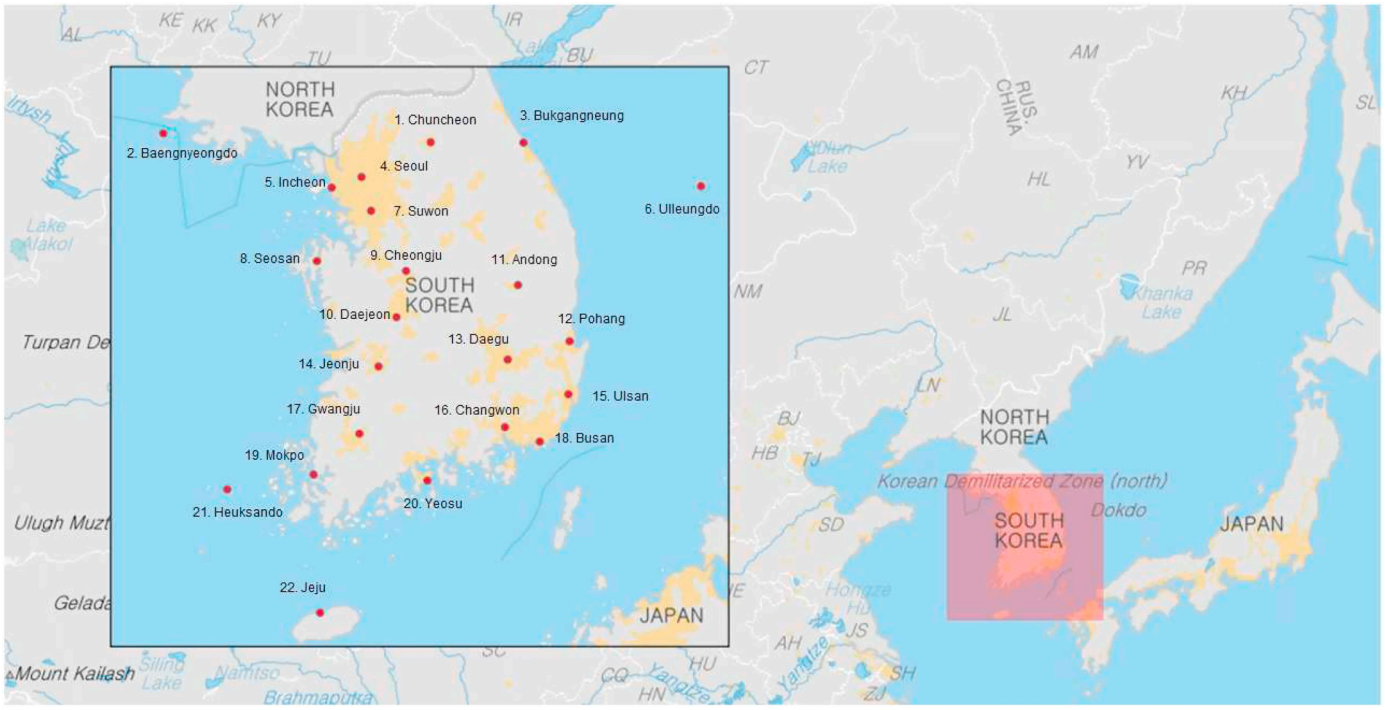

Fig. 1. Locations of the 22 significant locations in South Korea.

Table 1
Monthly occurrences of each precipitation type at 22 significant locations. The numbers of occurrences were counted by precipitation observed in 3-hour intervals from 2013 to 2015.

Month	Rain		Snow		Sleet	
Jan	867	(53%)	626	(38%)	153	(9%)
Feb	1044	(57%)	666	(36%)	122	(7%)
Mar	1181	(90%)	90	(7%)	39	(3%)
Apr	2019	(99%)	1	(0%)	18	(1%)
May	1289	(100%)	0	(0%)	0	(0%)
Jun	1760	(100%)	0	(0%)	0	(0%)
Jul	2782	(100%)	0	(0%)	0	(0%)
Aug	2530	(100%)	0	(0%)	0	(0%)
Sep	1683	(100%)	0	(0%)	0	(0%)
Oct	1229	(100%)	0	(0%)	0	(0%)
Nov	2233	(87%)	220	(9%)	112	(4%)
Dec	1074	(45%)	980	(41%)	335	(14%)

defined by

$$\mathrm{U}(A,B) = \frac{2\cdot \mathrm{I}(A,B)}{\mathrm{H}(A) + \mathrm{H}(B)}$$

The symmetric uncertainty normalizes mutual information so that its value lies between 0 and 1.

3.2. Correlation-based feature selection

The goal of feature selection is to obtain the subset of the candidate predictors from which a classifier can be constructed which is small but discriminatory. Effective feature selection shortens training time and reduces overfitting. There are various criteria that can be used when selecting candidate predictors (Kwak and Choi, 2002; Estévez et al., 2009; Hall, 1999; Peng et al., 2005). We use the correlation-based feature selection (CFS) (Hall, 1999) because it performed best in our preliminary experiments. The CFS is based on the hypothesis that a good feature subset contains features that are highly correlated with a class label, and largely uncorrelated with each other. Given a subset of features $S \subseteq \{X_1, X_2, \ldots, X_M\}$, merit of that subset is expressed as follows:

$$\mathrm{Merit}_S = \frac{\sum_{X_i \in S} \mathrm{U}(X_i, Y)}{\sqrt{\sum_{X_i \in S} \sum_{X_j \in S} \mathrm{U}(X_i, X_j)}}$$

The numerator measures the ability of S to predict the class label while the denominator measures the amount of information which is redundant between the selected features. To find the feature subset with the greatest merit, CFS uses the best-first search algorithm (Rich et al., 2009).

3.3. Multinomial logistic regression

Logistic regression is a statistical method that models the probability of a binary dependent variable. It assumes a linear relationship between the log odds of the dependent variable and the independent variables. Logistic regression is widely used in many atmospheric applications: Collino et al. (2009) assessed the ground effect of severe convective storms in Northern Italy; Gijben et al. (2017) predicted daily lightning threats over Southern Africa; Melcon et al. (2017) detected hail for Southwestern France; Moon et al. (2019) developed an early warning system for very short-term heavy rainfall.

Multinomial logistic regression is a generalization of logistic regression to problems involving more than two classes: in this case rain, snow, and sleet. In multinomial logistic regression with K classes, one class is chosen as a 'pivot' and $K - 1$ independent binary logistic regression models are constructed. If class K is selected as the pivot, then the model for class K is

$$\ln \frac{\Pr(Y=k)}{\Pr(Y=K)} = \mathbf{b}_k \cdot \mathbf{x} \tag{1}$$

where Y is the outcome random variable, $\mathbf{b}_k$ is the set of regression coefficients associated with class k, and $\mathbf{x}$ is a vector of observed weather variables. Then the probability that $\mathbf{x}$ belongs to class k can be expressed as follows:

$$\Pr(Y=k) = \Pr(Y=K) e^{\mathbf{b}_k \cdot \mathbf{x}} \tag{2}$$

Since the probabilities that $\mathbf{x}$ belongs to each class sum to 1, the probability that $\mathbf{x}$ belongs to class K becomes

$$\begin{aligned} \Pr(Y=K) &= 1 - \sum_{k=1}^{K-1} \Pr(Y=k) \\ &= 1 - \sum_{k=1}^{K-1} \Pr(Y=K) e^{\mathbf{b}_k \cdot \mathbf{x}} \\ &= 1 - \Pr(Y=K) \sum_{k=1}^{K-1} e^{\mathbf{b}_k \cdot \mathbf{x}} \end{aligned}$$

To eliminate the probability on the right-hand side, we can rewrite the above equation as follows:

$$\Pr(Y=K) = \frac{1}{1 + \sum_{k=1}^{K-1} e^{\mathbf{b}_k \cdot \mathbf{x}}} \tag{3}$$

and Eq. (2) as follows:

$$\Pr(Y=k) = \frac{e^{\mathbf{b}_k \cdot \mathbf{x}}}{1 + \sum_{k=1}^{K-1} e^{\mathbf{b}_k \cdot \mathbf{x}}} \tag{4}$$

Given observational data $\mathbf{x}$, multinomial logistic regression outputs a class label y such that:

$$y = \arg\max_k \Pr(Y=k)$$

The regression coefficients $\mathbf{b}_k$ are typically estimated by the maximum likelihood method (Hosmer et al., 2013). We use the ridge estimator (Cessie and Houwelingen, 1992) to prevent overfitting and instability. A discrete variable with p categorical values is converted to p binary (0 or 1) variables, each of which indicates whether or not the value of the variable falls into a certain category. Multinomial logistic regression usually uses all given features, however, there are recent studies (Ouyed and Allili, 2018a, 2018b) that embed feature weighting in multinomial logistic regression.

3.4. Performance criteria

Accuracy, expressed as the number of correct precipitation predictions divided by the total number of predictions, is the main performance criterion. In addition, we used Heidke skill score (HSS) to evaluate predictions of each type of precipitation individually. Table 2 shows a contingency table of prediction results for a binary event. Here, proportion correct (PC) for the event is:

Table 2
Contingency table for forecasts of a binary event. The numbers of occurrences in each category are denoted by a, b, c, and d.

Forecast	Observed		
	Yes	No	Total
Yes	a	b	$a + b$
No	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d = n$

$$\mathrm{PC} = \frac{a+d}{n}$$

The HSS is an adjusted PC that is scaled by the portion of correct forecasts due to random chance in the absence of forecast skill. The probability that the predictions will hit by chance is:

$$\mathrm{E} = \left(\frac{a+c}{n}\right)\left(\frac{a+b}{n}\right) + \left(\frac{b+d}{n}\right)\left(\frac{c+d}{n}\right)$$

which is the sum of the probabilities that a random forecast predicting *yes* is correct by chance and that a random forecast predicting *no* is correct by chance. Then, HSS is defined as:

$$\mathrm{HSS} = \frac{\mathrm{PC} - \mathrm{E}}{1 - \mathrm{E}}$$

Perfect forecast skill has an HSS value of 1 while no skill has a value of 0. The HSS was used by Behrangi et al. (2018) to compare various methods for determining precipitation phase.

We also used Peirce skill score (PSS) to evaluate the predictive performance for each precipitation type. The PSS can be defined by hit rate (HR) and false alarm rate (FAR). The HR is the proportion of events that were correctly predicted:

$$\mathrm{HR} = \frac{a}{a+c}$$

and FAR is the proportion of non-events that were incorrectly predicted:

$$\mathrm{FAR} = \frac{b}{b+d}$$

Then, PSS is defined as:

$$\mathrm{PSS} = \mathrm{HR} - \mathrm{FAR}$$

The PSS value of 1 indicates perfect forecast skill, while 0 indicates no skill. The PSS is useful in rare-event situations such as the occurrence of sleet since it is sensitive to the climatological frequency of the event (Gandin and Murphy, 1992). Please refer to Jolliffe and Stephenson (2003) and Wilks (2011) for general guidance on forecast verification including HSS and PSS.

4. Result

4.1. Experiment setup

We used the short-range weather forecast data from ECMWF and RDAPS for 22 significant locations in South Korea from 2013 to 2015. A 3-fold cross-validation was performed to evaluate different methods of forecasting precipitation types. The 2013 and 2014 data was used to train a model, which was then evaluated on the 2015 data. This procedure was repeated twice, using the 2013 and 2015 data for training and the 2014 data for evaluation, and then using the 2014 and 2015 data for training and the 2013 data for evaluation. The results from these procedures were averaged to produce a single estimation of performance for each forecasting model. The cross-validation was repeated for each 3-hour interval, from 3 to 72 h after the time of the forecast. Separate evaluations were performed for the ECMWF and RDAPS data.

Variable 69 (precipitation type) in the ECMWF and RDAPS short-range forecasts were used as a benchmark measure of performance. The results were also compared against those obtained using the improved Matsuo scheme (Lee et al., 2014), which is an algorithm developed by meteorologists to predict winter precipitation types in South Korea from air temperature, relative humidity, and the thickness of the 1000–850 layer. The performance of these two predictions were evaluated over all 3-hour intervals.

We tested the performance of multinomial logistic regression without any preprocessing methods, and that of multinomial logistic regression preceded by PCA (Behrangi et al., 2018; Moon et al., 2019),

Table 3
Frequently selected variables by CFS. Only those selected to participate in more than 50% of all models are shown.

No.	Variable name	Frequency[a]			U(*X*, *Y*)[b]
		ECMWF	RDAPS	Average	
64	Snow (kg/m^2)	100.00%	100.00%	100.00%	0.2495
92	Thickness of geopotential height at 1000–700 hPa (gpm)	100.00%	100.00%	100.00%	0.4540
91	Thickness of geopotential height at 1000–850 hPa (gpm)	97.22%	100.00%	98.61%	0.4568
14	Specific humidity at surface (kg/kg)	91.67%	93.06%	92.36%	0.4408
93	Thickness of geopotential height at 1000–500 hPa (gpm)	93.06%	90.28%	91.67%	0.3891
65	Equivalent potential temperature at 925 hPa (K)	88.89%	94.44%	91.67%	0.4445
6	Temperature at 850 hPa (K)	88.89%	88.89%	88.89%	0.3675
34	North-east wind at surface (m/s)	76.39%	77.78%	77.08%	0.3409
1	Latitude of the target location (°)	68.06%	76.39%	72.22%	0.0829
33	South wind at 500 hPa (m/s)	58.33%	80.56%	69.44%	0.3614
69	Precipitation type ({rain, sleet, snow})	45.83%	87.50%	66.67%	0.2495
13	Relative humidity at 500 hPa (%)	69.44%	48.61%	59.03%	0.2644
23	Dew point depression at 500 hPa (K)	38.89%	65.28%	52.08%	0.2640
16	Specific humidity at 850 hPa (kg/kg)	45.83%	56.94%	51.39%	0.3891

[a] Frequency with which variables in the ECMWF and RDAPS datasets were selected by CFS.
[b] U(X, Y) denotes the average symmetric uncertainty between the variable (X) and the precipitation type (Y).

which uses an orthogonal transformation to convert possibly correlated features into linearly uncorrelated features. The PCA projects high-dimensional features to a lower dimensional space by introducing a new coordinate system, and therefore reduces the number of features in a dataset. We also tested the C4.5 decision tree algorithm (Quinlan, 1993), which is based on the concept of entropy in information theory. It builds a decision tree using a learning technique in which it recursively chooses the feature at each node of the tree that best differentiates between instances of the training set. The improved Matsuo scheme used as a benchmark can also be seen as a handcrafted decision tree. The starting point of a decision tree is the root node, and the next node is determined by values of the selected feature. The tree traversal ends when it reaches a leaf node containing a class label.

We implemented the improved Matsuo scheme in C, and used the Waikato Environment for Knowledge Analysis (WEKA) package due to Hall et al. (2009) to implement PCA, CFS, C4.5, and multinomial logistic regression.

4.2. Analysis of feature selection

The CFS selects different features to build the model for each lead time. Table 3 lists the variables that were selected by CFS for at least 50% of the models, trained on either the ECMWF or the RDAPS data. Variables 64 (snow) and 92 (the thickness of the 1000-700 hPa layer) were selected for all models, and other variables related to humidity, temperature, and wind were frequently selected. Among variables relating to location, only latitude was selected for more than 60% of the models. When using the RDAPS dataset, CFS usually included Variable 69, which is the prediction of precipitation type embedded in the data, in its choices. However, Variable 69 was selected for fewer than half the models when using the ECMWF dataset. As shown later, this is understandable because the RDAPS prediction of precipitation is more accurate. On average, CFS chose 16 candidate predictors using the ECMWF dataset, and 18 when the RDAPS dataset was in use.

At each step of feature selection, CFS chooses a variable that is highly correlated with the precipitation types, yet uncorrelated with the already selected variables. The symmetric uncertainty in Table 3 suggests that the most discriminatory variables as a single predictor were Variables 91 (the thickness of the 1000-850 hPa layer), 92 (the thickness of the 1000-700 hPa layer), 65 (equivalent potential temperature at 925 hPa), and 14 (specific humidity at surface). Variables 64 (snow) and 1 (latitude) were selected frequently, despite their relatively low

Table 4
Multinomial logistic regression for rain. The coefficient, standard error, z-value, and p-value of each variable are shown. The p-values less than 0.05 are denoted with an asterisk.

No.	Variable name	Coefficient	Std. Error	z-Value	p-Value
1	Latitude of the target location (°)	−0.6426	0.1765	−3.6410	0.0003*
5	Temperature at 925 hPa (K)	0.0599	0.0441	1.3580	0.1746
6	Temperature at 850 hPa (K)	−0.7041	2.2168	−0.3180	0.7508
13	Relative humidity at 500 hPa (%)	0.4317	0.1443	2.9920	0.0028*
14	Specific humidity at surface (kg/kg)	30.3777	11.6734	2.6020	0.0093*
16	Specific humidity at 850 hPa (kg/kg)	79.7856	30.9575	2.5770	0.0100*
23	Dew point depression at 500 hPa (K)	0.1859	0.2255	0.8240	0.4098
24	East wind at surface (m/s)	−0.2069	0.3743	−0.5530	0.5805
33	South wind at 500 hPa (m/s)	−0.2260	0.5572	−0.4060	0.6850
34	North-east wind at surface (m/s)	−0.5023	0.2545	−1.9740	0.0484*
44	Wind speed at surface (m/s)	−0.0178	0.1619	−0.1100	0.9126
51	Dew point temperature at surface (K)	−0.0830	0.0231	−3.5930	0.0003*
64	Snow (kg/m^2)	−0.1904	0.4056	−0.4690	0.6388
65	Equivalent potential temperature at 925 hPa (K)	−28.7840	11.1677	−2.5770	0.0100*
66	Equivalent potential temperature at 850 hPa (K)	0.3439	0.1498	2.2950	0.0217*
87	Potential vorticity at 850 hPa (m^{2}s^{-1}Kkg^{-1})	−0.0057	0.0288	−0.1990	0.8422
91	1000–850 hPa thickness (gpm)	−0.2048	0.2013	−1.0170	0.3091
92	1000–700 hPa thickness (gpm)	0.1379	0.0505	2.7330	0.0063*
93	1000–500 hPa thickness (gpm)	0.0079	0.0267	0.2950	0.7684
N/A	Intercept[a]	7799.4942	3144.3072	2.4810	0.0131*

[a] Intercept from the regression coefficients associated with rain.

Table 5
Multinomial logistic regression for snow. The coefficient, standard error, *z*-value, and *p*-value of each variable are shown. The *p*-values less than 0.05 are denoted with an asterisk.

No.	Variable name	Coefficient	Std. Error	*z*-value	*p*-value
1	Latitude of the target location (°)	0.4235	0.1100	3.8510	0.0001*
5	Temperature at 925 hPa (K)	−0.0819	0.0326	−2.5100	0.0121*
6	Temperature at 850 hPa (K)	0.9596	1.5213	0.6310	0.5282
13	Relative humidity at 500 hPa (%)	0.0039	0.0964	0.0410	0.9676
14	Specific humidity at surface (kg/kg)	7.5060	13.9197	0.5390	0.5897
16	Specific humidity at 850 hPa (kg/kg)	18.0430	36.8381	0.4900	0.6243
23	Dew point depression at 500 hPa (K)	−0.3616	0.1632	−2.2160	0.0267*
24	East wind at surface (m/s)	−0.3189	0.2572	−1.2400	0.2150
33	South wind at 500 hPa (m/s)	0.1398	0.3543	0.3940	0.6932
34	North-east wind at surface (m/s)	−0.1617	0.1665	−0.9720	0.3312
44	Wind speed at surface (m/s)	−0.0167	0.1021	−0.1640	0.8697
51	Dew point temperature at surface (K)	−0.0002	0.0142	−0.0130	0.9894
64	Snow (kg/m^2)	1.2201	0.3320	3.6750	0.0002*
65	Equivalent potential temperature at 925 hPa (K)	−6.9486	13.3716	−0.5200	0.6033
66	Equivalent potential temperature at 850 hPa (K)	0.5828	0.1328	4.3900	0.0000*
87	Potential vorticity at 850 hPa ($m^2 s^{-1} K kg^{-1}$)	0.0133	0.0210	0.6350	0.5256
91	1000–850 hPa thickness (gpm)	−0.2221	0.1494	−1.4870	0.1371
92	1000–700 hPa thickness (gpm)	0.0107	0.0311	0.3450	0.7303
93	1000–500 hPa thickness (gpm)	0.0104	0.0174	0.6000	0.5487
N/A	Intercept[a]	1965.3755	3747.6967	0.5240	0.6000

[a] Intercept from the regression coefficients associated with snow.

symmetric uncertainty. The CFS prefers variables that can add new information to the already selected variables to avoid redundancy.

4.3. Analysis of multinomial logistic regression

Multinomial logistic regression explains the relationship between the possible output of the dependent variable and each independent variable. We examined the best model on the ECMWF data to see whether or not CFS and multinomial logistic regression generated a reasonable classifier. The CFS selected 19 variables from 93 variables, and multinomial logistic regression was applied to the selected ones.

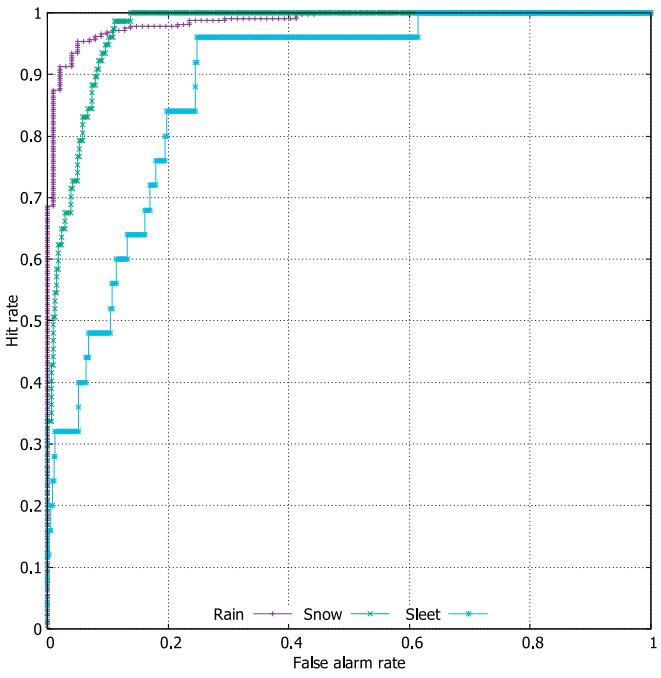

Fig. 2. ROC curves of the best model on the ECMWF data.

Table 4 shows the coefficients of the equation for rain, and Table 5 presents those for snow. The coefficient and its standard error of each variable are estimated by the maximum likelihood method. The *z*-value is the test statistic for the null hypothesis that the coefficient of the variable is 0, and it is calculated as the coefficient divided by its standard error. The *p*-value is computed from the *z*-value. A low *p*-value indicates that we can reject the null hypothesis that there is no relationship between the dependent variable and the corresponding independent variable. The equation for sleet is not provided since sleet was chosen as the pivot class.

When calculating the probability of the occurrence of rain or snow, 12 variables among the 19 selected ones were considered statistically significant at the significance level of 0.05. The coefficient of each variable also seems reasonable. For example, lower latitude and expected snowfall indicate a higher chance of rain, whereas higher latitude and expected snowfall indicate a higher chance of snow. Multiple logistic regression was able to generate effective equations that classify the types of precipitation.

A receiver operating characteristic (ROC) curve illustrates the ability of a binary classifier at various threshold settings. It is a plot of HR against FAR for the different possible thresholds. Since multinomial logistic regression predicts probabilities of each possible outcome, it can be interpreted as a binary classifier for each class label. Therefore, we can draw ROC curves for each precipitation type using the probabilities as decision thresholds. Fig. 2 compares the capabilities of the best model for each precipitation type. In general, a larger area under the ROC curve implies a better classification characteristic. This ROC curve shows that sleet is the most difficult to predict.

4.4. Comparative analysis

During non-winter seasons in South Korea, it is not difficult to achieve high accuracy in precipitation type predictions since precipitation occurs mostly in the form of rain. In winter, however, it is not easy to achieve

Table 6
Means and standard deviations of the accuracy of precipitation type predictions in ECMWF and RDAPS short-range forecasts, for all lead times.

	All seasons	Non-winter	Winter
ECMWF	0.8990 ± 0.0044	0.9701 ± 0.0033	0.6854 ± 0.0164
RDAPS	0.9056 ± 0.0050	0.9717 ± 0.0033	0.7077 ± 0.0112

high accuracy since various types of precipitation can occur. Table 6 shows the accuracy of precipitation type predictions included in the short-range forecasts of ECMWF and RDAPS between 2013 and 2015. The average accuracies for all seasons were about 90%, and over 97% in non-winter, but about 70% in winter. A significant improvement in overall accuracy can only be obtained by improving winter forecasts. Fig. 3 shows the accuracy of ECMWF and RDAPS predictions against lead time. The RDAPS forecasts are more accurate than those from ECMWF, in terms of precipitation type. As expected, accuracy decreases as the lead time increases. Wilcoxon signed-rank tests were performed on the accuracy pairs of ECMWF and RDAPS forecasts for all lead times to see if the accuracies of the forecasts from these two agencies have the same distribution. The tests indicated that RDAPS forecasts were more accurate than those from ECMWF, across all seasons ($p < 0.001$), and also in non-winter ($p < 0.01$) and winter ($p < 0.001$).

Experiments were conducted to improve the precipitation predictions of each forecast. We compared six different schemes for predicting precipitation type, based on a range of weather variables taken from the ECMWF forecast data, and compared the results with the ECMWF's own precipitation type forecast. Fig. 4 compares the accuracy of the methods and gives the HSS for each type of precipitation. All the schemes outperformed the ECMWF precipitation except that the improved Matsuo scheme was sometimes less accurate than ECMWF in sleet. The proposed method, multinomial logistic regression combined with CFS, has the highest performance, and multinomial logistic regression with PCA has the second highest performance for all measures except HSS for sleet. When we classify precipitation types, CFS which utilizes class labels were more effective than PCA which is a class-agnostic procedure. The predictive performance of the proposed method with the ECMWF data for each location individually is presented in Table C.1 in Appendix C.

Table 7 includes the results of Wilcoxon signed-rank tests for accuracy and HSS. The null hypothesis of the Wilcoxon signed-rank test is that the medians of the performance measure are the same for the proposed method and the method with which it is compared. A lower p-value indicates a greater difference between the performance of the two methods. These results show that the proposed method achieves a significant improvement in accuracy and HSS for rain and snow. C4.5 was the best at predicting sleet, followed by the proposed method. However, the p-value suggests that the difference in performance is not significant.

Table 8 compares the proposed method with the other schemes with respect to HR, FAR, and PSS for each precipitation type. The table gives the average of each measure for all lead times. The proposed method outperformed the other schemes in terms of PSS for rain and snow. For sleet, however, C4.5 and logistic regression performed better than the proposed method. Since PSS gives more rewards for a correct detection of rare events than HSS, schemes with a higher HR that offsets FAR achieved a higher PSS. Instead, the proposed method produced less false alarms than the two methods.

Fig. 5 shows the performance of the six schemes in predicting precipitation types from RDAPS data. Again, most schemes produce better predictions than RDAPS, except for those of the improved Matsuo scheme for sleet. This figure also shows that the overall performance of multinomial logistic regression combined with CFS or PCA is superior to that of the other schemes. Table 9 presents the results for RDAPS in

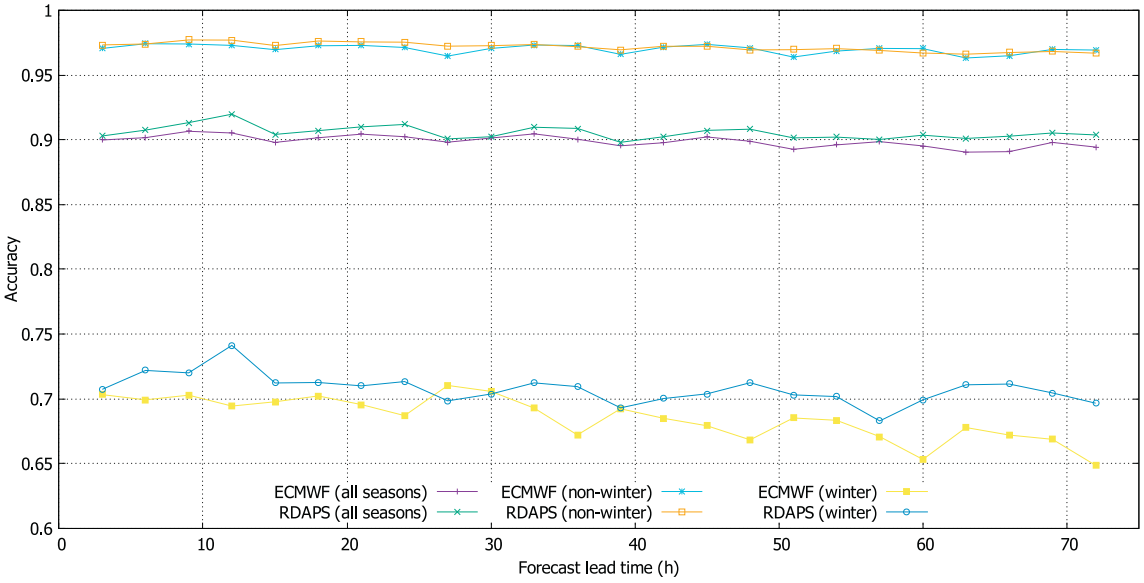

Fig. 3. Accuracies of ECMWF and RDAPS precipitation type predictions for different lead times.

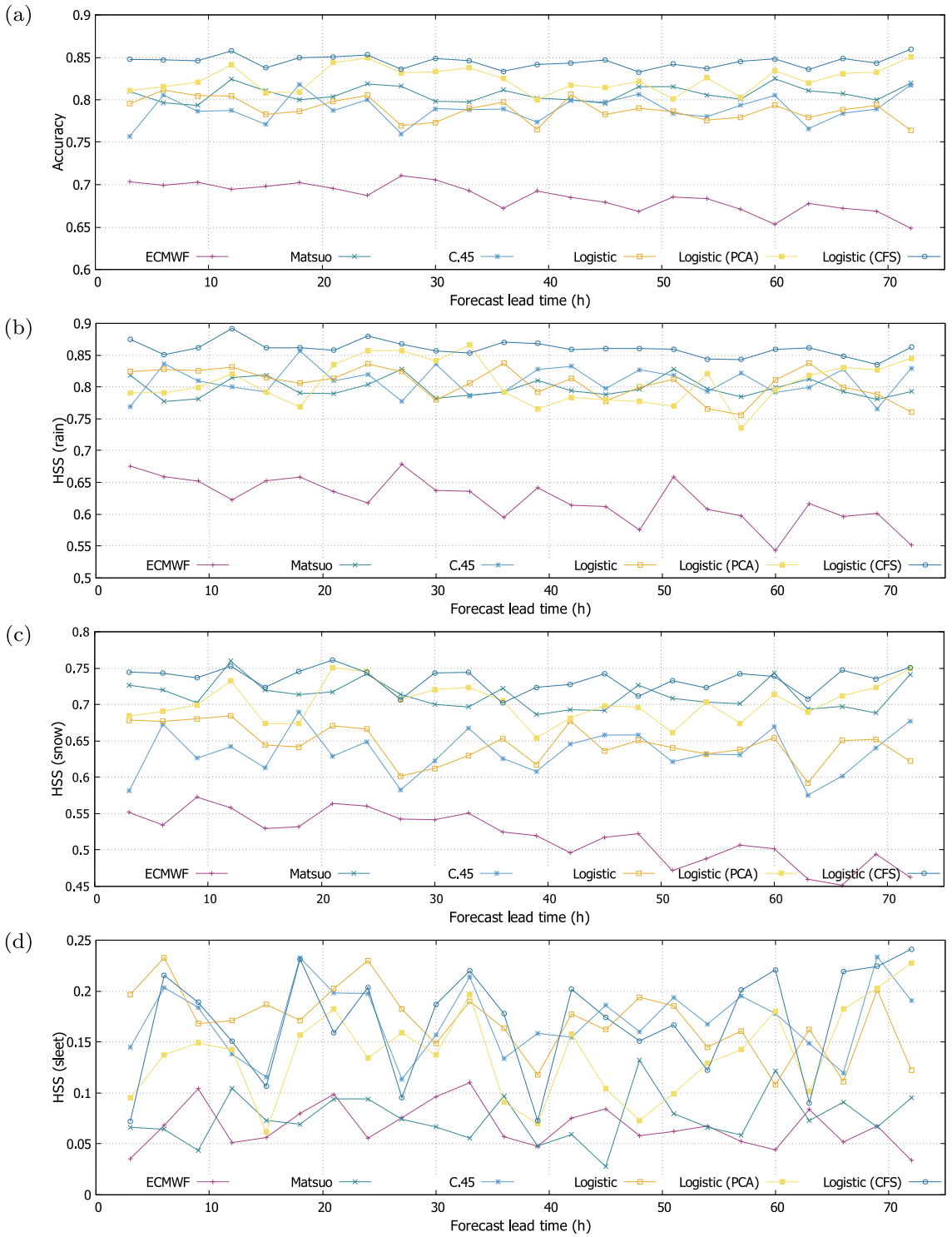

Fig. 4. Comparison of wintertime precipitation type predictions using ECMWF data: (a) accuracy, (b) HSS for rain, (c) HSS for snow, and (d) HSS for sleet. 'Logistic' denotes multinomial logistic regression, and is followed by the method used for preprocessing the data.

Table 7
Accuracy and HSS on the ECMWF dataset. The highest average values are shown in bold type. The *p*-values for each performance measure are calculated from the Wilcoxon signed-rank tests, with the null hypothesis that there is no difference between the proposed method and the method being compared with it.

	Accuracy		HSS (rain)		HSS (snow)		HSS (sleet)	
	Average	*p*-value	Average	*p*-value	Average	*p*-value	Average	*p*-value
Logistic (CFS)	**0.8449**		**0.8603**		**0.7346**		0.1705	
ECMWF	0.6854	< 0.001	0.6224	< 0.001	0.5189	< 0.001	0.0674	< 0.001
Matsuo	0.8073	< 0.001	0.7983	< 0.001	0.7127	< 0.001	0.0759	< 0.001
C4.5	0.7890	< 0.001	0.8090	< 0.001	0.6341	< 0.001	**0.1716**	0.8887
Logistic	0.7884	< 0.001	0.8058	< 0.001	0.6460	< 0.001	0.1705	0.8887
Logistic (PCA)	0.8241	< 0.001	0.8067	< 0.001	0.7027	< 0.001	0.1380	0.0016

The best values are shown in bold type.

Table 8
HR, FAR, and PSS on the ECMWF dataset. The best values are shown in bold type.

	Rain			Snow			Sleet		
	HR	FAR	PSS	HR	FAR	PSS	HR	FAR	PSS
Logistic (CFS)	**0.9458**	0.0865	**0.8593**	**0.8898**	0.1373	**0.7525**	0.1616	0.0311	0.1304
ECMWF	0.7783	0.1529	0.6254	0.6762	0.1624	0.5138	0.2416	0.1559	0.0858
Matsuo	0.8802	**0.0800**	0.8003	0.8892	0.1550	0.7343	0.1306	0.0641	0.0665
C4.5	0.9140	0.1054	0.8086	0.7543	0.1248	0.6296	0.2703	0.0922	0.1781
Logistic	0.9045	0.0983	0.8062	0.7636	**0.1216**	0.6420	**0.2812**	0.0987	**0.1825**
Logistic (PCA)	0.9295	0.1243	0.8051	0.8652	0.1468	0.7184	0.1291	**0.0276**	0.1014

The best values are shown in bold type.

the same way as Table 4. Again, the proposed method has the highest average score for all performance measures. The results of the Wilcoxon signed-rank test show that the proposed method outperforms the other methods with a significance level of 0.05, except for multinomial logistic regression with PCA. The proposed method has a higher average HSS than multinomial logistic regression with PCA but these results are not statistically significant for this dataset, although the corresponding accuracy result is statistically significant. Table 10 shows that the propose method is also superior to the other schemes on the RDAPS data in terms of PSS for rain and snow. For sleet, logistic regression has the highest PSS, but the proposed method has the lowest FAR. The predictive performance of the proposed method with the RDAPS data for each location individually is presented in Table C.2 in Appendix C.

We compared the performance of the proposed method with the ECMWF and RDAPS data, with the results shown in Table 11. Wilcoxon signed-rank tests were performed on all pairs of results, with the null hypothesis that the dataset makes no difference to performance. Sleet was better predicted from the RDAPS data than from the ECMWF data. Apart from sleet, the results from ECMWF data are better, but without statistical significance. The average values of HSS for sleet are much lower than the others. Most machine learning algorithms sacrifice performance on small classes to overall performance (Cardie and Howe, 1997; Chawla et al., 2002). Sleet only makes up 9.13% of the total number of precipitation instances, which could be expected to reduce the effectiveness with which sleet is forecast. Indeed, although there are 3510 observations of sleet in the data, the proposed method only predicts 1520 occurrences from the ECMWF data, and 1691 from the RDAPS data.

5. Conclusions and future work

It is not easy to predict precipitation types in winter due to the chaotic characteristics of winter weather. Most methods of predicting precipitation type use a small number of candidate predictors, mainly temperatures. We applied machine learning to the 93 meteorological variables of the ECMWF and RDAPS short-range weather forecasts for South Korea, and used CFS to choose a discriminatory subset of these variables. We believe that this is the first application of feature selection to precipitation type forecasting.

Multinomial logistic regression is applied to the selected features to predict winter precipitation types. The resulting prediction of precipitation found to be 15% more accurate than those within the ECMWF forecasts, and 13% better than those in the RDAPS forecasts. Our method also outperformed the improved Matsuo scheme (Lee et al., 2014), which is specialized for South Korean precipitation forecasts.

We used HSS and PSS to evaluate predictive performance for each type of precipitation. Sleet had the lowest scores, which can be attributed to its relative infrequency. In future work we plan to explore the use of undersampling (Liu et al., 2009), oversampling (Chawla et al., 2002), or boosting (Sun et al., 2007) techniques to improve the sleet forecast, while minimizing any reduction in the accuracy with which other precipitation types are predicted.

Declaration of Competing Interest

None.

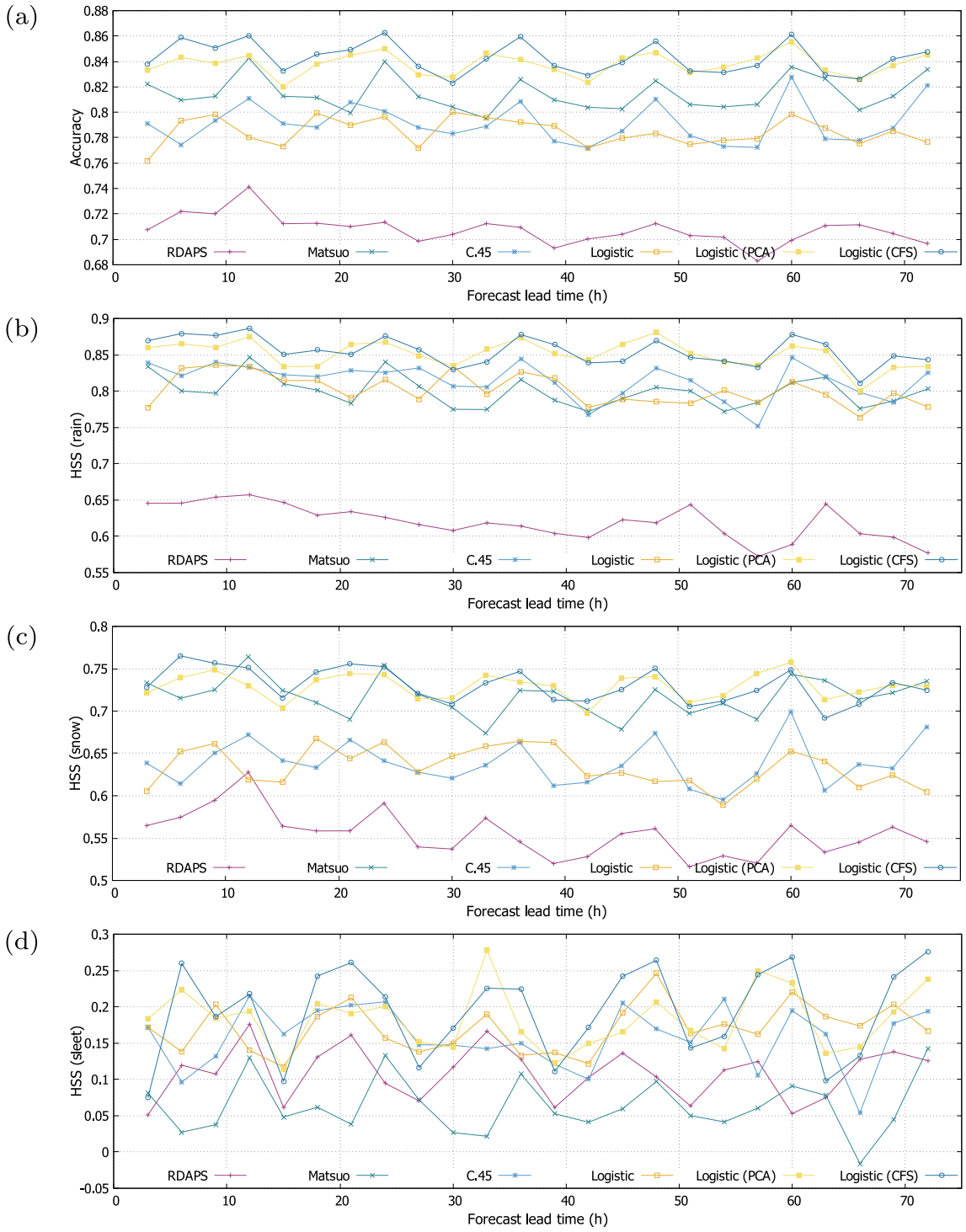

Fig. 5. Comparison of wintertime precipitation type predictions using RDAPS data: (a) accuracy, (b) HSS for rain, (c) HSS for snow, and (d) HSS for sleet. 'Logistic' denotes multinomial logistic regression, and is followed by the method used for preprocessing the data.

Table 9
Accuracy and HSS on the RDAPS dataset. The highest average values are shown in bold type. The *p*-values for each performance measure are calculated from the Wilcoxon signed-rank tests, with the null hypothesis that there is no difference between the proposed method and the method being compared with it.

	Accuracy		HSS (rain)		HSS (snow)		HSS (sleet)	
	Average	*p*-value	Average	*p*-value	Average	*p*-value	Average	*p*-value
Logistic (CFS)	**0.8428**		**0.8554**		**0.7303**		**0.1934**	
RDAPS	0.7077	< 0.001	0.6197	< 0.001	0.5547	< 0.001	0.1086	< 0.001
Matsuo	0.8149	< 0.001	0.7997	< 0.001	0.7173	0.0434	0.0636	< 0.001
C4.5	0.7913	< 0.001	0.8147	< 0.001	0.6386	< 0.001	0.1589	0.0244
Logistic	0.7847	< 0.001	0.8019	< 0.001	0.6340	< 0.001	0.1702	0.0466
Logistic (PCA)	0.8379	0.0041	0.8512	0.0989	0.7294	0.6965	0.1825	0.1236

The best values are shown in bold type.

Table 10
HR, FAR, and PSS on the RDAPS dataset. The best values are shown in bold type.

	Rain			Snow			Sleet		
	HR	FAR	PSS	HR	FAR	PSS	HR	FAR	PSS
Logistic (CFS)	**0.9415**	0.0870	**0.8545**	**0.8835**	0.1367	**0.7468**	0.1852	**0.0338**	0.1515
ECMWF	0.8111	0.1906	0.6205	0.6862	0.1401	0.5461	0.2546	0.1268	0.1279
Matsuo	0.9201	0.1216	0.7985	0.8616	0.1322	0.7294	0.1007	0.0496	0.0511
C4.5	0.9152	0.1007	0.8145	0.7643	0.1282	0.6361	0.2523	0.0897	0.1626
Logistic	0.9025	0.1003	0.8022	0.7569	**0.1266**	0.6303	**0.2807**	0.0984	**0.1824**
Logistic (PCA)	0.9356	**0.0850**	0.8506	0.8771	0.1329	0.7442	0.1917	0.0428	0.1489

The best values are shown in bold type.

Table 11
Performance of the proposed method on each dataset, showing the mean, standard deviation of the accuracy, and the *p*-value of the Wilcoxon signed-rank test for the corresponding performance measure.

	Dataset		*p*-value
	ECMWF	RDAPS	
Accuracy	0.8449 ± 0.0070	0.8428 ± 0.0122	0.2757
HSS (rain)	0.8603 ± 0.0121	0.8554 ± 0.0190	0.1188
HSS (snow)	0.7346 ± 0.0157	0.7303 ± 0.0200	0.2113
HSS (sleet)	0.1705 ± 0.0525	0.1934 ± 0.0634	0.0096

Acknowledgments

Many thanks to Dr. JunTae Choi at the National Institute of Meteorological Sciences (NIMS) of Korea for his valuable suggestions, which improved this paper. We also thank the anonymous reviewers for their helpful comments and suggestions. The present research has been conducted by the Research Grant of Kwangwoon University in 2020. This work was supported by Research on the Improved Precipitation Forecast of Numerical Weather Prediction Models (II), through NIMS of Korea, in 2016.

Appendix A. Meteorological variables

Table A.1 lists the meteorological variables used in this study. Both ECMWF and RDAPS forecasts use the same set of variables.

Table A.1
List of meteorological variables.

No.	Variable	No.	Variable
1	Latitude of the target location (°)	2	Altitude of the target location (°)
3	Elevation of the target location (m)	4	Temperature at surface (K)
5	Temperature at 925 hPa (K)	6	Temperature at 850 hPa (K)
7	Temperature at 700 hPa (k)	8	Temperature at 500 hPa (K)
9	Relative humidity at surface (%)	10	Relative humidity at 925 hPa (%)
11	Relative humidity at 850 hPa (%)	12	Relative humidity at 700 hPa (%)
13	Relative humidity at 500 hPa (%)	14	Specific humidity at surface (kg/kg)
15	Specific humidity at 925 hPa (kg/kg)	16	Specific humidity at 850 hPa (kg/kg)
17	Specific humidity at 700 hPa (kg/kg)	18	Specific humidity at 500 hPa (kg/kg)
19	Dew point depression at surface (K)	20	Dew point depression at 925 hPa (K)
21	Dew point depression at 850 hPa (K)	22	Dew point depression at 700 hPa (K)
23	Dew point depression at 500 hPa (K)	24	East wind at surface (m/s)
25	East wind at 925 hPa (m/s)	26	East wind at 850 hPa (m/s)

(*continued on next page*)

Table A.1 (*continued*)

No.	Variable	No.	Variable
27	East wind at 700 hPa (m/s)	28	East wind at 500 hPa (m/s)
29	South wind at surface (m/s)	30	South wind at 925 hPa (m/s)
31	South wind at 850 hPa (m/s)	32	South wind at 700 hPa (m/s)
33	South wind at 500 hPa (m/s)	34	North-east wind at surface (m/s)
35	North-east wind at 925 hPa (m/s)	36	North-east wind at 850 hPa (m/s)
37	North-east wind at 700 hPa (m/s)	38	North-east wind at 500 hPa (m/s)
39	North-west wind at surface (m/s)	40	North-west wind at 925 hPa (m/s)
41	North-west wind at 850 hPa (m/s)	42	North-west wind at 700 hPa (m/s)
43	North-west wind at 500 hPa (m/s)	44	Wind speed at surface (m/s)
45	Wind speed at 925 hPa (m/s)	46	Wind speed at 850 hPa (m/s)
47	Wind speed at 700 hPa (m/s)	48	Wind speed at 500 hPa (m/s)
49	Temperature max at surface (K)	50	Temperature min at surface (K)
51	Dew point temperature at surface (K)	52	Dew point temperature at 925 hPa (K)
53	Relative humidity at 300 hPa (%)	54	Dew point depression at 300 hPa (K)
55	Gust at surface (m/s)	56	Accumulated relative humidity at 925–500 hPa (%)
57	Accumulated relative humidity at 925–700 hPa (%)	58	Low cloud cover (%)
59	Total cloud cover (%)	60	Total column water vapor (kg/m^2)
61	Precipitable water at 500 hPa (kg/m^2)	62	Convective available potential energy (J/kg)
63	Precipitation (kg/m^2)	64	Snow (kg/m^2)
65	Equivalent potential temperature at 925 hPa (K)	66	Equivalent potential temperature at 850 hPa (K)
67	Equivalent potential temperature at 700 hPa (K)	68	Sky cover ({clear, scatter, broken, overcast})
69	Precipitation type ({rain, sleet, snow})	70	Depth of wet layer (DWL) at 1000–200 hPa (gpm)
71	Height of wet layer (HWL) at 1000–200 hPa (gpm)	72	Specific humidity of DWL at 1000–200 hPa (%)
73	Specific humidity of HWL at 1000–200 hPa (%)	74	Index for rainfall forecast at 1000–200 hPa
75	K-index	76	Lifted index
77	Showalter stability index	78	Lifted condensation level at 925 hPa (hPa)
79	Lifted condensation level at 850 hPa (hPa)	80	Lifted condensation level at 700 hPa (hPa)
81	Lapse rate at 850–500 hPa (°C/km)	82	Lapse rate at 850–700 hPa (°C/km)
83	Lapse rate at 925–850 hPa (°C/km)	84	Lapse rate at 950–850 hPa (°C/km)
85	Lapse rate at 950–925 hPa (°C/km)	86	Lapse rate at 1000–925 hPa (°C/km)
87	Potential vorticity at 850 hPa ($m^2s^{-1}Kkg^{-1}$)	88	Potential vorticity at 700 hPa ($m^2s^{-1}Kkg^{-1}$)
89	Potential vorticity at 500 hPa ($m^2s^{-1}Kkg^{-1}$)	90	Potential vorticity at 300 hPa ($m^2s^{-1}Kkg^{-1}$)
91	1000–850 hPa thickness (gpm)	92	1000–700 hPa thickness (gpm)
93	1000–500 hPa thickness (gpm)		

Table B.1
List of 22 significant locations in South Korea. The latitude, longitude, altitude, monthly temperature statistics, and the number of occurrences of each precipitation type are presented. 'Non-winter' signifies March to November, and 'winter' signifies December to February. The monthly temperatures shows the range of average monthly temperatures during each period. The number of occurrences of each precipitation type refers to precipitation observed in a 3-hour interval from 2013 to 2015.

No.	Name	Lat. (°)	Lon. (°)	Alt. (m)	Monthly temperatures (°C)		Occurrences in winter		
					Non-winter	Winter	Rain	Snow	Sleet
1	Chuncheon	37.9	127.7	77.7	[4.4, 26.7]	[−7.2, 0.4]	77	134	13
2	B aengnyeongdo	38.0	124.6	144.9	[1.8, 24.9]	[−2.2, 3.5]	64	195	29
3	Bukgangneung	37.8	128.9	78.9	[6.4, 27.2]	[−1.2, 4.4]	75	156	21
4	Seoul	37.6	127.0	85.8	[5.1, 27.7]	[−3.4, 1.9]	95	86	11
5	Incheon	37.5	126.6	71.4	[4.1, 26.9]	[−2.9, 2.4]	87	89	14
6	Ulleungdo	37.5	130.9	222.8	[6.2, 26.6]	[0.5, 5.4]	173	508	84
7	Suwon	37.3	127.0	34.1	[4.8, 27.4]	[−3.4, 2.4]	106	81	16
8	Seosan	36.8	126.5	28.9	[3.8, 26.6]	[−2.8, 2.8]	107	140	34
9	Cheongju	36.6	127.4	57.2	[6.4, 28.0]	[−2.7, 2.9]	114	123	17
10	Daejeon	36.4	127.4	68.9	[6.6, 27.8]	[−2.6, 3.1]	141	120	36
11	Andong	36.6	128.7	140.1	[4.5, 27.4]	[−3.2, 1.9]	96	66	4
12	Pohang	36.0	129.4	2.3	[8.8, 28.8]	[0.9, 6.3]	173	42	17
13	Daegu	35.9	128.7	49.0	[4.8, 29.0]	[−0.1, 4.5]	64	10	1
14	Jeonju	35.8	127.1	61.4	[7.0, 28.5]	[−1.3, 4.4]	161	78	25
15	Ulsan	35.6	129.3	34.6	[8.9, 29.0]	[0.9, 6.5]	170	35	17
16	Changwon	35.2	128.6	37.2	[8.9, 28.0]	[1.3, 5.5]	132	11	7
17	Gwangju	35.2	126.9	72.4	[7.6, 28.4]	[0.0, 5.3]	167	137	42
18	Busan	35.1	129.0	69.6	[9.5, 28.0]	[2.5, 7.9]	141	13	2
19	Mokpo	34.8	126.4	38.0	[6.1, 27.8]	[0.6, 6.3]	166	111	65
20	Yeosu	34.7	127.7	64.6	[8.3, 27.0]	[1.8, 7.0]	141	14	7
21	Heuksando	34.7	125.5	76.5	[6.0, 26.9]	[3.3, 7.2]	164	70	64
22	Jeju	33.5	126.5	20.5	[10.0, 29.1]	[5.6, 10.0]	371	53	84

Appendix B. The locations used in the study

Table B.1 lists the latitude, longitude, altitude, monthly temperature statistics, and winter precipitation data for the 22 significant South Korean locations.

Appendix C. Performance of the proposed method for each location

Tables C.1 and C.2 give the performance of the proposed method for each location, using the ECMWF and RDAPS dataset respectively. Locations with more frequent precipitation can be expected to have a larger influence on aggregate results.

Table C.1
Predictive performance of precipitation types for 22 significant locations in South Korea using the ECMWF dataset. 'PC' denotes the proportion correct, which is the proportion of correct forecasts.

No.	Name	Accuracy	Rain		Snow		Sleet	
			PC	HSS	PC	HSS	PC	HSS
1	Chuncheon	0.8631	0.8969	0.7735	0.8915	0.7708	0.9377	0.0006
2	Baengnyeongdo	0.8364	0.9369	0.8297	0.8450	0.6333	0.8909	0.0011
3	Bukgangneung	0.8175	0.8879	0.7331	0.8407	0.6510	0.9064	0.0014
4	Seoul	0.8807	0.9171	0.8343	0.9053	0.8107	0.9390	0.0005
5	Incheon	0.8862	0.9422	0.8842	0.9101	0.8201	0.9202	0.0368
6	Ulleungdo	0.8473	0.9266	0.8086	0.8858	0.7381	0.8822	0.0268
7	Suwon	0.8495	0.9071	0.8136	0.8822	0.7605	0.9097	0.0001
8	Seosan	0.8080	0.9113	0.8177	0.8352	0.6716	0.8695	0.0724
9	Cheongju	0.8505	0.9201	0.8391	0.8717	0.7440	0.9092	0.0007
10	Daejeon	0.8241	0.9309	0.8619	0.8442	0.6865	0.8731	0.0109
11	Andong	0.8880	0.9108	0.8128	0.9015	0.7893	0.9637	0.0006
12	Pohang	0.8113	0.8749	0.6765	0.8644	0.5460	0.8833	0.1950
13	Daegu	0.9505	0.9646	0.7858	0.9505	0.6687	0.9858	0.2447
14	Jeonju	0.8528	0.9366	0.8662	0.8698	0.7101	0.8993	0.0015
15	Ulsan	0.8533	0.9310	0.7945	0.8947	0.5735	0.8809	0.0733
16	Changwon	0.9135	0.9476	0.7070	0.9338	0.5016	0.9455	0.0009
17	Gwangju	0.8423	0.9623	0.9245	0.8601	0.7120	0.8621	0.0539
18	Busan	0.9421	0.9746	0.8365	0.9492	0.5454	0.9604	0.2208
19	Mokpo	0.7630	0.9523	0.9046	0.7834	0.5074	0.7904	0.2274
20	Yeosu	0.9452	0.9831	0.9204	0.9562	0.7260	0.9512	0.2451
21	Heuksando	0.7896	0.9780	0.9552	0.8070	0.4814	0.7942	0.3355
22	Jeju	0.8311	0.9165	0.7797	0.8930	0.4973	0.8526	0.3639

Table C.2
Predictive performance of precipitation types for 22 significant locations in South Korea using the RDAPS dataset. 'PC' denotes the proportion correct, which is the proportion of correct forecasts.

No.	Name	Accuracy	Rain		Snow		Sleet	
			PC	HSS	PC	HSS	PC	HSS
1	Chuncheon	0.8746	0.9154	0.8155	0.8992	0.7884	0.9346	0.0072
2	Baengnyeongdo	0.8278	0.9350	0.8224	0.8339	0.6069	0.8866	0.0087
3	Bukgangneung	0.8155	0.8806	0.7130	0.8414	0.6508	0.9091	0.0151
4	Seoul	0.8725	0.9126	0.8252	0.8980	0.7958	0.9344	0.0011
5	Incheon	0.8817	0.9422	0.8840	0.8963	0.7931	0.9248	0.1008
6	Ulleungdo	0.8431	0.9200	0.7940	0.8840	0.7356	0.8822	0.0212
7	Suwon	0.8538	0.9209	0.8411	0.8788	0.7529	0.9080	0.0020
8	Seosan	0.8150	0.9252	0.8454	0.8327	0.6669	0.8720	0.0966
9	Cheongju	0.8587	0.9270	0.8531	0.8730	0.7467	0.9174	0.0065
10	Daejeon	0.8093	0.9103	0.8205	0.8353	0.6686	0.8731	0.0178
11	Andong	0.8703	0.8942	0.7778	0.8838	0.7512	0.9627	0.0006
12	Pohang	0.8148	0.8728	0.6849	0.8623	0.5823	0.8945	0.2066
13	Daegu	0.9410	0.9552	0.7711	0.9552	0.7414	0.9717	0.0004
14	Jeonju	0.8509	0.9287	0.8496	0.8620	0.6930	0.9111	0.0991
15	Ulsan	0.8613	0.9237	0.7817	0.9041	0.6471	0.8947	0.1169
16	Changwon	0.9177	0.9487	0.7169	0.9359	0.5369	0.9509	0.0308
17	Gwangju	0.8304	0.9539	0.9076	0.8497	0.6891	0.8571	0.0494
18	Busan	0.9502	0.9736	0.8378	0.9604	0.6862	0.9665	0.2188
19	Mokpo	0.7645	0.9558	0.9116	0.7869	0.5097	0.7864	0.2440
20	Yeosu	0.9422	0.9801	0.9086	0.9522	0.7292	0.9522	0.0943
21	Heuksando	0.7745	0.9710	0.9413	0.7936	0.4589	0.7843	0.2968
22	Jeju	0.8314	0.9149	0.7704	0.9063	0.3879	0.8417	0.4179

The dataset being used seems to have no significant effect on the accuracy of the prediction for any location. For example, predictions for Mokpo and Heuksando were less than 80% accurate, but those for Daegu, Changwon, Busan, and Yeosu were more than 90% accurate, irrespective of the dataset. Although the PCs for sleet were generally higher than those for snow, the HSSs for sleet were generally much lower: the sleet forecasts are better than random predictions only at some locations. As already discussed, the accurate forecasting of sleet is difficult because it occurs relatively infrequently. For example, sleet occurred only once in Daegu on our dataset, and thus HSS for sleet can vary greatly depending on the success of the prediction for the event. In the case of rain and snow, however, the proposed method has good forecast skill since PCs and HSSs had high values in all the locations.

References

Behrangi, A., Yin, X., Rajagopal, S., Stampoulis, D., Ye, H., 2018. On distinguishing snowfall from rainfall using near-surface atmospheric information: comparative analysis, uncertainties and hydrologic importance. Q. J. Roy. Meteor. Soc. 144 (S1), 89–102.

Box, J.E., Fettweis, X., Stroeve, J.C., Tedesco, M., Hall, D.K., Steffen, K., 2012. Greenland ice sheet albedo feedback: thermodynamics and atmospheric drivers. Cryosphere 6 (4), 821–839.

Cardie, C., Howe, N., 1997. Improving minority class prediction using case-specific feature weights. In: Proceedings of the 14th International Conference on Machine Learning (ICML), pp. 57–65.

Cessie, S.L., Houwelingen, J.C.V., 1992. Ridge estimators in logistic regression. J. Roy. Stat. Soc. C Appl. Stat. 41 (1), 191–201.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

Chen, R.S., Liu, J.F., Song, Y.X., 2014. Precipitation type estimation and validation in China. J. Mt. Sci. 11 (4), 917–925.

Collino, E., Bonelli, P., Gilli, L., 2009. ST-AR (STorm-ARchive): a project developed to assess the ground effects of severe convective storms in the Po Valley. Atmos. Res. 93 (1), 483–489.

Cover, T.M., Thomas, J.A., 2006. Elements of Information Theory. Wiley-Interscience.

Dai, A., 2008. Temperature and pressure dependence of the rain-snow phase transition over land and ocean. Geophys. Res. Lett. 35 (12).

Ding, B., Yang, K., Qin, J., Wang, L., Chen, Y., He, X., 2014. The dependence of precipitation types on surface elevation and meteorological conditions and its parameterization. J. Hydrol. 513, 154–163.

Estévez, P.A., Tesmer, M., Perez, C.A., Zurada, J.M., 2009. Normalized mutual information feature selection. IEEE Trans. Neural Netw. 20 (2), 189–201.

Froidurot, S., Zin, I., Hingray, B., Gautheron, A., 2014. Sensitivity of precipitation phase over the Swiss Alps to different meteorological variables. J. Hydrometeorol. 15 (2), 685–696.

Gandin, L.S., Murphy, A.H., 1992. Equitable skill scores for categorical forecasts. Mon. Weather Rev. 120 (2), 361–370.

Gao, X., Ye, B., Zhang, S., Qiao, C., Zhang, X., 2010. Glacier runoff variation and its influence on river runoff during 1961–2006 in the Tarim river basin, China. Sci. China Earth Sci. 53 (6), 880–891.

Gijben, M., Dyson, L.L., Loots, M.T., 2017. A statistical scheme to forecast the daily lightning threat over Southern Africa using the Unified Model. Atmos. Res. 194, 78–88.

Hall, M.A., 1999. Correlation-Based Feature Selection for Machine Learning. Ph.D. Thesis. University of Waikato.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. SIGKDD Explor. 11 (1), 10–18.

Hosmer, D.W., Lemeshow, S., Sturdivant, R.X., 2013. Applied Logistic Regression. Wiley-Interscience.

Hynčica, M., Huth, R., 2019. Long-term changes in precipitation phase in Europe in cold half year. Atmos. Res. 227, 79–88.

Jennings, K.S., Winchell, T.S., Livneh, B., Molotch, N.P., 2018. Spatial variation of the rain-snow temperature threshold across the Northern Hemisphere. Nat. Commun. 9 (1), 1148.

Jolliffe, I.T., Stephenson, D.B., 2003. Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wiley.

Keeter, K.K., Cline, J.W., 1991. The objective use of observed and forecast thickness values to predict precipitation type in North Carolina. Weather Forecast. 6 (4), 456–469.

Kienzle, S.W., 2008. A new temperature based method to separate rain and snow. Hydrol. Process. 22 (26), 5067–5085.

Kwak, N., Choi, C.H., 2002. Input feature selection for classification problems. IEEE Trans. Neural Netw. 13 (1), 143–159.

Lee, S.M., Han, S.U., Won, H.Y., Ha, J.C., Lee, Y.H., Lee, J.H., Park, J.C., 2014. A method for the discrimination of precipitation type using thickness and improved Matsuo's scheme over South Korea. Atmosphere 24, 151–158.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., Bergström, S., 1997. Development and test of the distributed HBV-96 hydrological model. J. Hydrol. 201 (1), 272–288.

Liu, X.Y., Wu, J., Zhou, Z.H., 2009. Exploratory undersampling for class-imbalance learning. IEEE Trans. Syst. Man Cybern. Part B Cybern. 39 (2), 539–550.

Liu, S., Yan, D., Qin, T., Weng, B., Lu, Y., Dong, G., Gong, B., 2018. Precipitation phase separation schemes in the Naqu river basin, eastern Tibetan plateau. Theor. Appl. Climatol. 131 (1), 399–411.

Melcon, P., Merino, A., Sanchez, J.L., Lopez, L., Garcia-Ortega, E., 2017. Spatial patterns of thermodynamic conditions of hailstorms in Southwestern France. Atmos. Res. 189, 111–126.

Moon, S.H., Kim, Y.H., Lee, Y.H., Moon, B.R., 2019. Application of machine learning to an early warning system for very short-term heavy rainfall. J. Hydrol. 568, 1042–1054.

Norrman, J., Eriksson, M., Lindqvist, S., 2000. Relationships between road slipperiness, traffic accident risk and winter road maintenance activity. Clim. Res. 15 (3), 185–193.

Ouyed, O., Allili, M.S., 2018a. Feature weighting for multinomial kernel logistic regression and application to action recognition. Neurocomputing 275, 1752–1768.

Ouyed, O., Allili, M.S., 2018b. Recognizing human interactions using group feature relevance in multinomial kernel logistic regression. In: Proceedings of the 31st International Florida Artificial Intelligence Research Society Conference, pp. 541–546.

Peng, H., Long, F., Ding, C., 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. 27 (8), 1226–1238.

Quinlan, R., 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

Ralph, F.M., Rauber, R.M., Jewett, B.F., Kingsmill, D.E., Pisano, P., Pugner, P., Rasmussen, R.M., Reynolds, D.W., Schlatter, T.W., Stewart, R.E., Tracton, S., Waldstreicher, J.S., 2005. Improving short-term (0-48 h) cool-season quantitative precipitation forecasting: recommendations from a USWRP workshop. B. Am. Meteorol. Soc. 86 (11), 1619–1632.

Reeves, H.D., Ryzhkov, A.V., Krause, J., 2016. Discrimination between winter precipitation types based on spectral-bin microphysical modeling. J. Appl. Meteorol. Climatol. 55 (8), 1747–1761.

Rich, E., Knight, K., Shivashankar, N.B., 2009. Artificial Intelligence. McGraw-Hill.

Sims, E.M., Liu, G., 2015. A parameterization of the probability of snow-rain transition. J. Hydrometeorol. 16 (4), 1466–1477.

Sun, Y., Kamel, M.S., Wong, A.K., Wang, Y., 2007. Cost-sensitive boosting for classification of imbalanced data. Pattern Recogn. 40 (12), 3358–3378.

Wigmosta, M.S., Vail, L.W., Lettenmaier, D.P., 1994. A distributed hydrology-vegetation model for complex terrain. Water Resour. Res. 30 (6), 1665–1679.

Wilks, D.S., 2011. Statistical Methods in the Atmospheric Sciences. Academic Press.

Yang, Z.L., Dickinson, R.E., Robock, A., Vinnikov, K.Y., 1997. Validation of the snow submodel of the biosphere-atmosphere transfer scheme with Russian snow cover and meteorological observational data. J. Clim. 10 (2), 353–373.

Zhong, K., Zheng, F., Xu, X., Qin, C., 2018. Discriminating the precipitation phase based on different temperature thresholds in the Songhua river basin, China. Atmos. Res. 205, 48–59.