Ecole Polytechnique Fédérale de Lausanne

Master Thesis

Coastal atmospheric temperature prediction in Greenland using support vector regression

---

Author:
Matthew Parkan

Supervisors:
Dr. Devis Tuia
Prof. François Golay

Laboratory of Geographic Information Systems (LASIG)

June 29, 2012

Abstract

In the recent years, global climate change has induced evergrowing loss of sea ice in the Arctic. As the sea ice disappears, albedo diminishes and the sea surface is more likely to be warmed by incoming solar radiation. With the right wind conditions, this extra heat may also be advected towards the shore and thus influence coastal atmospheric temperatures. Thus, knowing how coastal atmospheric temperature is related to offshore conditions is paramount to help predict inshore effects. To study this relation, an exploratory approach using machine learning algorithms is proposed. Based on a combination of daily in situ (i.e. wind velocity, sea level pressure) and remotely sensed (i.e. sea surface temperature, sea ice concentration) data, a series of predicting features are constructed for the years 1981-2010. Two implementations of support vector regression (SVR), one with a linear kernel and the other with a combination of gaussian and histogram intersection kernels are then applied. Results of the SVR indicate that prediction root mean squared errors of less than $5°C$ are routinely achievable. Prediction errors are also found to be the smallest in summer months and/or at lower latitudes. Finally, the relative importance (ranking) of features appears to be highly variable, depending both on the location and the period of the year.

Keywords: Greenland, Atmospheric temperature prediction, Support vector regression

Résumé

Suite au changement climatique récent, la banquise arctique voit sa surface diminuer d'année en année. La disparition de la glace provoque une réduction de l'albédo, augmentant ainsi l'exposition de la surface de l'océan au rayonnement solaire. Sous certaines conditions, le vent peut permettre l'advection de la chaleur absorbée vers le littoral et donc influencer la température de l'air côtier. Il est ainsi essentiel de comprendre la relation entre les conditions au large et la température côtière pour en prédire les éventuels effets à l'intérieur des terres. Dans le but d'étudier cette relation, une approche exploratoire employant des algorithmes d'apprentissage automatique est proposée. Une combinaison de données journalières in situ (i.e. velocité du vent, pression atmosphérique) et issues de la télédétection (température à la surface de l'océan, concentration de glace) sont utilisées dans la construction de prédicteurs pour la période 1981-2010. Deux implémentations de la régression à vecteurs de support, l'une avec une fonction kernel linéaire et l'autre avec une combinaison de fonctions kernels gaussiennes et histogramme intersecté, sont appliquées. Les résultats indiquent qu'une prédiction avec une erreur moyenne en dessous de $5°C$ est régulièrement réalisable. Il est aussi déterminé que les erreurs de prédiction sont systématiquement plus faibles lors des mois d'été et/ou aux latitudes plus basses. Finalement, l'importance relative (classement) des prédicteurs semble être très variable en fonction du lieu et de la période de l'année.

Mots-clés: Groenland, Prédiction de température atmosphérique, Régression à vecteurs de support

Acknowledgements

I would like to thank Dr. Devis Tuia for introducing me to the field of machine learning, as well as for his patient and kind supervision. My gratitude is also extended to Dr. Loris Foresti whose advice has been a great help in converging on the subject of temperature prediction.
Special thanks also go to Prof. Michael Lehning for an enlightening discussion on Greenland and to Giona Matasci for his technical suggestions on support vector machines.
Finally, I wish to thank Prof. Marco Tedesco for pointing me to the article that spawned this project.

Contents

1 Introduction

1.1 Rationale

The main motivation for attempting to apply machine learning algorithms to atmospheric temperature prediction in Greenland came after reading an article by Rennermalm et al. 2009 about the possible correlation between sea ice concentration and the melt of the Greenland ice sheet. In their writings, Rennermalm et al. suggest that:

> *In principle, sea ice can be linked to ice sheet surface-melt through a chain of high correlations between sea ice and ocean temperatures (Comiso 2002), ocean and coastal temperatures (Hanna, Cappelen 2003a), and coastal temperatures and ice sheet surface melt(Abdalati, Steffen 2001,Mote 2007).*

To explain this chain of correlations, the authors suggest the following hypothesis:

> *[...] reduced offshore sea ice concentration, i.e. greater open-water fraction, warms the ocean mixed layer and increases onshore advection of sensible and turbulent heat fluxes, in turn raising air temperatures over the ice sheet and the probability of surface-melt occuring.*

It is important to understand that only correlations are suggested, not causes. Attributing the cause of increased ice sheet melt to decreasing sea ice is not straightforward. This effect which is known as polar amplification (that is the higher sensitivity of Polar regions to warming due to their temperature dependant albedo) is complex. In fact, determining the causality of melt involves the computation of energy and mass balances in turn dependant on multiple factors (such as precipitation). For these reasons, it was decided to restrict this study to a less complex set by focusing on the prediction of atmospheric temperatures, the penultimate link in the chain of correlations described by Rennermalm et al. Following the line of thought suggested by these authors, it is theoretically possible to estimate the atmospheric temperature at a particular location, by using local and regional offshore geophysical variables as predictors. Specifically, sea ice concentration, sea surface temperature, wind speed and direction are prime candidates (all linked to heat advection) to be considered for this purpose.

One of the main characteristics of atmospheric temperature is its non-linearity in time. Indeed, in addition to daily, seasonal and long term trends, air temperature at a particular location may exhibit very diverse behaviors depending on local and regional features such as dominant wind velocity, proximity to the ocean, topography, and so on. This non-linearity means that predictions with physical models become highly complex and sometimes unpractical (both in terms of computational costs and number of equations/parameters entering into account). This is when the use of statistical modelling can be advantageous. The statistical approach can be used either to restrict the number of parameters considered for physical models or as a standalone prediction method. Here, it is the latter that will catch our attention. Specifically, the standalone use of a statistical modelling approach known as machine learning. Guyon, Elisseeff 2006 define machine learning for predictive modelling as follows:

> *Machine learning problems occur when a task is defined by a series of cases or examples rather than by predefined rules. Given a number of training examples (also called data points, samples, patterns or observations) associated with desired outcomes, the machine learning process consists of finding the relationship between the patterns and the outcomes using solely the training examples.*

One of the main benefits of this approach is that unlike the use of physical models, no a priori knowledge of the relative importance (weight) of involved features is required. That is, the prediction is based only on the input data and the machine learning algorithm ranks the features according to the amount of information they bring. Thus, the model emerges from the data. A machine learning algorithm called support vector regression (SVR) is employed. Essentially, given a set of predictor values ("features" in machine learning

terminology) and their related label (i.e. atmospheric temperature), SVR will attempt to find a suitable model to explain the relationship between the former and latter.

After processing all the datasets into features, conducting a preliminary analysis and applying suppport vector regression, several questions are addressed in this study:

- Using a statistical approach, can a link be found between offshore geophysical variables and coastal atmospheric temperature?
- Using combined in situ and remotely sensed based features, what is the atomospheric temperature prediction accuracy that can be achieved using support vector regression?
- How does this prediction accuracy vary in space and time?
- How does the related feature ranking vary in space and time?
- Is there any time lag between specific features and temperature?

As a first step before investigating these questions, a short introduction to the coastal weather conditions of Greenland is presented. This is followed by a review of similar temperature prediction studies (using support vector regression). All the datasets are then described, as well as the processing applied to them. Afterwards, a preliminary exploratory data analysis is conducted, followed by the results of the SVR predictions and their interpretation in relation to the aforementioned questions. Finally, a list of further investigation possiblities is proposed.

1.2 Coastal weather conditions in Greenland

This section is intended to give an overview of the coastal conditions in Greenland. It is mainly concerned with coastal wind regimes and ocean currents. These topics where chosen among many others because they will provide a basic knowledge of important phenomenon occuring at the interface between ocean and land. They are in no way an exhaustive coverage. Much of the processes that occur in the Arctic are complex interactions, some of which are not yet fully understood and explaining them goes beyond the scope of this section.

Coastal wind regime

Coastal regions are in close contact with open water areas (at least during warm months) and can be directly affected by offshore heat advection. This advection is possible when the wind is blowing from the sea. Of course, this is not always the case and land winds occur frequently. In particular, katabatic winds may exert an influence on atmospheric temperature:

> *In its most generic sense, a katabatic wind (from the Greek word katabikos - to go down) refers to any downslope wind flowing from high elevation mountains, plateaus or hills down to valley or plains. But more commonly, the term is reserved for winds that, despite the effects of adiabatic compression during descent, are colder than the air displaced at the bottom of the incline. [...] The strongest katabatic winds are observed in western Greenland north of 70° N and in eastern Greenland north of 75° N. Katabatic storms are well known along the southeastern coast of Greenland and in the east coast valleys near Angmassalik. By disrupting the surface based temperature inversion, katabatic wind events can cause rapid changes in surface temperature.* (Serreze, Barry 2005)

Another important characteristic of Greenland is the presence of a barrier effect. That is the influence of the orography on the direction and speed of winds. The following lines are taken from the daily expedition journal[1] of Dallas Murphy (Woods Hole Oceanographic Institution) which illustrates the effect:

> *"This is interesting," said Ben, one of the atmospheric scientists. "It's blowing 43 knots at sea level along the Greenland barrier. But look, above 3,000 meters [the elevation of Greenland] the wind is blowing 23 knots and from a different direction." There it was, in black and white, so to say: conclusive evidence of the barrier effect.*

The orography may also cause regional wind accelerations along the coast (discussed further, see 3.1.2) known as tip jets.

[1] Cf. `http://www.whoi.edu/page.do?pid=29175`

Coastal ocean currents

The three main ocean currents along the coast of Greenland are the East Greenland current, the Irminger current and the West Greenland current (see fig. 1). The East Greenland current is cold with a low salinity and flows from the North along the East coast. It connects the Arctic to the Northern Atlantic, linking water masses from polar and temperate regions. It is composed of three water masses: Polar water (top 150 m, cold and low salinity), Atlantic water (from 150m to about 1000 m depth, relatively warm and salty) and deep water (from 1000 m to the bottom, constantly cold and salty). This current is known for the large quantity of sea ice it carries into the Atlantic Ocean.

Along the Southeast coast, the East Greenland current meets the relatively warm and salty Irminger current which branches off the North Atlantic Current (derived from the Gulf Stream). The two currents flow in parallel towards the South down to Cape Farewell (the Southern tip of Greenland). After that, the Irminger water mass dives below the Polar water, with some mixing in the process. The balance between the the two currents determines the hydrographical conditions of the West Greenland Current. Thus, changes in Atlantic circulation may influence local sea surface temperatures[2]. A detailed description of ocean currents is provided in Appendix C.

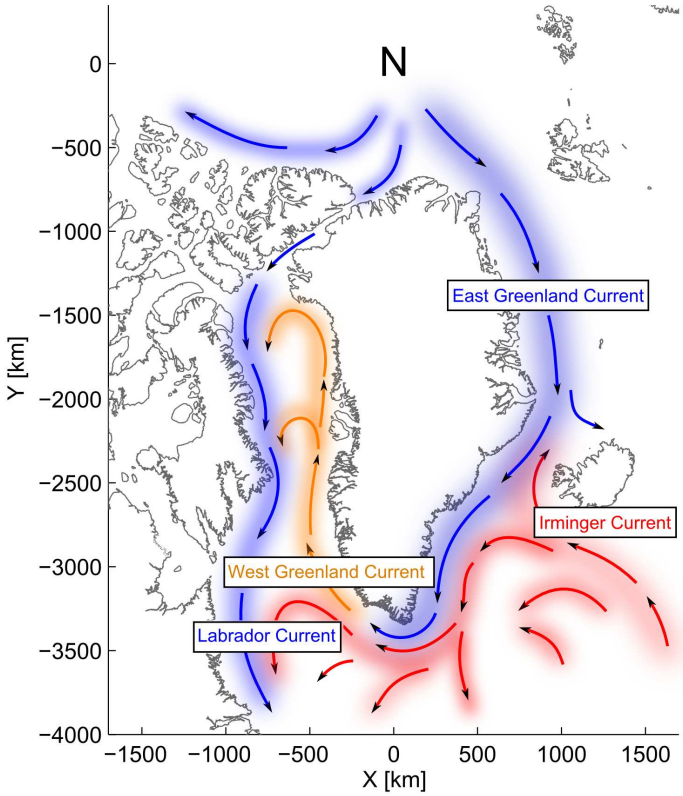

Figure 1: The main ocean surface currents along the coast of Greenland

[2]Cf. Greenland Institute of Natural Resources, *West Greenland Current temperatures remain high*, 8 Aug 2011, [http://www.natur.gl/en/fish-and-shellfish/sea-temperatures/]

1.3 Similar studies

Studies on non-spatialized atmospheric temperature prediction using SVR have usually focused on time series forecasting based on previous temperature observations. Most of the research in this field is recent, it includes:

- Radhika, Shashi 2009 who used SVR to predict the maximum temperature of the next day at a specific location based on the daily maximum temperatures for a span of previous n days. The MSE obtained on their predictions ranged from $7.07°C^2$ to $7.56°C^2$.
- Xue et al. 2009 who used genetic algorithms to optimize the free parameters of SVR which was than applied to meteorological prediction. Using classical SVR (without the addition of genetic algorithms to find the free parameters) on a series of 11 days in August for three different locations they obtained normalized RMSE ranging from 4.11% to 9.51%. When using genetic algorithms to find the free parameters, they obtained slightly better RMSE ranging from 3.01% to 8.10%.
- In Chevalier 2008, SVR models were created for air temperature prediction from one to twelve hours ahead. These models were more accurate than Artificial Neural Network models. Chevalier found Mean absolute errors ranging from $0.516°C$ (for the 1 hour horizon length) and $1.906°C$ for (for the 12 hour horizon length).
- Paniagua-Tineo et al. 2011 who focused on different measuring stations in Europe, from which different meteorological variables were obtained, including temperature, precipitation, relative humidity and air pressure. Two more variables were also included, specifically synoptic situation of the day and monthly cycle. Using this pool of prediction variables, it was shown that the SVR algorithm is able to give an accurate prediction of the maximum temperature 24h later. These authors were able to obtain a mean RMSE ranging from $1.5483°C$ and $2.8318°C$ depending on the location.
- Ortiz-Garcia et al. 2012 who presented a novel system for addressing problems of local very short term (up to a time prediction horizon of 6 h) temperature prediction based on Support Vector Regression algorithms. Using different models, Ortiz-Garcia et al. obtained mean RMSE in the $0.6°C$ to $1.38°C$ range for their predictions.

Given the good results obtained by these authors, SVR seems to be well adapted for atmospheric temperature prediction. Nonetheless, these studies are based directly on previous (short term) temperature observations and not on indirect prediction using more heteroclit predictors. No article was found that addresses such a prediction problem.

2 Data and feature construction

Features for use in machine learning are constructed by processing both in situ and remotely sensed data. Of course, before any feature construction can occur, suitable datasets need to be found. That is datasets which exhibit both sufficiently high spatiotemporal resolution and temporal coverage. Thanks to the free availability of works produced by United States governmental agencies, a very large amount of historical and current in situ and remote sensing geophysical data can easily be obtained. This makes the exploratory evaluation of different datasets possible. Given its very large archive of climate data covering the Greenland coast, the U.S. National Climatic Data Center became the de facto source of in situ data. Concerning remote sensing data, the diversity of products made the choice more delicate. The important cloud cover over the Arctic means that optical observations are unfrequent and microwave observations (which go through clouds) are more adequate. Two datasets, both based on a fusion of observations made by different satellites, emerged as particulary adapted: the optimally interpolated sea surface temperatures developped by Reynolds et al. 2007 and the bootstrap sea ice concentrations developped by Comiso 1999, updated 2012.

2.1 NOAA weather station data

Source: All weather data was obtained through the NOAA[3] National Climatic Data Center[4].

Summary: The variables measured by NOAA weather stations depend on the location and period of observation. However stations at most of the locations have been sampling wind direction, wind speed, atmospheric temperature, dewpoint and sea level pressure regardless of the period. Observation availability and frequencies are highly variable, depending on the station.
A limited number of weather stations was chosen using a clustering procedure which is described further on (see section 2.1.1). Among the chosen stations, the data was processed to keep only to most reliable observations. A brief description [5] of the weather station data is given in table 1.

Usage: This dataset was employed to construct daily wind speed, wind direction and sea level pressure features as well as daily atmospheric temperature labels.

[3] I.e. National Oceanic and Atmospheric Administration
[4] Cf. `http://www.ncdc.noaa.gov/cdo-web/`
[5] A detailed description can be found in appendix D

Table 1: NOAA weather station metadata

Type	**Attribute**	**Unit/Format**	**Description**
Identification	USAF	######	U.S. Air Force identifier
	NCDC	#####	National Climate Data Center identifier
Time	Date	[YYYYMMDD]	Year, Month, Day
	HrMn	[HHMM]	Coordinated Universal Time Code (UTC)
Wind	I	#	Identification data source flag
	Type	Obs.	Type of geophysical surface observation
	Dir	Angular degrees	Direction angle
	Q	[0...9]	Quality code
	I	#	Type code
	Spd	m/s	Speed rate
	Q	[0...9]	Quality code
Temperature	Temp	°C	Air temperature
	Q	[0...9]	Quality code
Dewpoint	Dewpt	°C	Dewpoint air temperature
	Q	[0...9]	Quality code
Sea level pressure	SLP	hPa	Sea level atmospheric pressure
	Q	[0...9]	Quality code

2.1.1 Weather station selection

A preliminary screening of all coastal NOAA weather stations in Greenland was made to eliminate those with too few observations in the period 1978-2012. For the remaining stations, the choice of which ones to study was done using the following procedure:

1. Aggregation of similar temperature time series into clusters using the K-means algorithm (see fig. 2). A "top-down" clustering method which implies specifying the number of desired clusters in a dataset (2 to 7 clusters here). The algorithm then proceeds to attribute a cluster to each element of the dataset. The aggregation obtained with K-means clustering was complemented with a simple dendrogram analysis (based on euclidean distance) which allowed a quantitative visualization of similarity between temperature time series (see fig. 3). The dendrogram is based on the Ward algorithm which aggregates the two most similar station time series at each step in a "bottom-up" progression.

2. Amongst each clusters recursively present with the K-means and Ward aggregations, selecting the station with the most complete (a priori) observation record.

A total of six weather stations were kept for the study (see fig. 4 and table 2), after eliminating coarsly redundant time series.

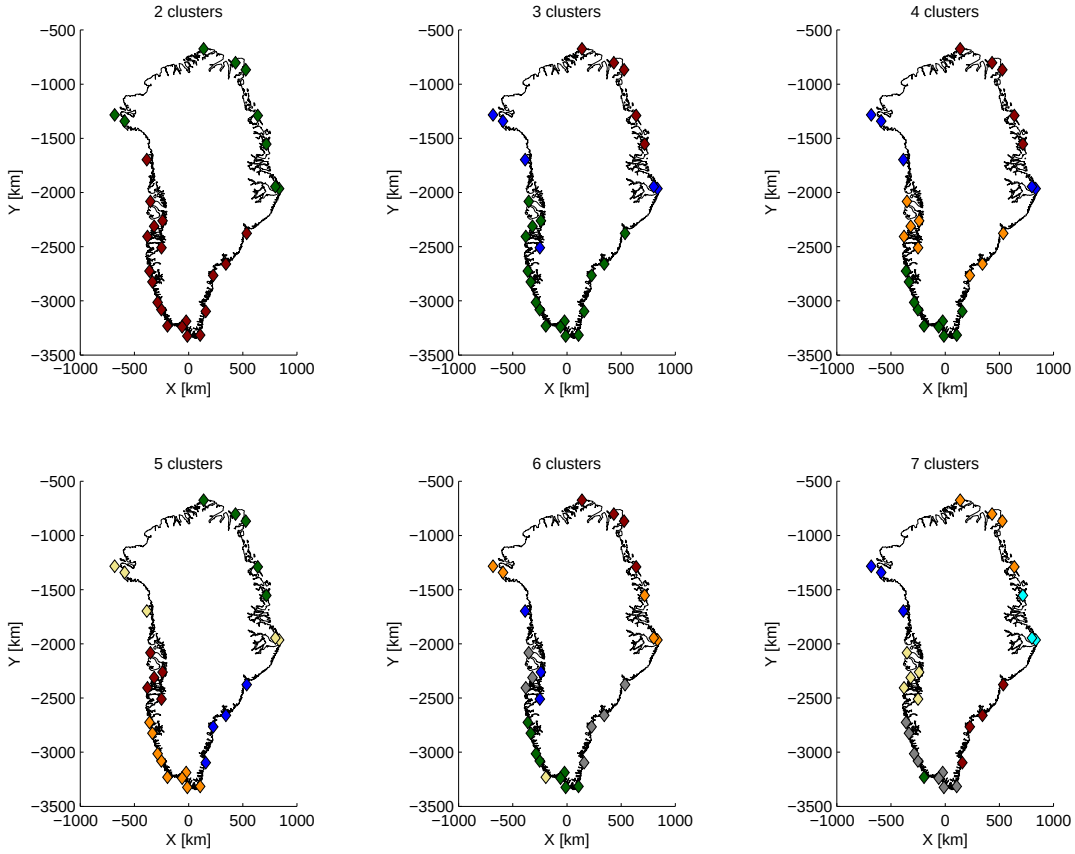

Figure 2: Air temperature time series aggregation using the K-means algorithm with squared euclidean distance (1978-2012). The North Pole is located at (0,0).

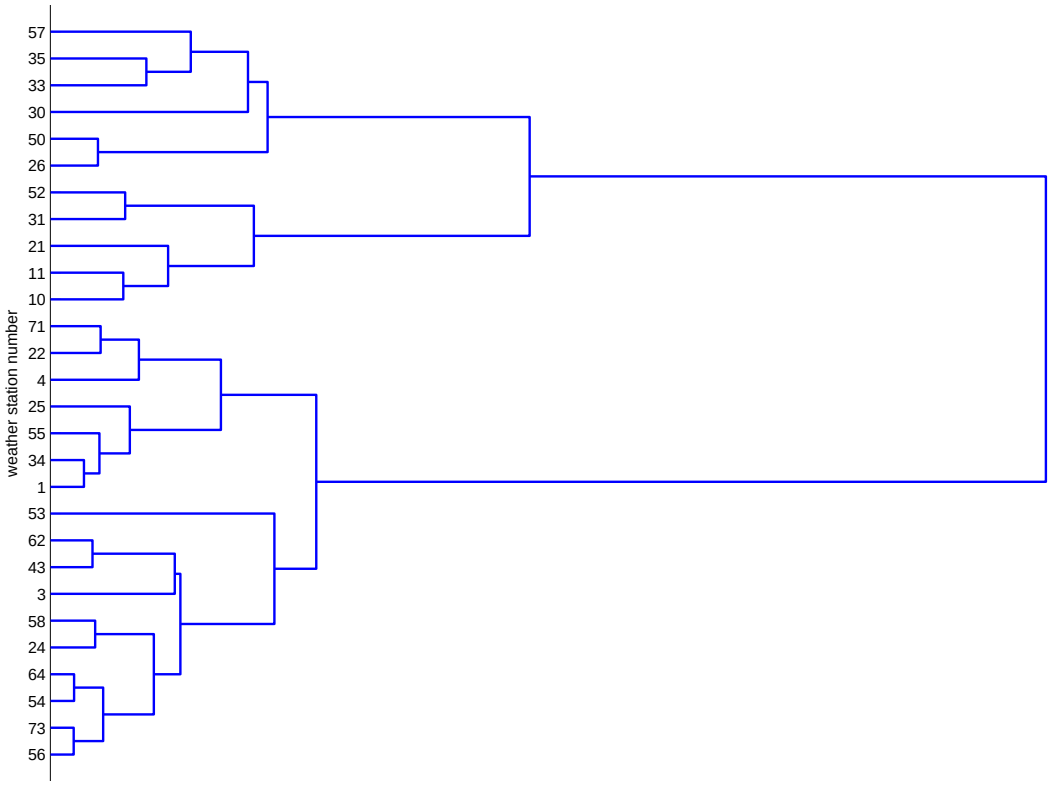

Figure 3: Air temperature time series aggregation dendrogram (using the Ward algorithm)

Unsurprisingly, it can be seen that clusters are mostly formed as a function of latitude. In other words, stations located in the same latitude band tend to have similar atmospheric temperature time series. A longitudinal effect can be observed as well; stations on the West coast in the same cluster as those on the East coast have higher latitudes. This may be due to the warming effect of the West Greenland current (see fig. 1). Some stations also exhibit particular behaviors likely due to local conditions. It was decided to avoid these outlier stations. The final choice of stations is shown in figure 4 and their geographic coordinates are provided in table 2.

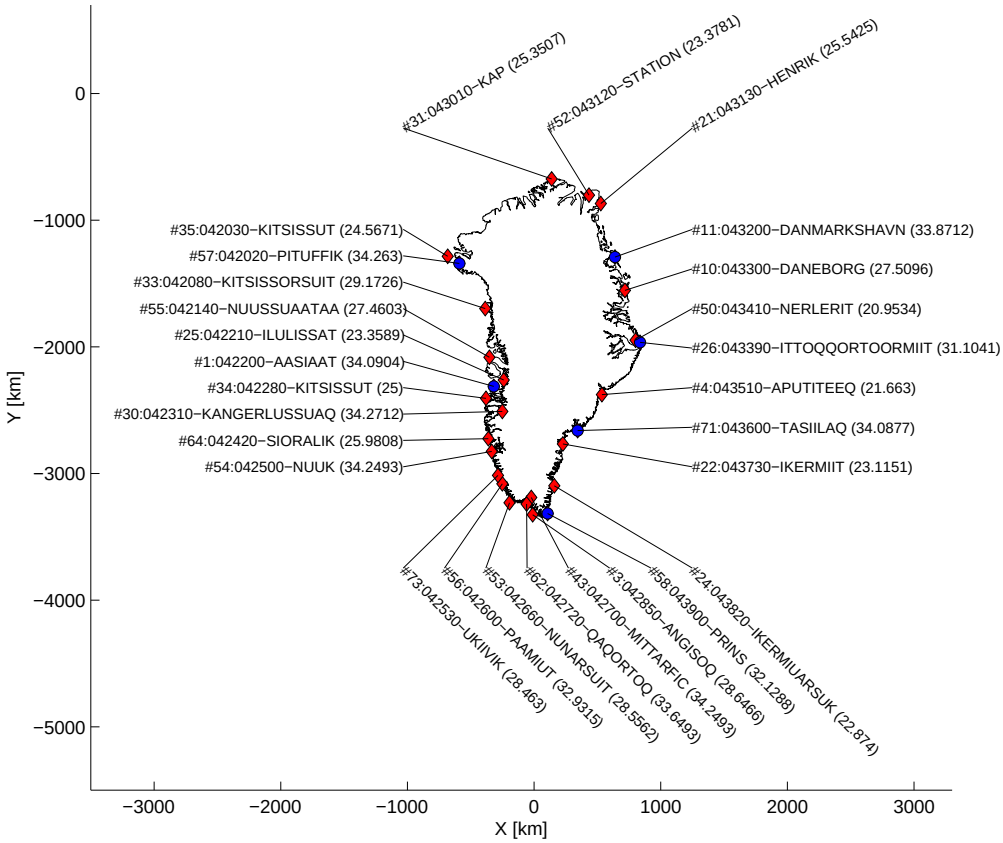

Figure 4: Map of NOAA weather stations in Greenland (North Pole located at (0,0)). Only stations with a sufficiently complete data record for the period of interest are represented. Blue diamonds indicate the chosen stations amongst each cluster. The values to the right of the station names are the cumulative years of available data after 1978.

Table 2: Location of the studied NOAA weather stations

Station name	**USAF#**	**Lat/Lon**	**Elevation**
Danmarkshavn	043200	76.767° N / 18.667° W	12 m
Ittoqqortoormiit	043390	70.483° N / 21.950° W	69 m
Tasiilaq / Ammassali	043600	65.600° N / 37.633° W	52 m
Prins Christian Sund	043900	60.050° N / 43.167° W	75 m
Aasiaat / Egedesmind	042200	68.700° N / 52.850° W	41 m
Pituffik (Thule Airbase)	042020	76.533° N / 68.750° W	59 m

2.1.2 Quality and sampling check

Concerning weather stations, the heterogeity of synoptic observation frequencies (see fig. 5) and data quality was an important constraint. In order to make sure only the most accurate data was used in all subsequent computations, both quality and sampling checks were applied to the original datasets.

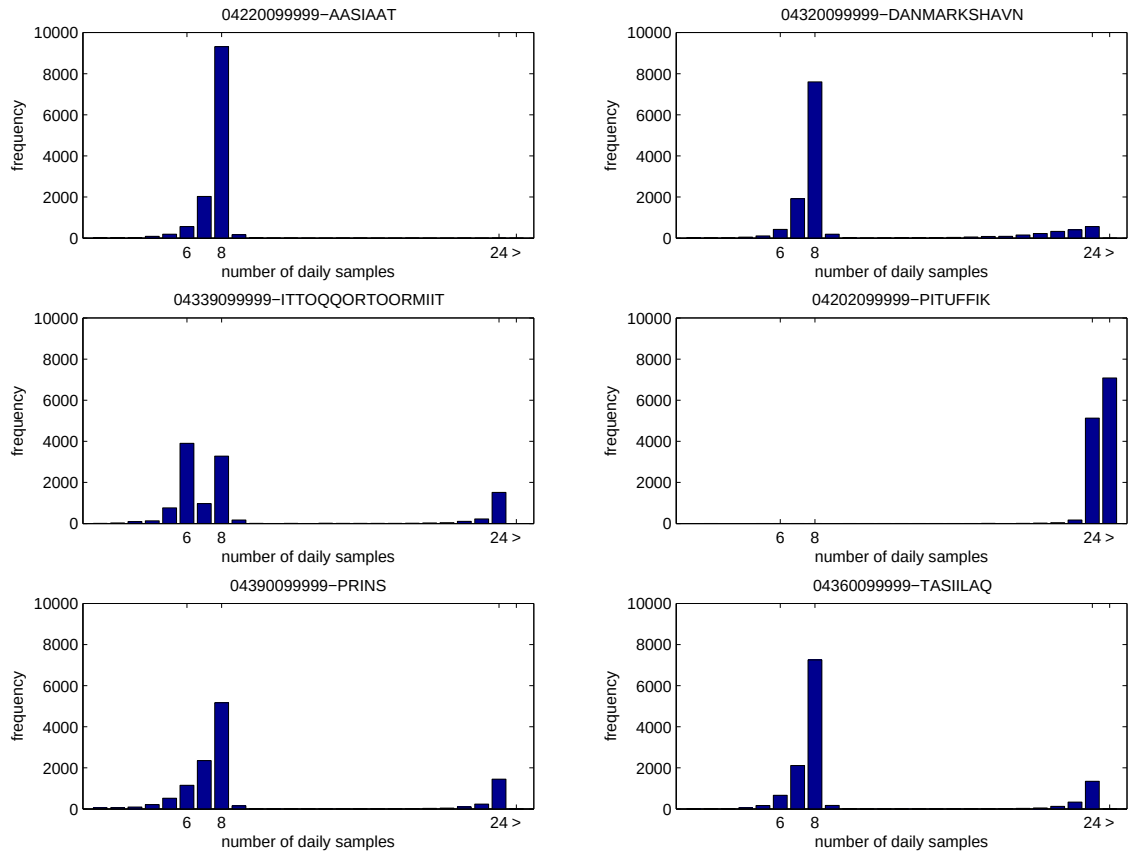

Figure 5: Heterogeneity of synoptic observation frequencies

Aasiaat:  | Danmarkshavn:  | Ittoqqortoormiit:  | Pituffik:  | Prins Christian Sund:  | Tasiilaq: 

It was decided to keep only daily measurements which were at least 80% similar to any of the standard hourly [6], 3-hourly [7] or 4-hourly [8] sampling patterns. All observations which were not sufficiently similar to these standard sampling patterns, either because they were incomplete or followed a non-standard sampling scheme, were discarded.

[6] i.e. 0000,0100,0200,0300,0400,0500,0600,0700,0800,0900,1000,1100,1200,1300,1400,1500,1600,1700,1800,1900,2000,2100,2200,2300

[7] i.e. 0000,0300,0600,0900,1200,1500,1800,2100

[8] i.e. 0000,0400,0800,1200,1600,2000

The similarity between any daily observation set ($\vec{H}_m$) and the standard observation patterns ($\vec{H}_s$) was computed with the following pseudocode:

Algorithm 2.1: SIMILARITY($\vec{H}_m, \vec{H}_s$)

$\vec{H}_m = [H_{m1}\ H_{m2}\ ...\ H_{m,N_m}]$

$\vec{H}_s = [H_{s,1}\ H_{s,2}\ ...\ H_{s,N_s}]$

$S = [6;\ 8;\ 24]$

for each $N_s \in \mathcal{S}$

do $Similarity_{N_s}(\vec{H}_m, \vec{H}_s) = \frac{\sum ismember(\vec{H}_m, \vec{H}_s)}{N_s}$

if $Similarity_{N_s}(\vec{H}_m, \vec{H}_s) \geq 0.8$

then $KEEP$

else $Similarity_{N_s}(\vec{H}_m, \vec{H}_s) = 0$

output $(logical(\sum Similarity_{N_s}))$

Where:

- N_m is the number of measurements in the tested sample
- N_s is the standard number of measurements for a specific sampling pattern

Below is an example of a valid sampling date at station Aasiaat/Egedesmind (on Feb 5^{th} 1978):

```
Identification                    WIND            TEMP     DEWPT    SLP
USAF   NCDC  Date     HrMn I Type  Dir Q I Spd   Q Temp   Q Dewpt  Q Slp    Q
042200,99999,19780205,0000,4,FM-12,080,1,N,  3.6,1, -15.0,1, -17.0,1,1036.1,1,
042200,99999,19780205,0300,4,FM-12,080,1,N,  3.1,1, -15.0,1, -17.0,1,1036.1,1,
042200,99999,19780205,0600,4,FM-12,080,1,N,  3.6,1, -15.0,1, -17.0,1,1035.4,1,
042200,99999,19780205,0900,4,FM-12,060,1,N,  4.6,1, -14.0,1, -16.0,1,1035.3,1,
042200,99999,19780205,1100,4,FM-12,999,9,9,999.9,9, 999.9,9,  -9.0,1,9999.9,9,
042200,99999,19780205,1200,4,FM-12,050,1,N,  4.1,1, -14.0,1, -16.0,1,1035.3,1,
042200,99999,19780205,1500,4,FM-12,050,1,N,  1.5,1, -14.0,1, -16.0,1,1034.2,1,
042200,99999,19780205,1800,4,FM-12,080,1,N,  1.0,1, -15.0,1, -17.0,1,1032.9,1,
042200,99999,19780205,2100,4,FM-12,040,1,N,  2.1,1, -14.0,1, -16.0,1,1032.2,1,
```

Notice that there are nine normal measurements in this series. The pattern is very similar to the standard 3-hourly sampling. However, there is an extra unvalid measurement at 11:00. Computing the similarity of this sample with the standard 3-hourly sampling pattern, we have:

$$\vec{H}_m = [0000\ 0300\ 0600\ 0900\ 1100\ 1200\ 1500\ 1800\ 2100]$$

$$\vec{H}_s = [0000\ 0300\ 0600\ 0900\ 1200\ 1500\ 1800\ 2100]$$

$$Similarity(\vec{H}_m, \vec{H}_s) = \frac{\sum[1\ 1\ 1\ 1\ 0\ 1\ 1\ 1\ 1]}{8} = 1 > 0.8 \rightarrow keep$$

This procedure is repeated with the 4-hourly and hourly sampling patterns. The three resulting values are sumed up. If the result is non null the measure is kept, otherwise it is discarded. It is important to note that only the measures with adequate quality flags are used.

2.2 Optimally interpolated sea surface temperatures

Source: All sea surface temperature (SST) data was obtained through the Physical Oceanography Distributed Active Archive Center [9] hosted by the Jet Propulsion Laboratory (CalTech).

Summary: The product is based on a fusion of data from the Advanced Very High Resolution Radiometer (AVHRR) infrared satellite SST and in situ data from ships and buoys. It includes a large-scale adjustment of satellite biases with respect to the in situ data. A summary of the metadata is provided in table 3.

Usage: This dataset was employed to construct median, maximum and minimum sea surface temperatures features within different distance ranges around the weather stations.

Table 3: Sea surface temperature metadata

Category	**Description**
Parameter	Sea surface temperature
Data format	NETCDF
Spatial coverage	Global
Temporal coverage	01 September 1981 - Present
Temporal resolution	Daily
Projection	Cylindrical Lat-Lon WGS 84 ellipsoid (a = 6378.137 km, e = 0.081819190)
Nominal grid resolution	0.25° (latitude) x 0.25° (longitude)
Journal reference	Reynolds et al. 2007

[9] Cf. `http://podaac.jpl.nasa.gov/dataset/NCDC-L4LRblend-GLOB-AVHRR_OI`

2.3 Bootstrap sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I-SSMIS

Source: All sea ice concentration data was obtained through the National Snow and Ice Data Center [10].

Summary: According to the official description, the sea ice concentration data set was derived using measurements from the Scanning Multichannel Microwave Radiometer (SMMR) on the Nimbus-7 satellite and from the Special Sensor Microwave/Imager (SSM/I) sensors on the Defense Meteorological Satellite Program's (DMSP) -F8, -F11, and -F13 satellites. Measurements from the Special Sensor Microwave Imager/Sounder (SSMIS) aboard DMSP-F17 are also included. The data set has been generated using the Advanced Microwave Scanning Radiometer (AMSR-E) - Earth Observing System Bootstrap Algorithm with daily varying tie-points. A summary of the metadata is provided in table 4.

Usage: This dataset was employed to construct sea ice concentration features within different distance ranges around the weather stations.

Table 4: Sea ice concentration metadata

Category	**Description**
Parameter	Sea ice concentration
Data format	Flat binary two-byte integer; little-endian byte order; scaled by 10
Spatial coverage	Upper left corner: 30.98° N / 168.35° E Upper right corner: 31.37° N / 102.34° E Lower left corner: 33.92° N / 279.26° E Lower right corner: 34.35° N / 350.03° E
Temporal coverage	26 October 1978 - 31 December 2010
Temporal resolution	Daily (every other day for SMMR data)
Projection	Polar stereographic Hughes ellipsoid (a = 6378.273 km, e = 0.081816153) True scale lat.: 70° N
Nominal grid resolution	25 km x 25 km
Journal reference	Comiso 1999, updated 2012

[10] Cf. `http://nsidc.org/data/docs/daac/nsidc0079_bootstrap_seaice.gd.html`

2.4 Feature construction

A total of 134 features were employed for the prediction. Sea, ice, wind and atmospheric pressure features (see table 5) were lagged in time using a lag of one and two days. Thus, each of these features appear three times and they are distinghished by the suffixes "D" (current day), "D-1" (previous day), "D-2" (two days before). For each weather station, spatial statistics (i.e. median and maximum) on remote sensing data were computed using the values inside two small circles[11] centered on the station: a short radius small circle (75 km) and a long radius small circle (300 km). The corresponding features are respectively designated by "short range" and "long range" in table 5. These different ranges were chosen in an attempt to separate local and regional offshore influences (see fig. 6). Figures 9 and 10 are an example of how sea ice concentration inside long and short range small circles varies during the year.

In order to avoid the problem of averaging sub-daily wind direction observations, it was decided to use the normalized frequencies at which particular wind directions appeared each day. Indeed, even if wind direction is decomposed into cartesian UV coordinates, averaging the components could yield misleading results. For example, if the wind was blowing half the time in a direction and the other half in the opposite direction, the average would be zero. However, this is clearly not the same weather condition as a calm day with no wind.

Daily wind speed and sea level atmospheric pressure features were obtained by simply averaging sub-daily observations.

Finally, since it is a cyclic feature, the day of year was transformed into cosine(day of year) and sine(day of year) features. This accounts for the fact that daily conditions at the beginning of a year are very similar to those at the end of the previous year (e.g. December 31^{st} 2011 is almost the same as January 1^{st} 2012).

Table 5: List of features (n.b. lagged features are not listed)

Type	**Feature**	**Unit**
Sea	Daily sea surface median temperature (short range)	°C
	Daily sea surface max temperature (short range)	°C
	Daily sea surface median temperature (long range)	°C
	Daily sea surface max temperature (long range)	°C
Ice	Daily sea ice fraction (short range)	%
	Daily sea ice fraction (long range)	%
Wind	Daily mean wind speed	m/s
	Daily wind fraction from azimuth 010°	%
	Daily wind fraction from azimuth 020°	%
	⋮	⋮
	Daily wind fraction from azimuth 360°	%
Pressure	Daily mean sea level pressure	hPa
Time	Sine(day of year)	[-]
	Cosine(day of year)	[-]

All the daily features presented above were computed for the period 1981-2010 and were then grouped by month. This resulted in 12 matrixes (i.e. one per month) containing all features for the period of interest (see fig. 7). After the construction of all the in situ and remote sensing daily features, only those which were simultaneously available for at least three consecutive days were selected to be used in support vector regression (see fig. 8). This restricted the features to years 1981-2010.

[11] A small circle is a circle defined as the intersection of a sphere and a plane when the plane does not contain the center of the sphere.

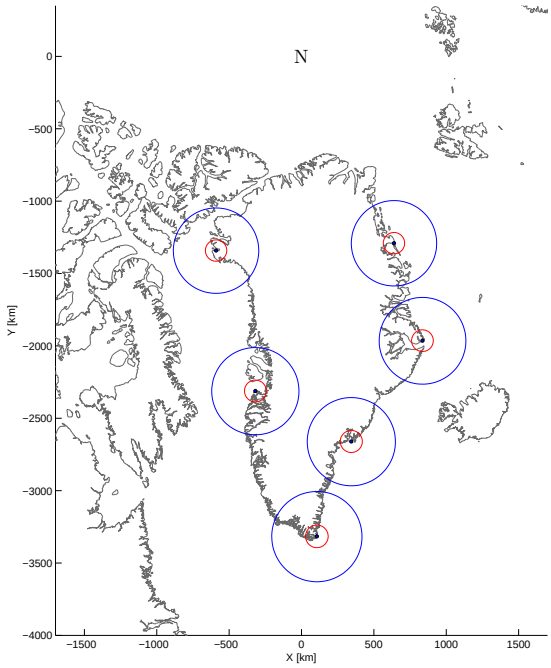

Figure 6: Weather stations with long range (300 km) and short range (75 km) small circles plotted around each station.

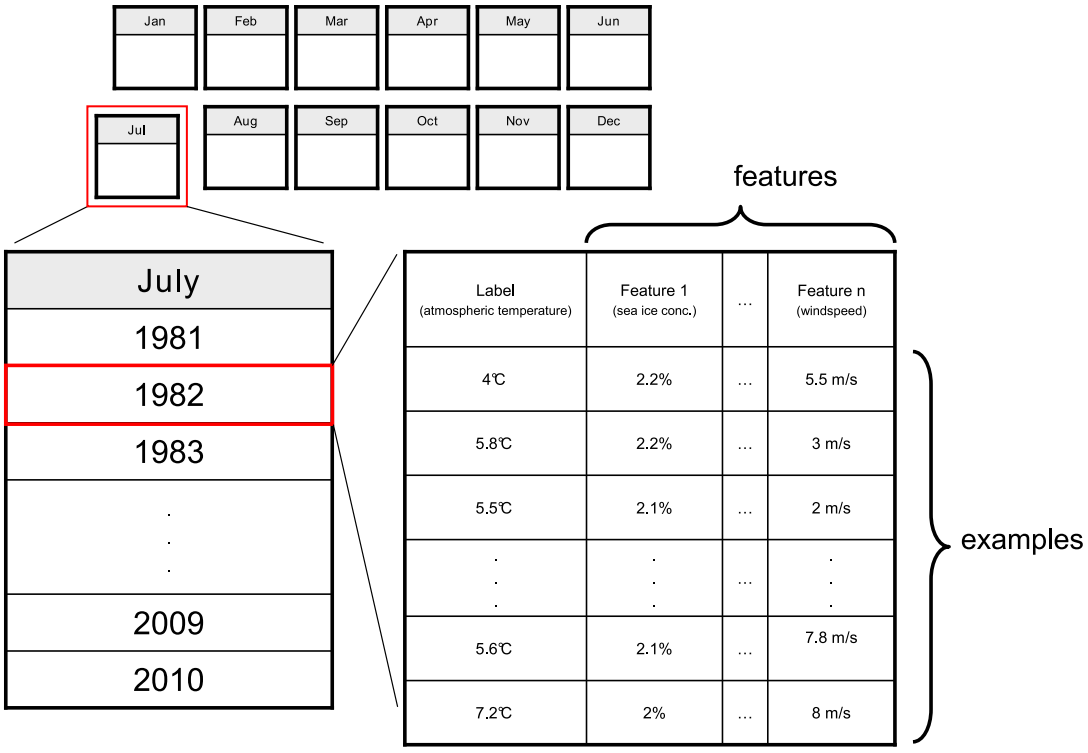

Figure 7: Features for 1981-2010 are organized by month and separated into twelve monthly matrixes.

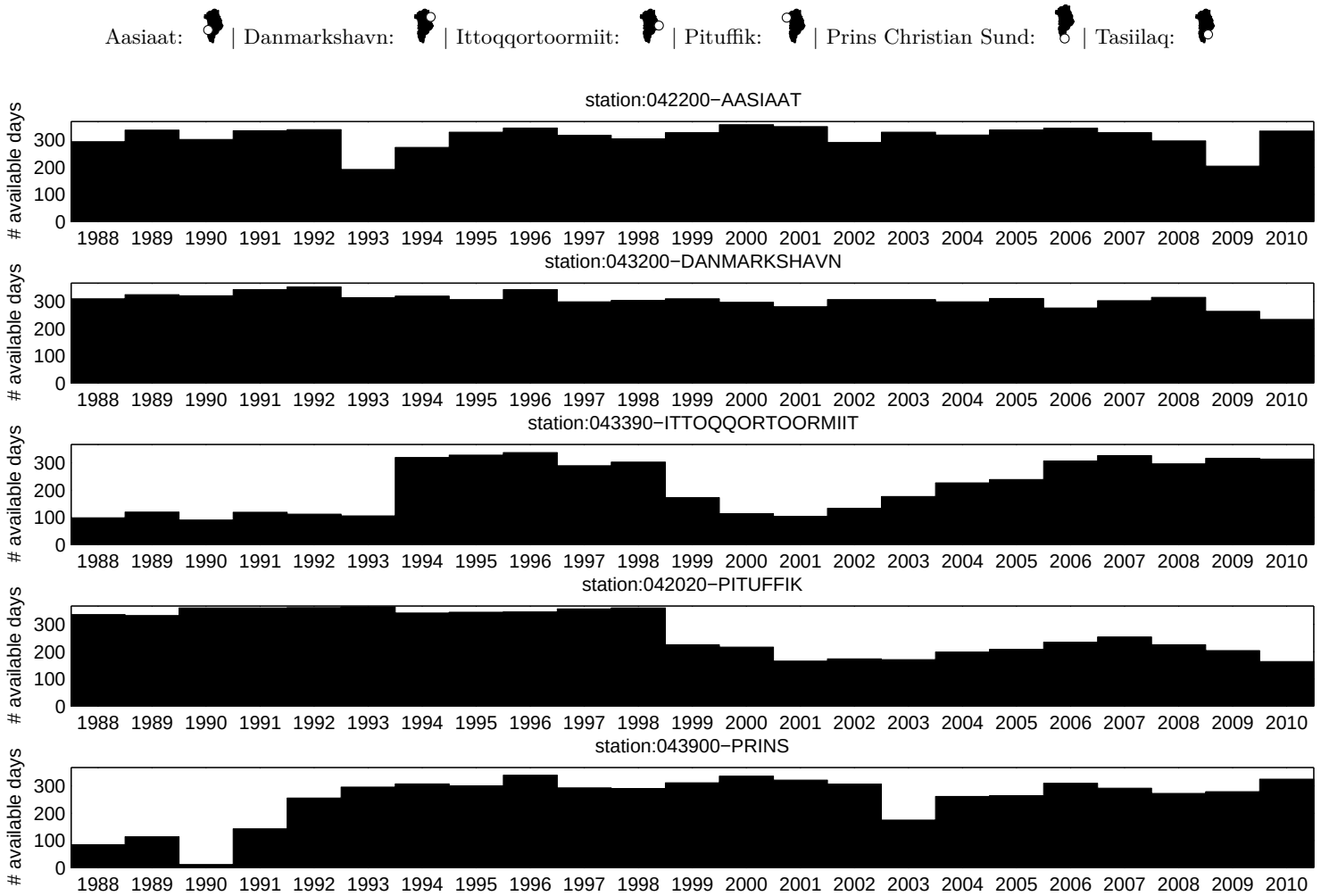

Figure 8: Combined feature availability. Some of the yearly availabilities are very low. This is normal given the very restrictive constraints which apply. Not only do valid daily observations have to be simultaneously available in all of the datasets (in situ and remotely sensed), but they also have to be available for at least 3 consecutive days (because features are lagged)

.

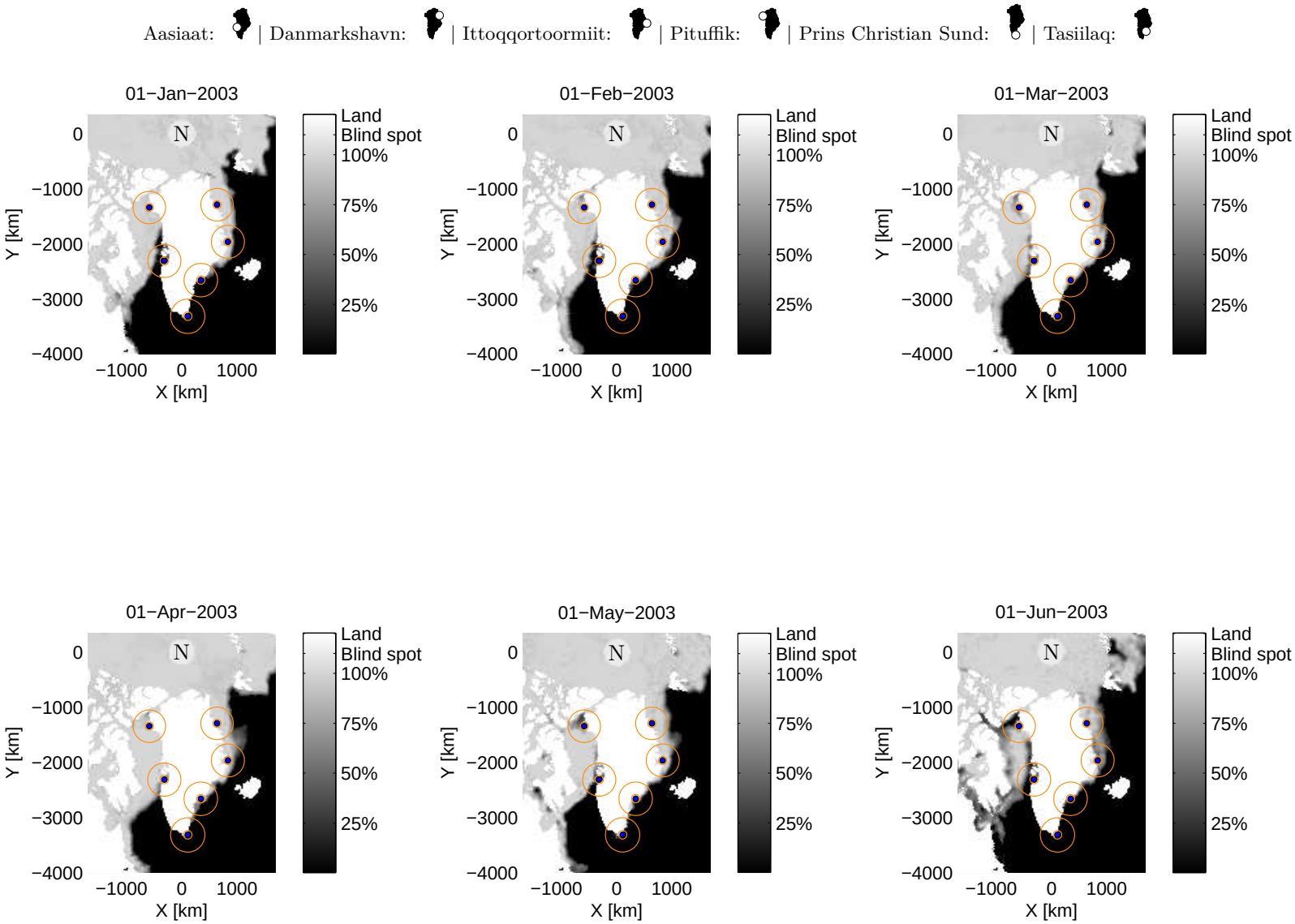

Figure 9: Sea ice concentration monthly sequence (January to June 2003)

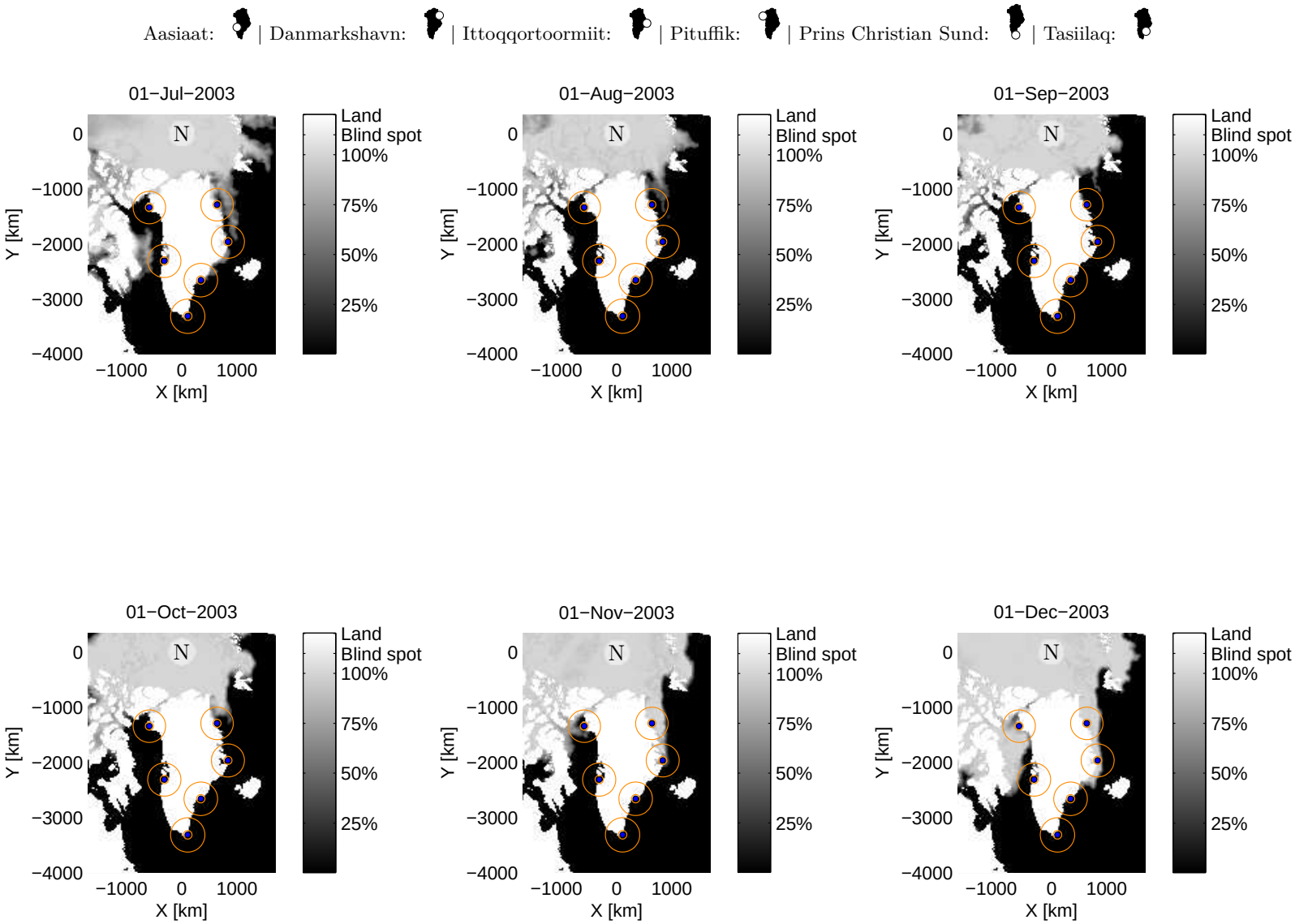

Figure 10: Sea ice concentration monthly sequence (July to December 2003)

3 Preliminary Analysis

This section provides a series of structural and comparative visualizations for weather station variables in the period 1978-2012. It is aimed at determining any particularly evident relations between wind and temperature regimes for each of the chosen weather stations. The cross-correlation between the atmospheric and regional sea surface temperature time series is also computed in search of a time lag. A simple seasonal atmospheric temperature model is created and the deseasonalized and detrended atmsopheric temperature time series are analyzed.

3.1 Station overview

3.1.1 Daily mean air temperature distributions

All of the atmospheric temperature distributions represented below (see fig. 11) are different.

This supports the choice of weather stations made previously. Two groups of distributions can be distinguished. The three most Northerly stations, Danmarkshavn , Ittoqqortoormiit and Pituffik have bimodal temperature distributions. On the other hand, stations Aasiaat , Prins Christian Sund and Tasiilaq which are located in the Southern half of Greenland exhibit a unimodal distribution. The bimodal temperature distribution is likely related to the yearly appearance and disappearance of sea ice in Northern Greenland. Excluding polynyas[12], these parts of the island have no open water areas nearby during more than half of the year.

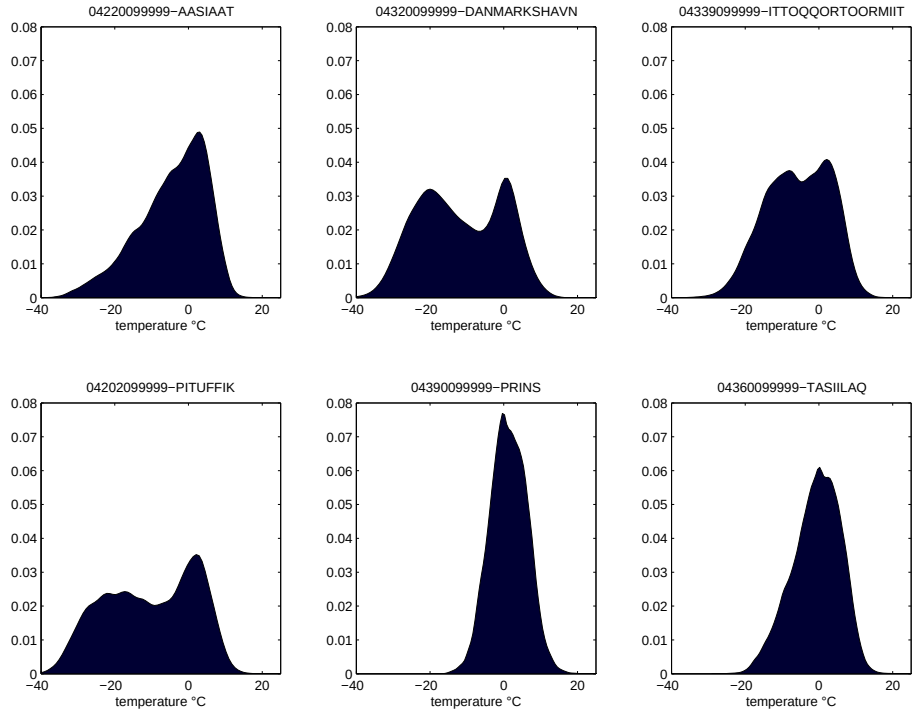

Figure 11: Kernel smoothing probability density estimates of daily mean air temperature

[12]Polynya, is a word from the Russian language which means "natural ice hole". It is used to describe areas of unfrozen sea within the ice pack. Such areas can be the results of strong winds and/or ocean currents exporting the sea ice, thus creating an opening in the ice pack. Such openings can also be caused by warm water upwelling.

3.1.2 Wind regime and ocean currents

The distribution of wind types appears to be highly dependant on location (see fig. 12). All of the stations except Ittoqqortoormiit exhibit a maximum sea wind fraction in summer. Stations Aasiaat and Danmarkshavn have a similar behavior. The fractions of sea and land winds are usually inversely proportional. Analysis of wind azimuth distributions (see appendix A) provides further explanation of the wind types at each station:

- **Station Aasiaat**  **(West)** is largely dominated by land winds. In summer the fraction of sea and land winds are equal. A glimpse at the wind direction histograms (see fig. 32) reveals that winds tend to blow from the South and the Northeast (and quite frequently from azimuths between the two) during winter. In summer a Southwesterly wind appears. The offshore area around this station is ice free most of the year, so the West Greenland Current is likely to have an influence.

- **Station Danmarkshavn**  **(Northeast)** (see fig. 33) has a relatively constant fraction of calm days. Sea winds become dominant in summer, as two new wind directions appear (from the South and from the East). The rest of the year, the wind blows from the Northwest (see fig. 33). The offshore area around this Northern station is covered in sea ice most of the year, so contact with the sea surface seems limited.

- **Station Ittoqqortoormiit**  **(East)** is dominated by sea winds all year long. Land winds are marginal at this location. Most of the year, the winds blows from the North-Northeast. In summer, winds with Southern components appear (see fig. 34). This station is surrounded by offshore sea ice most of the year (slightly less than Danmarkshavn).

- **Station Pituffik**  **(Northwest)** is almost exclusively dominated by land winds in winter. In summer, the fraction of sea winds becomes larger than land winds (see fig. 31). Despite its high Northern latitude, Pituffik remains relatively near open water about half of the year. This is because it is located near the North water polynya. According to Smith, Barber 2007, the largest polynya in the Canadian Arctic and one of the most biologically-productive polynyas in the Northern Hemisphere.

- **Station Prins Christian Sund**  **(South)** exhibits windy conditions all year (very low, nearly constant fraction of calm wind conditions). In summer the fraction of sea winds becomes larger than land winds. The location of the station at the Southern tip of Greenland explains the high exposure to wind. The wind direction histograms (see fig. 36) show two dominant wind directions. Westerly and Northeasterly winds which are known as tip jets occur very frequently. Vage et al. 2009 suggest that the Westerly tip jet arises from the interplay of the synoptic-scale flow evolution and the perturbing effects of Greenland's topography upon the flow. Station Prins Christian Sund, is hardly ever under the direct influence of the ice shelf, due to its exposure to the warm Irminger current.

- **Station Tasiilaq**  **(Southeast)** exhibits a particular behavior. Having almost equal fractions of sea and calm wind conditions. Land winds are dominant in winter. The wind direction histograms (see fig. 35) indicate there are two wind regimes. In fall and winter, Westerly winds (i.e. blowing from inland) are dominant. These are cold and high speed katabatic winds know as Piteraq. In spring and summer, Southeasterly winds become dominant. The waters of Tasiilaq are ice free most of the year and are partly influenced by the Irminger Current.

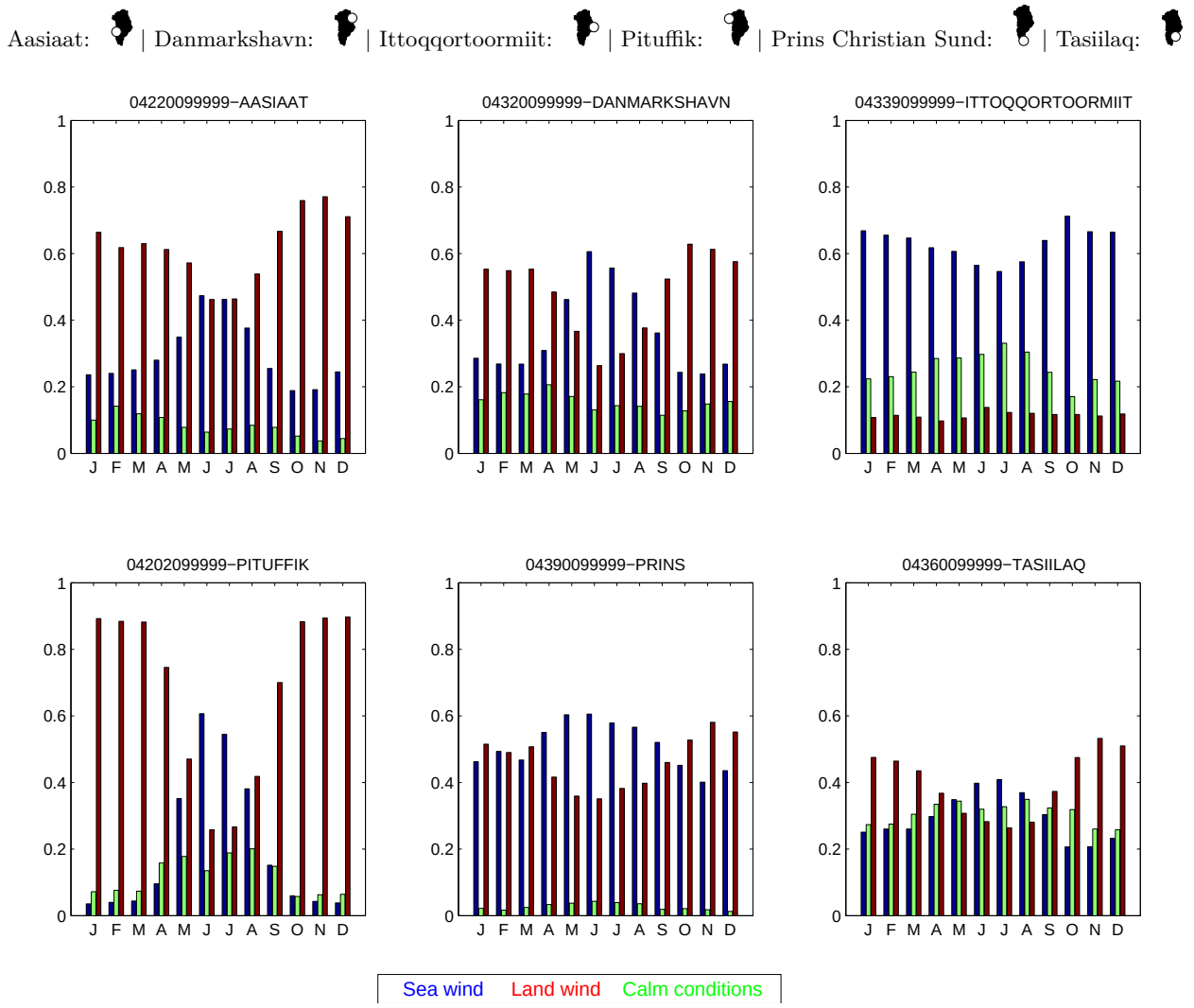

Figure 12: Monthly Sea/Land/Calm wind distributions

3.1.3 Air temperature distributions as a function of wind conditions

After visualizing the frequencies of different wind directions, the next step was to determine if they were associated to specific atmospheric temperature distributions. The kernel smoothing density estimates for different wind conditions using the temperature time series with both trend and seasonality is represented in figure 13. Stations Asiaat , Danmarkshavn and Pituffik clearly have higher temperatures when the wind is blowing from the sea (though this does not necessarily imply causality). For stations Prins Christian Sund and Tasiilaq , this relation is less pronounced. Finally, in Ittoqqortoormiit , land winds seem to be rather associated with higher temperatures. The observations suggest a relation between wind and temperature and motivate the use of wind features in the predictions.

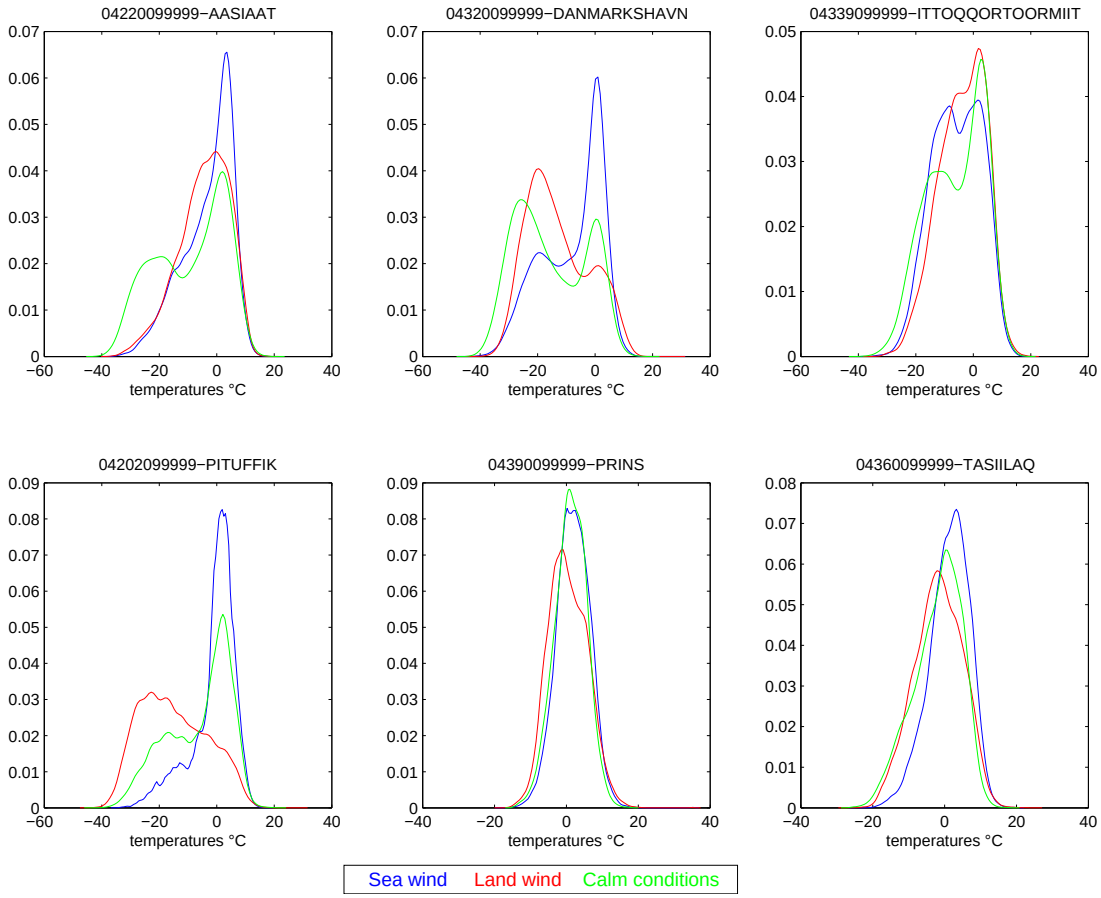

Figure 13: Kernel smoothing density estimates for different wind conditions

Aasiaat: | Danmarkshavn: | Ittoqqortoormiit: | Pituffik: | Prins Christian Sund: | Tasiilaq:

3.1.4 Trend and seasonality modelling

The long term trend and seasonality of atmospheric temperature were modelled separately for each station using a slightly modified version of the moving average approach described in Brockwell, Davis 2002.
The long term trend was determined by applying a centered moving average of order 365 to the daily temperature time series (with missing days linearly interpolated). Seasonality was modelled by averaging daily mean temperatures for each day of the year which resulted in the noisy curves shown in figure 14. In order to smooth the noise, a 10 day centered moving average filter was then applied recusively [13]. The autocorrelation functions of the residuals [14] were computed to verify the effectiveness of the procedure. All residuals show significant autocorrelation (see fig. 15). The use of an autoregressive integrated moving average (ARIMA) model would have been necessary to diminish this autocorrelation. However, as only the original data (i.e. not the residuals) is used in the prediction, the atmospheric temperature time series were not processed any further.

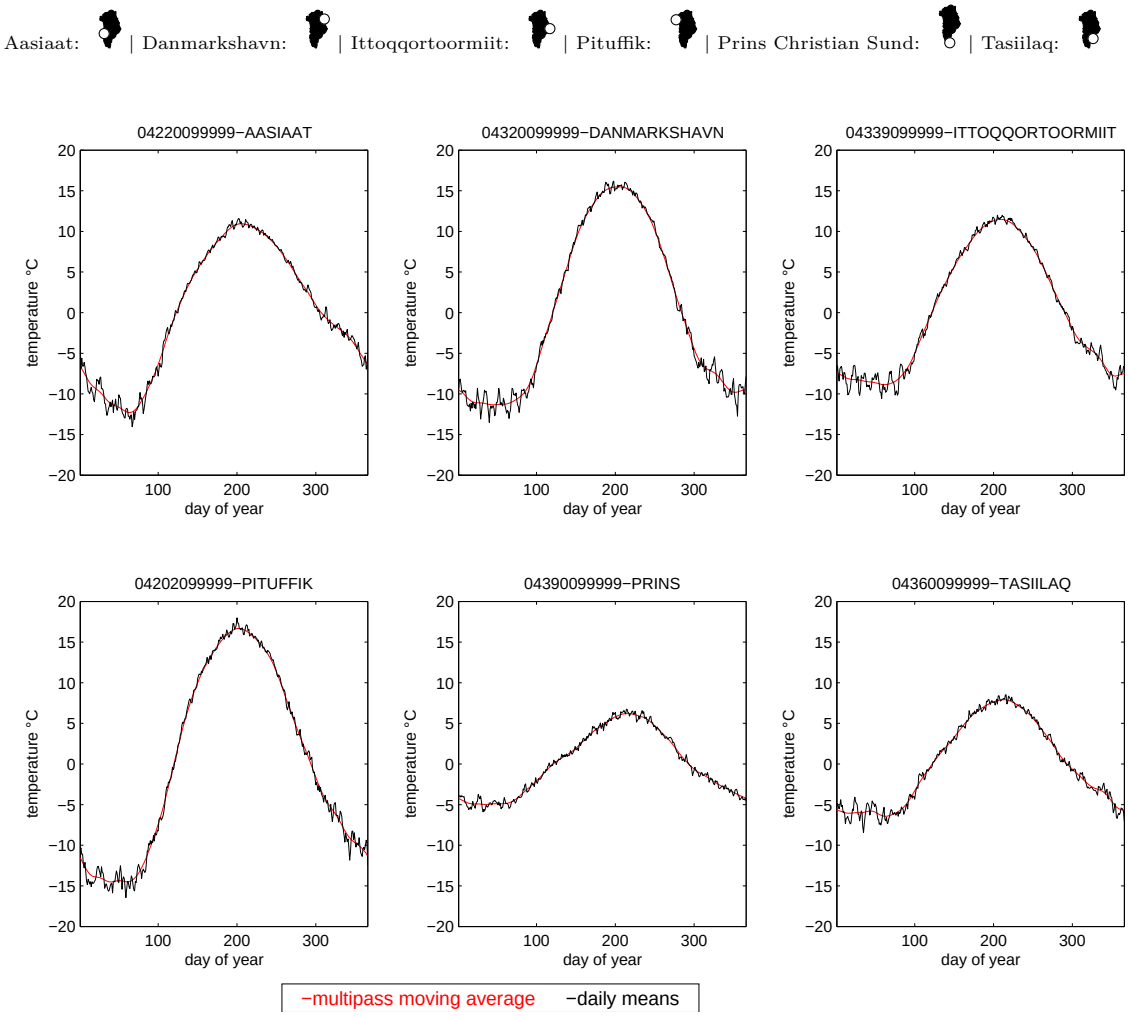

Figure 14: Estimated seasonal component of temperature based on 1978-2012 data. The noisy curves in black are the averaged temperatures of each day of the year for the period of interest. The red curves are the result of the smoothing multipass moving average filter.

[13] I.e. multiple pass moving average filter.
[14] I.e. the original time series minus trend and seasonality.

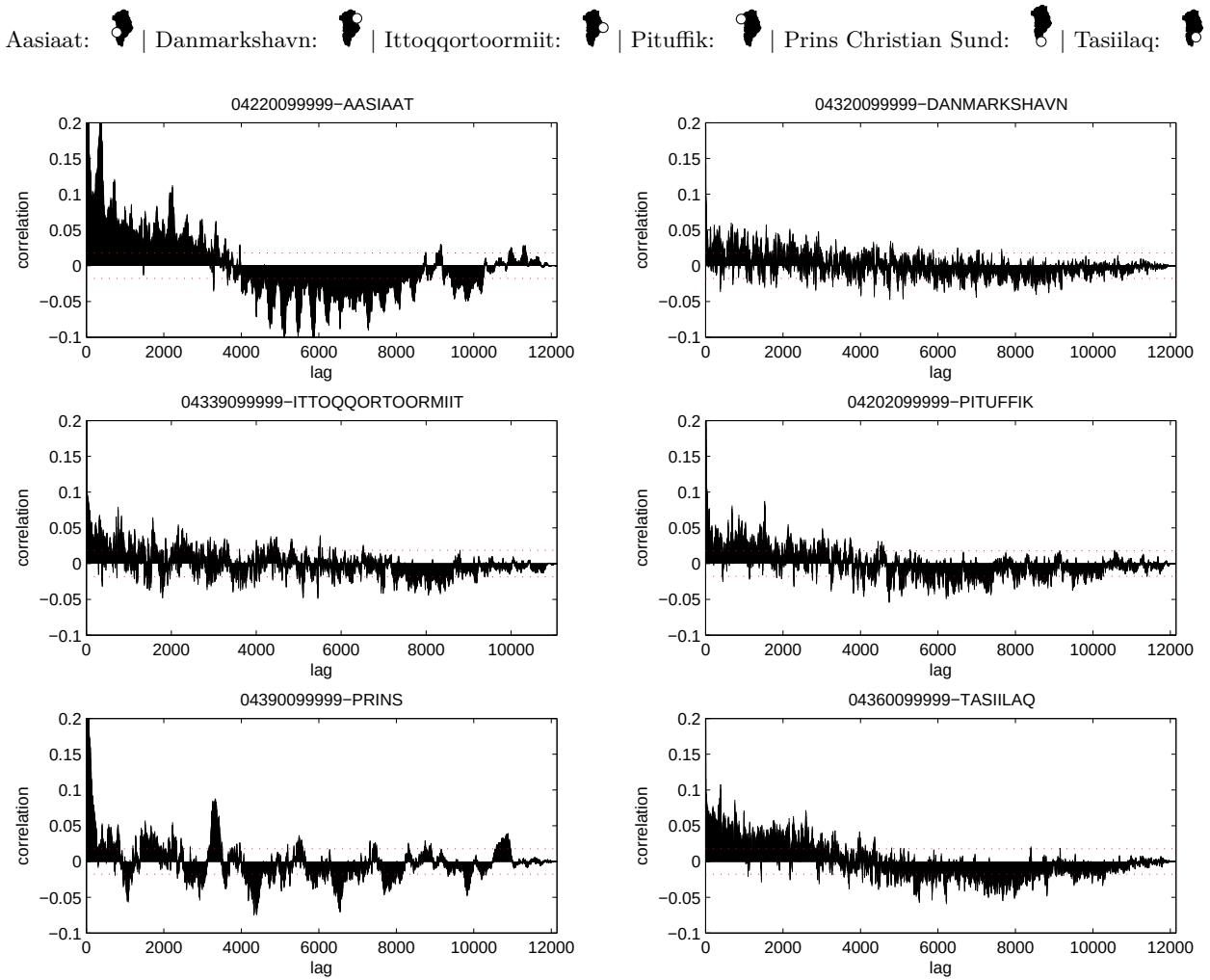

Figure 15: Autocorrelation function of the residuals (multipass moving average model) based on 1978-2012 data. The doted red lines indicate the approximate 95% confidence limits of the autocorrelation function of an IID process. Notice that significant autocorrelation remains in all of the time series, even after detrending and deseasonalizing.

3.1.5 Detrended air temperature distributions as a function of wind conditions

A new series of density estimates, this time without trend and seasonality (discussed previously) are presented in figure 16. The distributions do not show any large differences in temperature depending on wind types. In order to test if the daily mean air temperatures were different for days with sea winds, land winds and calm conditions, a statistical test on the mean was performed. To confirm the clear visual hint that the residuals are not normally distributed, a one-sample Kolmogorov-Smirnov test[15] ($\alpha = 0.05$) was performed. Unsuprisingly, all distributions were determined to be non-normal. Since the distributions of the temperature residuals are not normally distributed, Student's t-tests cannot be employed to determine if the mean temperatures are significantly different. The non-parametric two-sample Kolmogorov-Smirnov test[16] ($\alpha = 0.05$) was used instead. The results of this test showed that the distributions among each station were significantly different. The test did not allow to determine if sea winds were associated to higher temperature residuals. A visual examination seems to indicate a slight difference in the means but without conclusive evidence. However, this result cannot be considered fully reliable as the residuals still have significant autocorrelation (i.e. seasonaliy and trend are still present).

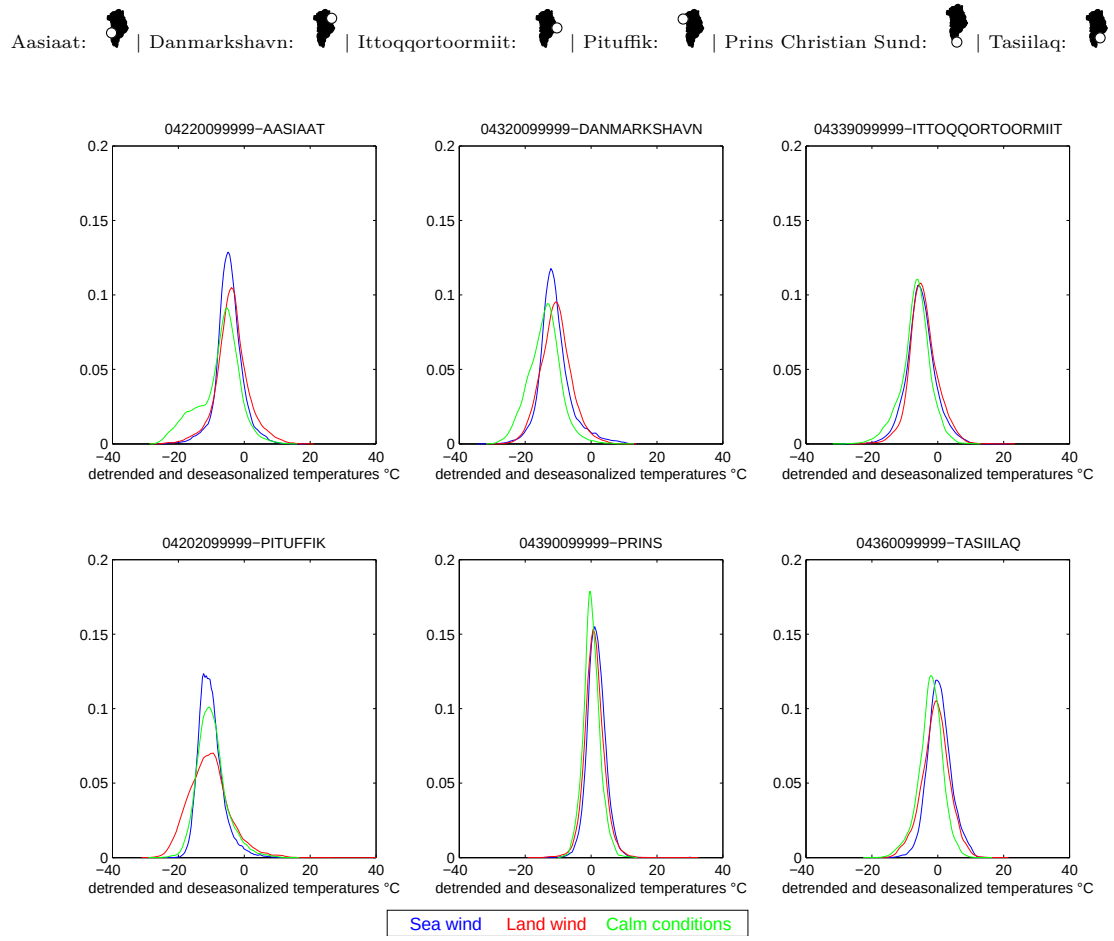

Figure 16: Kernel smoothing probability density estimates for different wind conditions (after detrending and deseasonalizing)

[15]The one-sample Kolmogorov-Smirnov test is used to compare the values in a dataset to a standard normal distribution. The null hypothesis is that the dataset has a standard normal distribution. The alternative hypothesis is that it does not have that distribution.

[16]The two-sample Kolmogorov-Smirnov test is used to compare the distributions of the values in the two datasets. The null hypothesis is that both datasets have the same continuous distribution. The alternative hypothesis is that they have a different continuous distributions.

3.1.6 Distributions of consecutive daily temperature and sea wind fraction differences

The daily mean temperature (see fig. 18) and sea wind fraction (see fig. 19) differences were computed for all consecutive days. It was determined that these differences are well modeled by t location-scale distributions[17]. Unfortunately, no suitable statistical test was found for this family of distributions. In the end, it was chosen to verify if a large difference in sea wind fraction (i.e. larger than one standard deviation σ) was linked to a large temperature difference. A visual inspection of the probability densities (see fig. 17) associated to wind type changes did not reveal such a link.

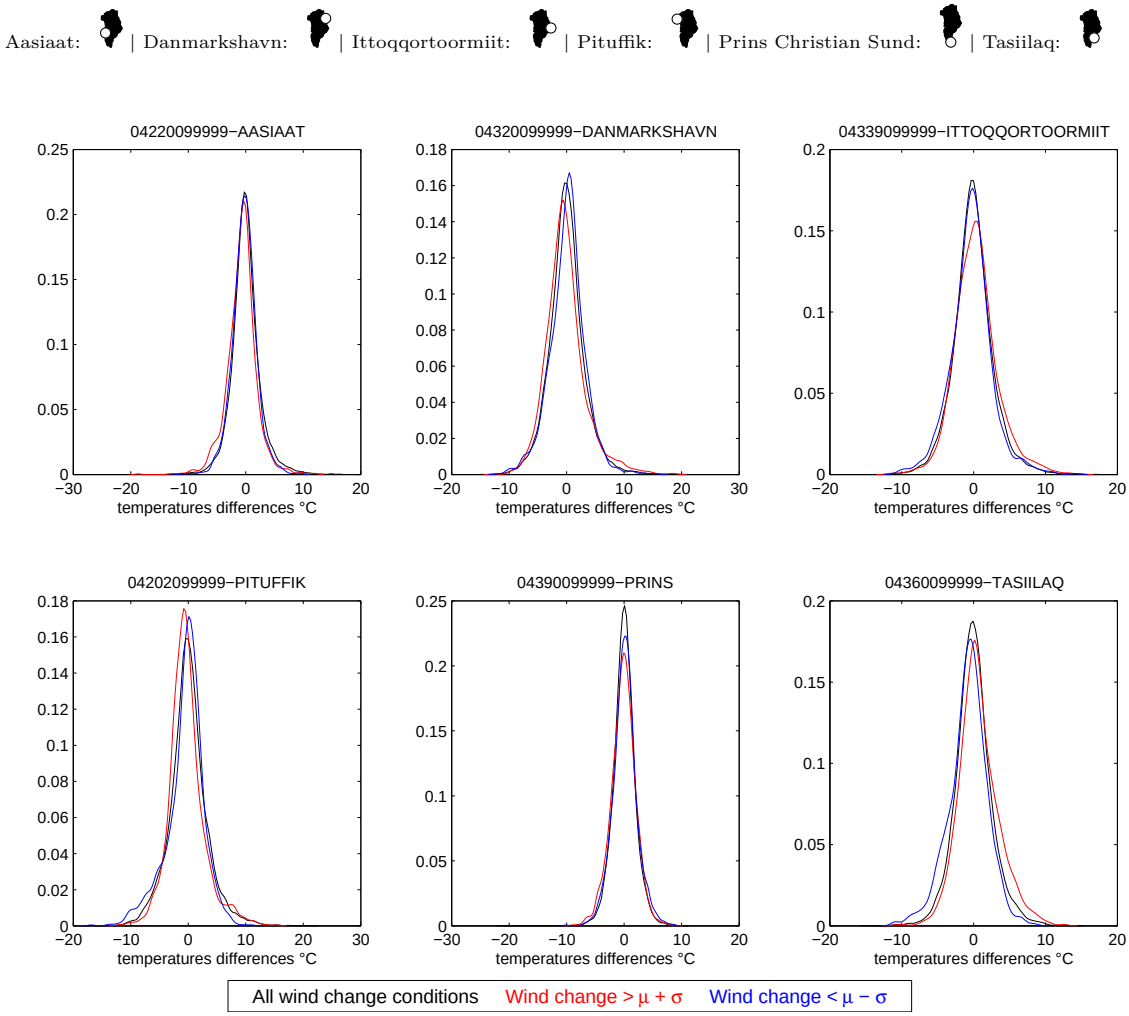

Figure 17: Kernel smoothing probability density estimates for temperature differences during extreme wind type changes

[17] A modified version of Student's t-distribution parametrized by a location parameter μ, a positive scale parameter σ and a shape parameter ν.

Aasiaat:  | Danmarkshavn:  | Ittoqqortoormiit:  | Pituffik:  | Prins Christian Sund:  | Tasiilaq: 

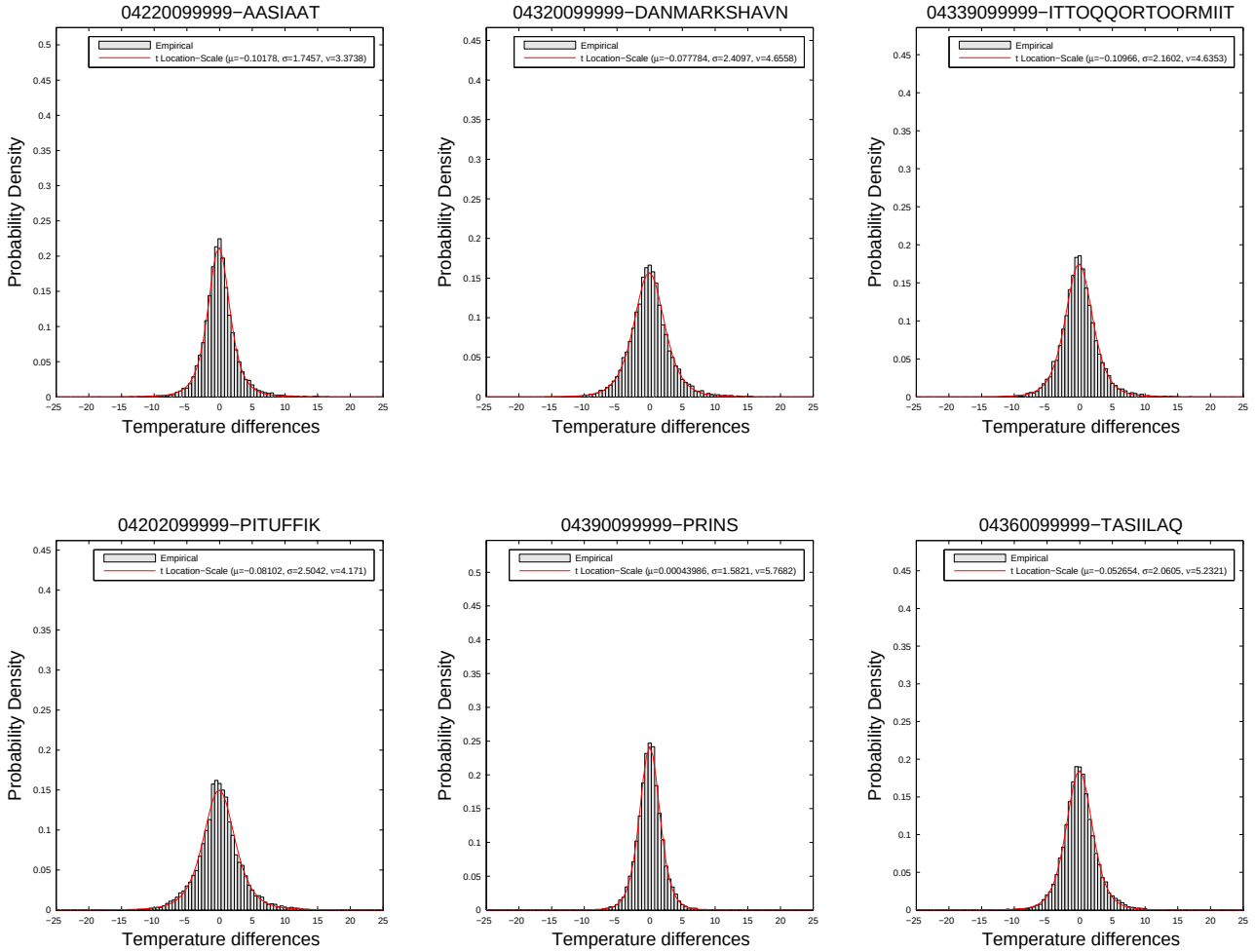

Figure 18: Consecutive daily temperature differences (empirical probability and model distributions)

Aasiaat:  | Danmarkshavn:  | Ittoqqortoormiit:  | Pituffik:  | Prins Christian Sund:  | Tasiilaq: 

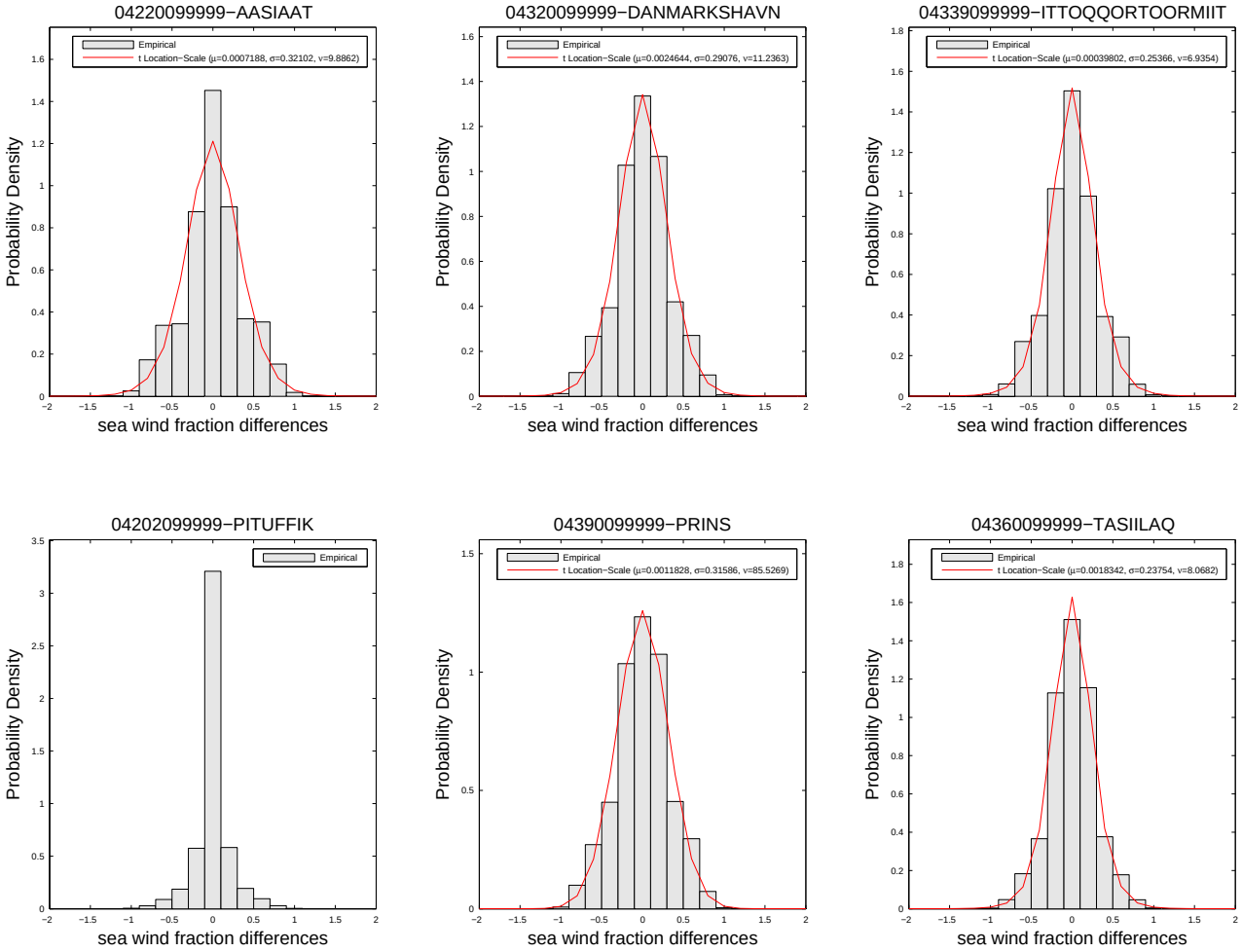

Figure 19: Consecutive daily sea wind fraction differences (empirical probability and model distributions)

3.1.7 Cross-correlation between daily sea surface and atmospheric temperatures

The cross-correlations between median sea surface temperatures (within a 300 km radius) and atmospheric temperatures were computed for a lag range of ±80 days (see fig. 20). In all cases the air temperature time series leads the sea surface temperature time series. All of the stations except Prins Christian Sund and Tasiilaq show a lag of about two weeks. Prins Christian Sund has a lag of about three weeks and Tasiilaq about six weeks. The large lag associated with Tasiilaq could be caused by the nearby East Greenland Current (a cold and low salinity current from the North). In summer atmospheric temperature may be relatively high at this location, but sea surface temperature remains cold longer because of the current. Altough this result gives us a hint on how the two time series are linked, the construction features with such large time lags would be unpractical due to missing observation days.

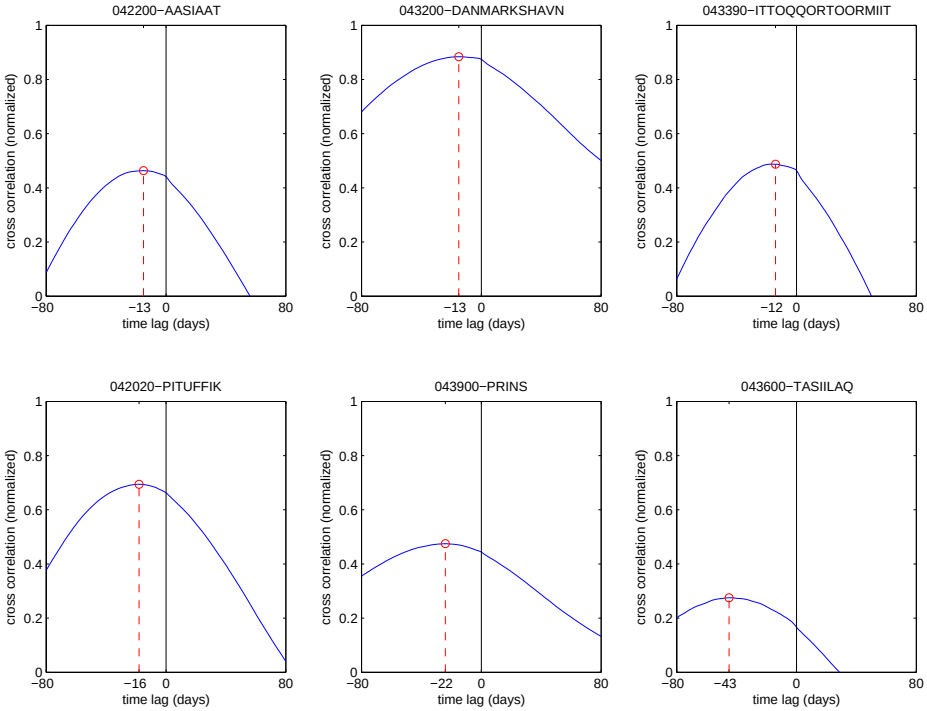

Figure 20: Cross correlations between daily sea surface temperature (lag) and air temperature (lead)

Aasiaat: | Danmarkshavn: | Ittoqqortoormiit: | Pituffik: | Prins Christian Sund: | Tasiilaq:

4 Modelling coastal temperatures with SVR

This section begins with an introduction to support vector regression stripped of its mathematical subtleties[18]. It is continued with the presentation of the SVR prediction results and their discussion.

4.1 Support vector regression

4.1.1 Theory

The classical form of linear regression involves adjusting a linear model to observed data by minimizing the sum of squared errors (i.e residuals between the observations and the model). Support vector regression (SVR) is an improvement over basic linear regression because it can be extended to determine non-linear models for high dimensional data. The term "support vector" refers to the fact that the model is determined using only a subset of the data points, those that bring the most information, called support vectors. The basic idea behind SVR is well explained by Hamel 2011:

> *Suppose that we have a regression problem where all the observations of the regression training set fit into a (hyper-) tube of width 2ϵ with $\epsilon > 0$. We can interpret this hypertube as a regression model by imagining that there is a hyperplane positioned right in the center of the tube that models the observations. Now typically there are many different ways to position the hypertube of width 2ϵ and still have all the training observations contained within the tube. However, there exists an optimal hypertube alignement such that as many observations are pushed as close to the outer boundaries of the hypertube as possible. In other words, the optimal hypertube alignement is achieved when the distances of the observations from the center hyperplane are maximized.*

Observations that are not contained in the 2ϵ hypertube are taken into account by introducing two slack variables ξ_i (for observations above the hypertube) and ξ_i' (for observations below the hypertube). These extra variables take values according to a loss function L which penalizes observations located outside the hypertube (see fig. 21).

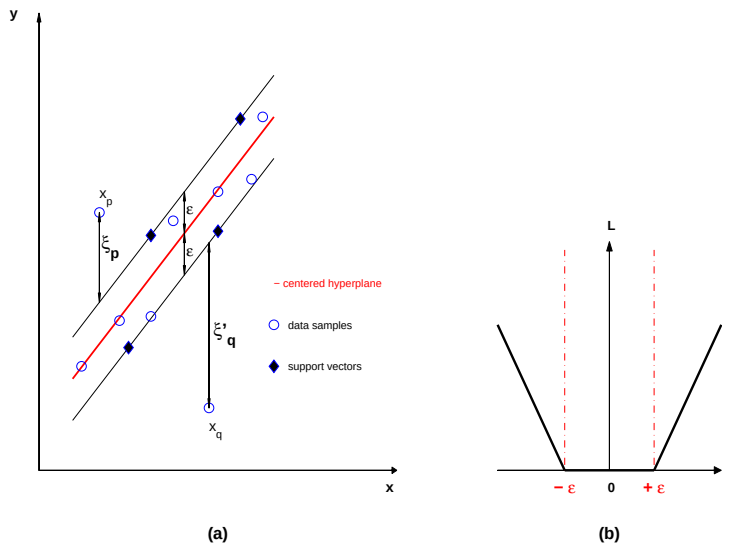

Figure 21: (a) The center hyperplane aligned inside the 2ϵ hypertube. The points on the edge of the hypertube are called support vectors. (b) The slack variables are equal to zero inside the 2ϵ tube and increase for points outside according to the loss function L (here a linear ϵ insensitive loss function)

[18] Readers interested in detailed demonstrations may consult Hamel 2011, Kanevski, Pozdnoukhov, Timonin 2009, Cherkassky, Mulier 2007, Cornuéjols, Miclet 2010 among others.

Mathematically, this is a constrained optimization problem where the objective is to minimize the values of the slack variables while maximizing the distance (also called margin) between the hyperplane and the observations within the hypertube. Maximizing the margin is equivalent to minimizing $\|\vec{w}^2\|$.

Specifically, given a set of m observation pairs (called examples) $D = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), ..., (\vec{x}_m, y_m)\}$ with dimension of $\vec{x}$ being the number of features, the objective is:

$$min \left[\underbrace{\frac{1}{2}\|\vec{w}^2\|}_{\text{inverse margin}} + \overbrace{C}^{\text{regularization parameter}} \cdot \underbrace{\sum_{i=1}^{m}(\xi_i + \xi_i')}_{\text{sum of slack variables}} \right] \tag{1}$$

subject to the constraints:

$$\begin{cases} y_i - \vec{w}^\top \cdot \vec{x}_i - b \leq \epsilon + \xi_i \\ (\vec{w}^\top \cdot \vec{x}_i + b) - y_i \leq \epsilon + \xi_i' \\ \xi_i, \xi_i' \geq 0, i = 1, ..., m \end{cases}$$

Notice that a parameter C has been included in equation (1). This parameter (called regularization) gives control on the balance between the margin maximization and the minimization of slack variables. A small value of C will limit the influence of slack variables and will favor a simple model (with large errors). On the other hand, a large value of C will produce a more complex model that fits observations better (smaller error). Altough they may perform well on a given set of observations, complex models bear the risk of not being well suited when additional observations are added to the problem (i.e overcomplexity leads to poor generalization). In the end, the best model is a trade-off between complexity and error. The influence of C on the predictive model is illustrated in figure 22 for a non-linear regression example (application of SVR to non-linear problems will be discussed shortly).

The minimization problem (1) with inequality constraints can be solved using the augmented method of Lagrange multipliers (which is not detailed here).

This method implies maximizing the following expression[19]:

$$max \left[\sum_{i=1}^{m} y_i(\alpha_i - \alpha_i') - \epsilon \sum_{i=1}^{m}(\alpha_i - \alpha_i') - \frac{1}{2}\sum_{i,j=1}^{m}(\alpha_i - \alpha_i')(\alpha_j - \alpha_j')(\vec{x}_i \cdot \vec{x}_j) \right] \tag{2}$$

subject to the constraints:

$$\begin{cases} \sum_{i=1}^{m} \alpha_i' = \sum_{i=1}^{m} \alpha_i \\ 0 \leq \alpha_i \leq C \\ 0 \leq \alpha_i' \leq C \end{cases}$$

This a quadratic optimization problem which can be solved numerically using different methods (which are not detailed here). In the end, the optimal regression model $f(x)$ is:

$$f(x) = \sum_{i=1}^{m}(\alpha_i - \alpha_i') \underbrace{(\vec{x}_i \cdot \vec{x})}_{\text{dot product}} + b \tag{3}$$

The optimized weights vector $\vec{w}^*$ attributed to the features provides the ability to rank features according to their relative importance (i.e. "How informative they were") in determining the outcome of the prediction. It is is given by:

$$\vec{w}^* = \sum_{i=1}^{m}(\alpha_i^* - \alpha_i'^*)\vec{x}_i \tag{4}$$

[19] In Lagrangian optimization this expression is called the dual form and α_i, α_i' are the Lagrangian coefficients.

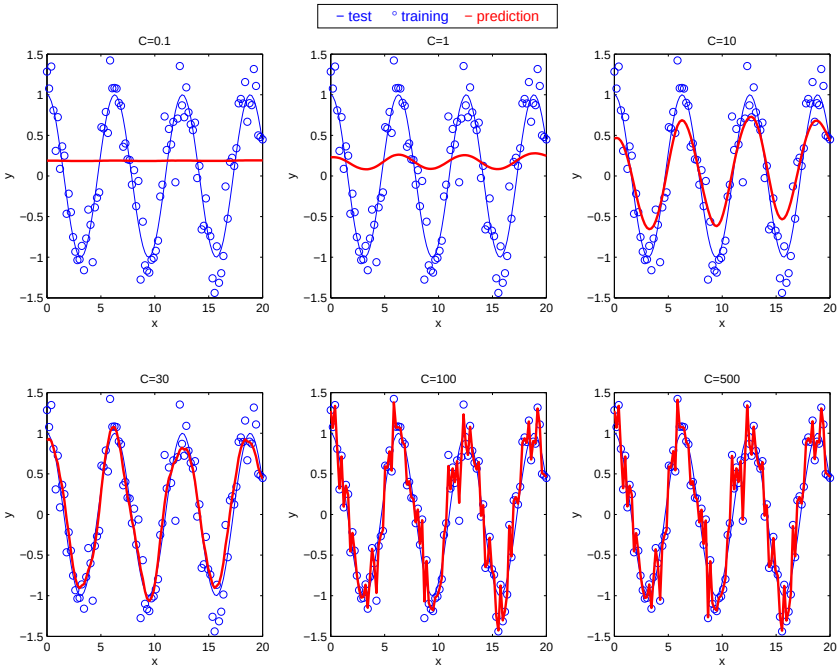

Figure 22: The effect of the regularization parameter C on model complexity. Higher values of C produce better fits but can lead to overfitting (thus bad generalization to new data). The influence of C can be seen in terms of bias and variance. A large C model has lower bias and high variance. A small C model has higher bias and low variance. Here, an overly high value of C produces a model that does no reflect the underlying sinusoidal structure.

Now that the basic concepts are in place for the linear case, let's see how the algorithm presented above can be adapted to non-linear regression. The extension to non-linear problems is made possible by exploiting a proof made by mathematician James Mercer (1883-1932) stating that some positive definite functions (called kernels, from German *Kern*, nucleus, core) can be expressed as a dot product in a high dimensional space. This is called the Mercer theorem. A practical implication of this theorem in SVR is that the observations can be mapped to a higher dimensional space where they have a linear structure, without explicitly computing the mapping. Indeed, looking at equation (3) it can be seen that it only depends on the dot product. By simply substituting the dot product in equation (3) by a kernel function, a new regression model (5) for non-linear cases is obtained:

$$f(x) = \sum_{i=1}^{m} (\alpha_i - \alpha_i') \underbrace{K(\vec{x}_i, \vec{x})}_{\text{kernel function}} + b \tag{5}$$

This substitution is called the kernel trick and it is at the heart of SVR. In machine learning terminology, the original space is named "input space" and the new space implicitly mapped using the kernel trick is named "feature space". Focusing a bit more on kernels, not only do they define the dot product of the feature space, but they can also be seen as a measure of similarity between two vectors. Indeed, they provide us with a unique scalar value describing the similarity between two vectors of same length. In the input space the kernel function K is simply the linear dot product (which is also called linear kernel[20], even if technically speaking it isn't really a kernel). It is important to note that not all functions can be used as kernels. According to Mercer, only positive definite functions are suitable. Table 6 provides the list of kernels that were used for atmospheric temperature predictions in this study. Generally speaking linear and Gaussian kernels are well suited for most type of features. Altough some features, such as the wind fraction from a specific direction, require more specialized kernels.

[20] When using a linear kernel the input space is the same as the feature space.

Table 6: The kernel functions used for temperature prediction. These functions are a measure of similarity between two vectors of same length $\vec{a}$ and $\vec{b}$ with $\vec{a} = [a_1\ a_2\ ...\ a_m]$, $\vec{b} = [b_1\ b_2\ ...\ b_m]$

Name	**Function**	**Free parameters**
Linear	$K(\vec{a}, \vec{b}) = \vec{a} \cdot \vec{b}$	none
Gaussian	$K(\vec{a}, \vec{b}) = \exp(\frac{\|\vec{a}-\vec{b}\|^2}{2\sigma^2})$	$\sigma\ (> 0)$
Histogram intersection	$K(\vec{a}, \vec{b}) = \sum_{i=1}^{m} min\{a_i, b_i\}$	none

Note that when the kernel is not linear (for example Gaussien) the optimized weight vector (see equation (4)) cannot be determined since it is hidden in the feature space. The use of linear and Gaussian kernels is relatively straightforward. On the other hand, the histogram intersection kernel is more specialized and deserves some extra attention. The basic idea of the histogram intersection kernel is to obtain a score of similarity between two histograms. To illustrate this, an example using wind direction is provided in table 7). The use of such a kernel was motivated by the fact that the wind direction features contain many null values when considered by separate 10° azimuths. Applying a histogram intersection to them means they are not considered independant any more and all the wind direction information is represented at the same time.
One final remark about the support vector algorithm discussed up to now is that it only allows the choice of one kernel for all features. This can be an important limitation when considering features whose similarity is not well described by a single kernel (for example not all features used here would be well described by a histogram intersection kernel). This problem has been overcome by a more sophisticated family of algorithms known a multiple kernel learning (MKL), briefly presented in the following subsection.

Table 7: Histogram intersection kernel applied to wind direction. The fraction of time when the wind was blowing from a specific azimuth ###° is given for three days ($\vec{w}_{day1}, \vec{w}_{day2}, \vec{w}_{day3}$). From the value of the kernel, it can be seen that wind directions were very similar on day 1 and 2, but totally different on day 2 and 3.

Wind azimuth	$\vec{w}_{day1}$	$\vec{w}_{day2}$	$\vec{w}_{day3}$	$min(\vec{w}_{day1}, \vec{w}_{day2})$	$min(\vec{w}_{day2}, \vec{w}_{day3})$
010°	0.25	0.15	0	0.15	0
020°	0.1	0.15	0	0.1	0
030°	0	0	0	0	0
040°	0	0	0	0	0
050°	0	0	0	0	0
060°	0	0	0	0	0
070°	0	0	0	0	0
080°	0	0	0.4	0	0
090°	0	0	0.6	0	0
⋮	⋮	⋮	⋮	⋮	⋮
350°	0	0	0	0	0
360°	0.65	0.7	0	0.65	0
				$\underbrace{\quad\quad}_{\sum\ = 0.9}$	$\underbrace{\quad\quad}_{\sum\ = 0}$
				$\Rightarrow K(\vec{w}_{day1}, \vec{w}_{day2}) = 0.9$	$\Rightarrow K(\vec{w}_{day2}, \vec{w}_{day3}) = 0$

4.1.2 Combination of kernels

In order to improve the performance of SVR and combine features with different types of kernel, it is possible to use multiple kernel learning (MKL). Put simply, instead of using a single kernel for all features, MKL allows using linear combinations of kernels:

$$K(\vec{x}_i, \vec{x}) = \sum_{j=1}^{m} d_j K_j(\vec{x}_i, \vec{x}) \text{ with } d_j \geq 0 \text{ and } \sum_{j=1}^{m} d_j = 1 \quad (6)$$

MKL optimizes the kernel weights d_j and the Lagrangian coefficients α_i, α_i' simultaneously (Rakotomamonjy et al. 2008, Canu et al. 2005, Foresti et al. 2009). The weights in (6) also represent the relative importance of different features (they sum up to 1). That is, the features with the largest weights are likely to be the most important drivers in the modelled phenomenon. Note that the $\vec{d}$ weight vector is the MKL equivalent of the linear SVR $\vec{w}$ weight vector.

This method is also useful because it allows the use of different kernels for groups of features with each kernel having its own list of parameters. In this way each feature or group of feature can be attributed an adapted kernel. For atmospheric temperature predictions, a combination of Gaussian and histogram intersection kernels was employed. Gaussian kernels with free parameters $\sigma = [0.001\ 0.05\ 0.1\ 0.2\ 0.4\ 0.55\ 0.6\ 0.8\ 0.9\ 1]$ were attributed to all features except those related to wind direction. For the latter, histogram intersection kernels were employed.

For the Gaussian kernels, MKL replicates σ-feature couples and selects the most informative couples for the prediction. For example with the σ values provided above, MKL will replicate each Gaussian kernel feature i 10 times, each time associating the same feature to a different value of σ:

$[F_i - \sigma_1\ , F_i - \sigma_2\ , F_i - \sigma_3\ , F_i - \sigma_4\ , F_i - \sigma_5\ , F_i - \sigma_6\ , F_i - \sigma_7\ , F_i - \sigma_8\ , F_i - \sigma_9\ , F_i - \sigma_{10}]$

Figure 23 illustrates how MKL proceeds to set non-important features to null weights, keeping only the most informative and thus providing a ranking.

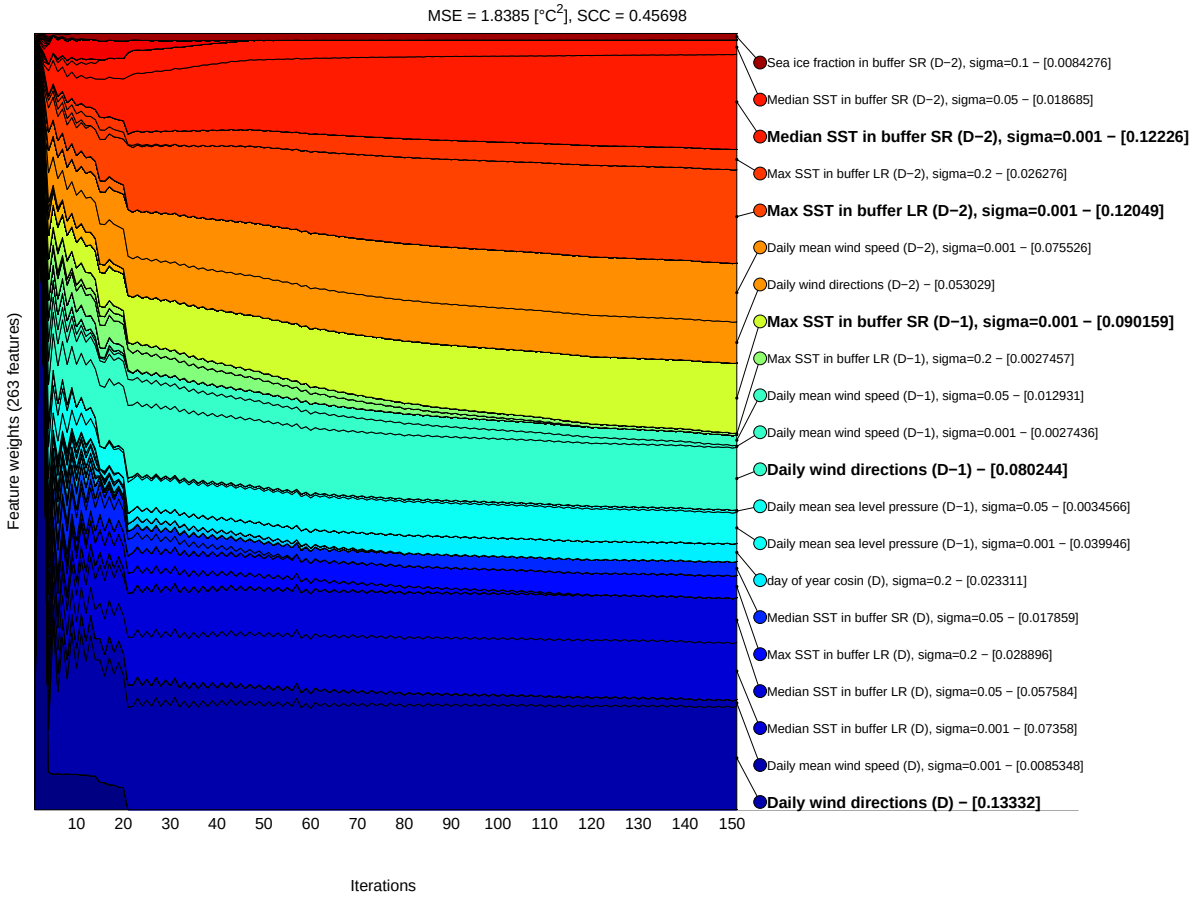

Figure 23: Example of feature weight evolution for station Aasiaat in August. The daily wind directions weights are computed using a histogram intersection kernel, all the other weights are computed with Gaussian kernels with the free parameters σ listed previously. Initially, all features are set to have the same weight. MKL then selects the most important. The values of the non-zero relative weights are indicated in brackets and the most suitable σ values are reported.

4.1.3 Feature set splitting, scaling and parameter optimization

Once all the features have been constructed, several additional steps are required to prepare them in a form suitable for the machine learning phase (i.e SVR). These preparation steps are feature set splitting and normalization.
This is followed by the optimization of the regularization and kernel parameters (when they are present).

Feature set splitting

The feature construction step presented above has left us with monthly features matrixes containing m training examples and n features (the first column being the label or target value for the prediction):

$$\begin{array}{c} \\ example\ 1 \\ example\ 2 \\ example\ 3 \\ \vdots \\ example\ m \end{array} \begin{array}{c} \begin{array}{ccccccc} Label & F_1 & F_2 & F_3 & F_4 & \cdots & F_n \end{array} \\ \begin{pmatrix} 0.4 & 1013.2 & 12 & 98.5 & 1 & \cdots & 5 \\ 0.6 & 1013.2 & 10 & 98.4 & 1 & \cdots & 4.6 \\ -2 & 1013.4 & 5 & 98.4 & 0 & \cdots & 4 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ -1.4 & 1013.5 & 5 & 98.3 & 1 & \cdots & 4.2 \end{pmatrix} \end{array}$$

When applying a machine learning algorithm, it is important to reorder the examples randomly and then divide the feature set into three subsets:

- A cross validation set used to optimize the free parameters.
- A training set used to train the algorithm with the optimal free parameters
- A test set to determine the algorithm's performance

The relative size of these datasets may vary. In this case, it was decided to split the datasets according to the following scheme:

$$\left.\begin{array}{ll} N_{crossval} = N_{train} = 0.25 \cdot m & \text{if } m \leq 500 \\ N_{crossval} = N_{train} = 125 & \text{if } m > 500 \end{array}\right\} N_{test} = m - N_{crossval} - N_{train}$$

Where
$N_{crossval}$ is the number of examples in the cross validation set
N_{train} is the number of examples in the training set
N_{test} is the number of examples in the test set

Without this splitting, the prediction error might be overly optimistic. Indeed, the model would be perfectly tuned to the observations, but would likely not generalize well to new observations.

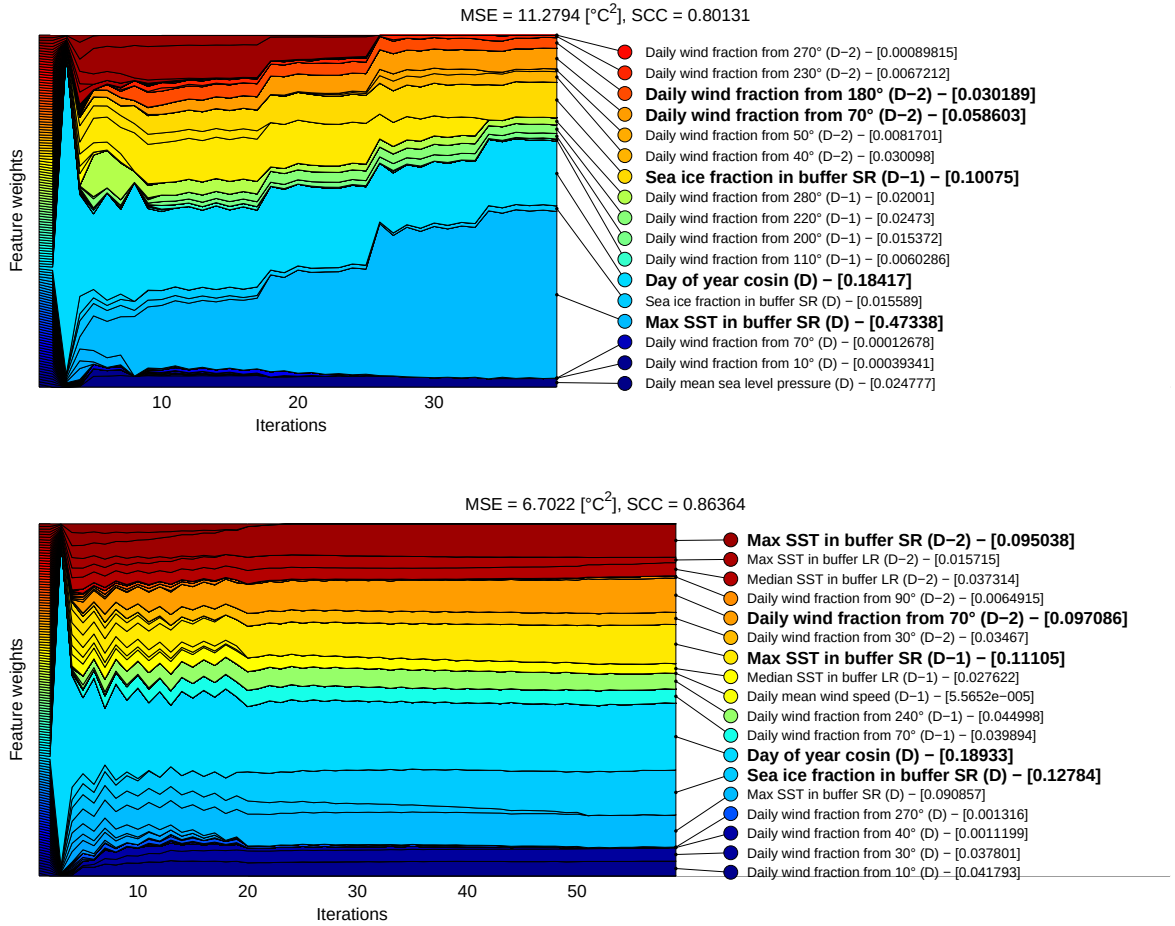

Figure 24: The effect of randomly reordering examples among the feature sets (using MKL with a single parameter Gaussian kernel). The relative weights of features as a function of the iteration number (the 5 most influent features are in bold) are represented. Initially, all features are set to have the same weight. As the number of iterations increases, most feature weights become null and a few emerge as dominant. Notice how the outcome varies between the top and bottom plots. Both of the predictions are based on the same feature dataset. However, since the training, cross-validation and test subsets are chosen randomly in the feature dataset, the initial conditions are different for each run. This means the results of feature ranking have to be based on the average or median of relative weights obtained for each run.

Scaling (normalization)

In order to facilitate the optimization, it is required that all features (and labels) be scaled. Since features with different measurement units are employed, the scaling step also helps constrict the features to a similar value range.
Scaling can simply be done using the so-called "*reduce to standard scores*" formula:

$$F_{scaled} = \frac{F - mean(F)}{std(F)}$$

Parameter optimization

When using a linear kernel, the optimal regularization parameter C can be determined through an iterative search. The SVR prediction is done using a range of different C parameters. The optimal C is the one which provides the smallest mean squarred error with the cross-validation set (see fig. 25).
When using a Gaussian kernel an extra parameter σ is introduced. In this case both C and σ have to be optimized and the easiest (and most widely used approach) is to perform a grid search. That is running the SVR algorithm with all combinations of C and σ and choosing the combination which yields the smallest mean squarred error.

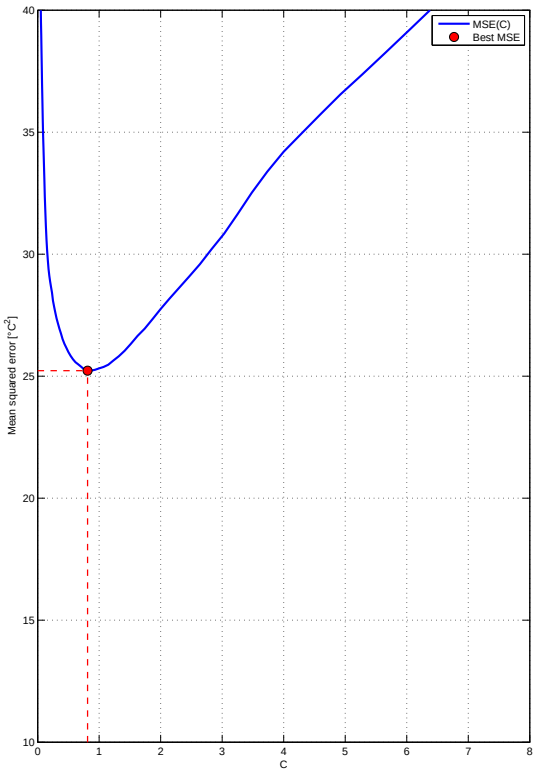

Figure 25: Example of an iterative search to find the optimal C parameter. This plot was obtained when performing linear SVR on station Aasiaat for February.

4.1.4 Performance and ranking metrics

The performance of SVR predictions was evaluated using the mean squarred error (MSE) obtained with the test set:

$$MSE[^{\circ}C^2] = \frac{\sum_{k=1}^{n}(T_{pred,k} - T_{obs,k})^2}{n} \tag{7}$$

with
T_{pred} : the predicted temperature [°C]
T_{obs} : the observed temperature [°C]
n : the number of samples in the test set

For presentation purposes, the root mean squarred error (RMSE) is employed. This performance metric is simply the square root of the MSE. Since several runs of SVR were applied to each set, the final performance measure was the median RMSE over all runs[21] (see fig. 24 for an explanation).

It was mentionned previously (see fig. 24) that the feature ranking changes with each SVR run, for this reason it is necessary to work with the median of weights obtained at each run. This results in a monthly median ranking (see fig. 26).

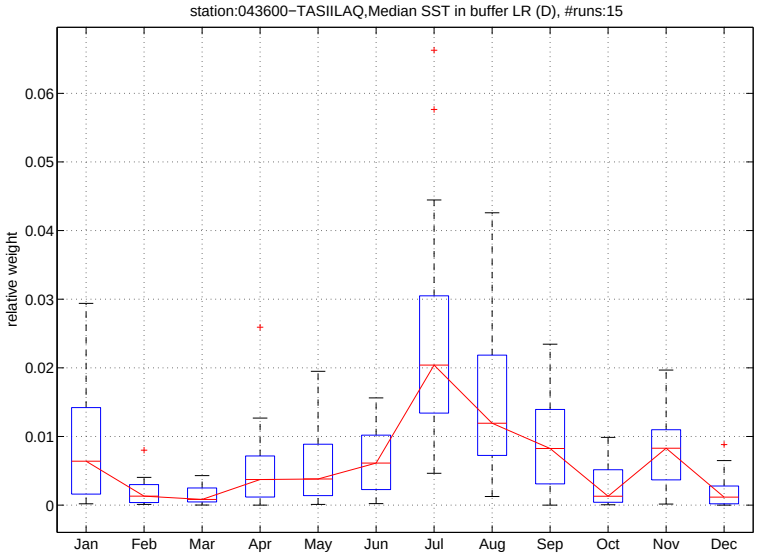

Figure 26: Median weight of the long range daily sea surface temperature feature as a function of the month (station Tasiilaq, 15 runs). This is applied to compensate the effect of the random sampling used when learning with the features sets.

[21] N.B. It is important not to confuse the terms "runs" and "iterations". For each run the algorithm undergoes several iterations until it converges and produces a set of feature weights.

4.1.5 Software

Two Matlab implementations of support vector machines were employed:

- LibSVM[22] (see Chang, Lin 2011) was used for the linear SVR predictions.
- SimpleMKL (part of Multiple Kernel Learning toolbox for Matlab)[23] (see Rakotomamonjy et al. 2008, Canu et al. 2005) was used for the multiple kernel (Gaussian and histogram intersection) SVR predictions.

[22] Cf. `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`
[23] Cf. `http://asi.insa-rouen.fr/enseignants/~arakotom/code/mklindex.html`

4.2 Predictions using a linear kernel

The following pages present the performance of linear SVR in terms of monthly RMSE. A visualization of the difference between observed and predicted temperatures is presented in an effort to determine the source of error. Finally, the monthly feature ranking is commented.

4.2.1 Performance

The SVR performances obtained using a linear kernel are presented in figure 27. The results are the median of the RMSE over 15 SVR runs. The smallest monthly median RMSE is 1.282 $°C$ (in August at Asiaat) and the largest is 5.632 $°C$ (in January at Danmarkshavn).

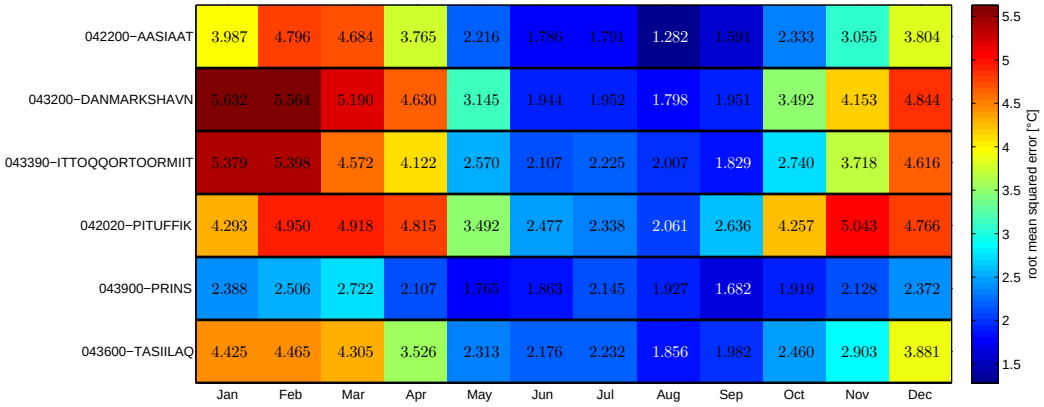

Figure 27: SVR prediction performance as a function of the station and the month. The performance metric is the median of the root mean squared error (RMSE) over 15 runs. The best performances of each station are highlighted in white. Note that smaller RMSE can be obtained as shown in appendix B.

The overall performances (yearly statistics) were computed from monthly results (see table 8). Clearly, summer months and stations located in the South yield the best predictions. The most stable performance is obtained at Prins Christian Sund and the least stable at Danmarkshavn . Generally speaking, predictions for Southern stations have lower variability. Taking a closer look at differences between observed and predicted temperatures for an example station (see fig. 28 and 29), it can be seen that most of the error comes from the bad modelling of extreme temperatures (cold and warm ends). One possible explanation is that since these are "extreme" values, they are less likely to be present in the training set and so the SVR is also less likely to be well trained in predicting them. Notice that SVR is unable to reproduce to noisy patterns of day to day temperature variation (it is not an exact interpolator).

Table 8: Overall yearly performance of the SVR using a linear Kernel (ranked from best to worst)

Station name	**RMSE yearly mean** $[°C]$	**RMSE yearly std** $[°C]$
Prins Christian Sund	2.127	0.317
Aasiaat / Egedesmind	2.924	1.248
Tasiilaq / Ammassali	3.044	1.013
Ittoqqortoormiit	3.440	1.347
Danmarkshavn	3.691	1.507
Pituffik (Thule Airbase)	3.837	1.163

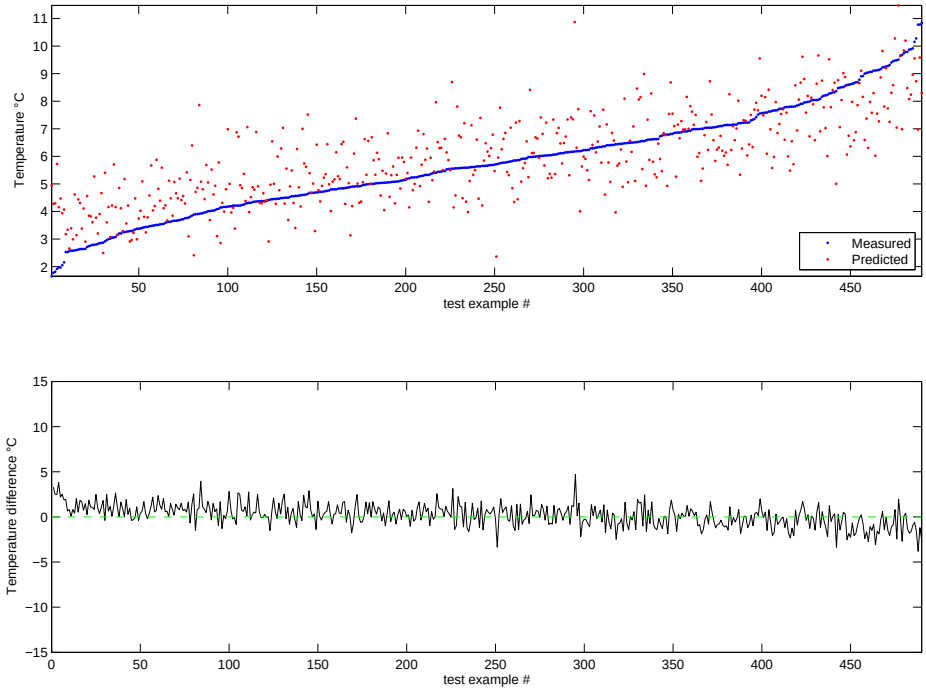

Figure 28: Difference between predicted and observed daily temperatures at Aasiaat in August. Top: Sorted measured temperatures and corresponding predictions for the test examples. Bottom: Sorted temperature differences (predicted minus measured) for the test examples.

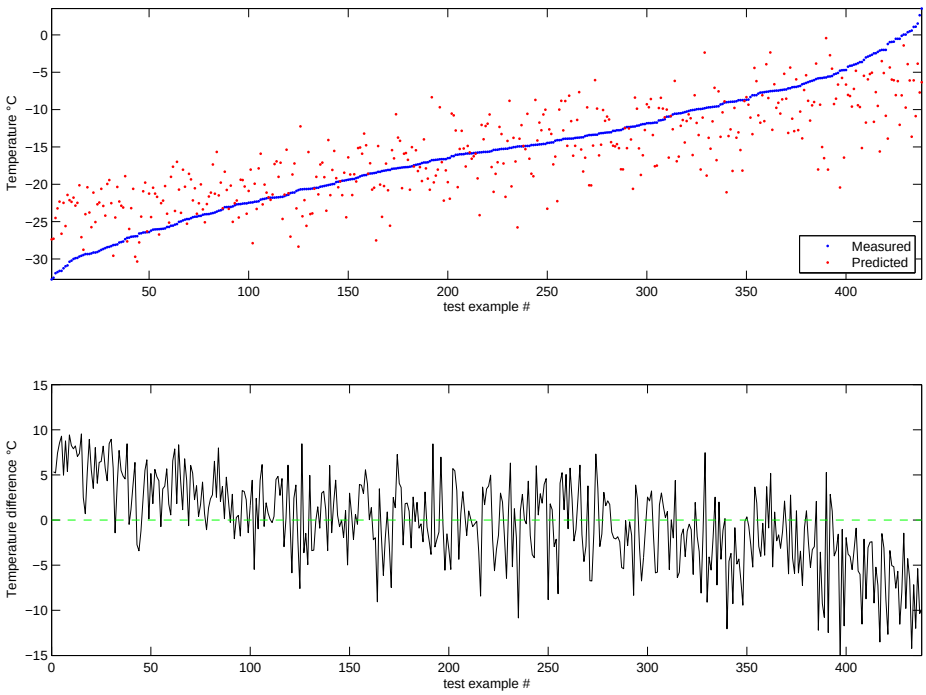

Figure 29: Difference between predicted and observed daily temperatures at Aasiaat in February. Top: Sorted measured temperatures and corresponding predictions for the test examples. Bottom: Sorted temperature differences (predicted minus measured) for the test examples.

4.2.2 Feature ranking

As discussed previously, representing the spatiotemporal change of feature ranking is difficult. For the sake of readability, only the five most influent features for each month were considered. Tables 9 and 10 show the monthly ranking and highlight feature types with different colors. However, these tables do not give any information about weight repartition among the features. Indeed, in some cases all of the weight might be attributed to a small group of features and in others cases the weight might be distributed among many features. In other words, the ranking can sometimes be based on very small weight differences (usually not the case here).

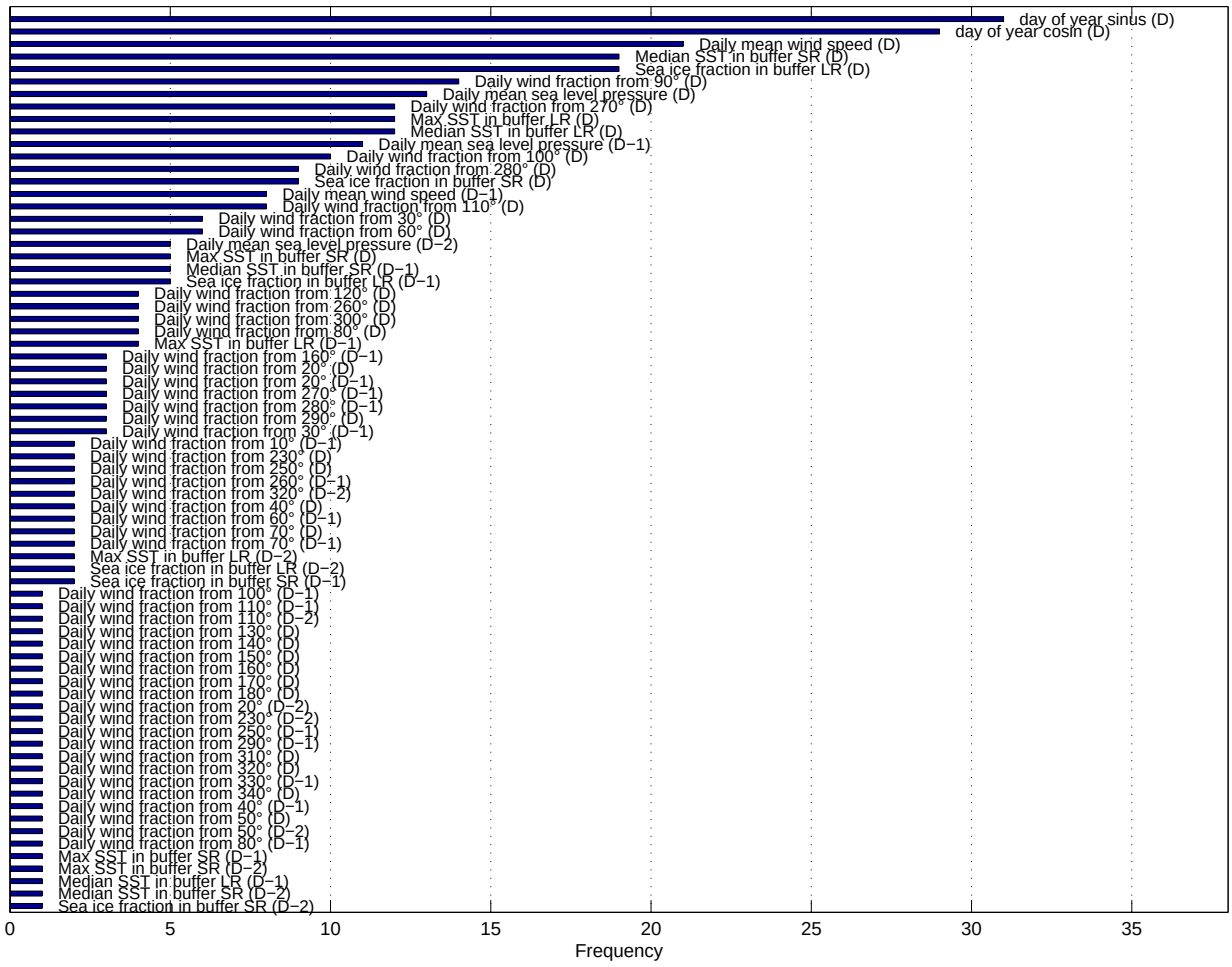

Figure 30: Frequency of feature appearance in the top 5 ranks when using a linear kernel

Looking at features that appear at least ten times in the top 5 ranks of figure 30, we see that the most dominant features are the day of the year, wind speed (D), short range median sea surface temperatures (D) and long range sea ice fraction (D). Focusing on the most frequent influent wind directions features, it can be noted that two directions prevail: wind from the East (090°, 100°) and wind from the West (270°, 280°). The high frequency of these directional features is particularly visible for stations Pituffik and Prins Christian Sund .

Table 9: The top 5 most influent features as a function of station and month. The color scheme indicates the feature type: wind, ice, water, sea level pressure and time

	Aasiaat / Egedesmind	Danmarkshavn	Ittoqqortoormiit
Jan	Sea ice fraction in buffer SR (D) Sea ice fraction in buffer LR (D) Daily mean wind speed (D) Max SST in buffer SR (D) Daily wind fraction from 60° (D-1)	Daily mean wind speed (D) Sea ice fraction in buffer LR (D) Daily mean sea level pressure (D-1) Daily mean sea level pressure (D) Daily mean wind speed (D-1)	Daily mean sea level pressure (D) Daily mean sea level pressure (D-1) Daily mean wind speed (D) Median SST in buffer SR (D) Sea ice fraction in buffer SR (D)
Feb	Median SST in buffer SR (D) Sea ice fraction in buffer SR (D) Median SST in buffer LR (D) Sea ice fraction in buffer LR (D) Daily mean wind speed (D)	Daily mean wind speed (D) Daily mean wind speed (D-1) Sea ice fraction in buffer LR (D) Daily mean sea level pressure (D) Daily mean sea level pressure (D-1)	Daily mean wind speed (D) Daily mean sea level pressure (D) Daily mean sea level pressure (D-1) Max SST in buffer SR (D) Sea ice fraction in buffer LR (D)
Mar	Sea ice fraction in buffer SR (D) Sea ice fraction in buffer LR (D) Median SST in buffer SR (D) Median SST in buffer SR (D-1) Median SST in buffer SR (D-2)	Daily mean wind speed (D) Daily mean wind speed (D-1) Sea ice fraction in buffer LR (D) Median SST in buffer LR (D) day of year cosin (D)	Daily mean wind speed (D) Daily mean sea level pressure (D) Median SST in buffer SR (D) Max SST in buffer SR (D-2) Median SST in buffer SR (D-1)
Apr	day of year cosin (D) Median SST in buffer SR (D) Sea ice fraction in buffer LR (D) day of year sinus (D) Sea ice fraction in buffer SR (D)	day of year cosin (D) day of year sinus (D) Sea ice fraction in buffer LR (D) Daily mean wind speed (D) Sea ice fraction in buffer LR (D-1)	day of year cosin (D) day of year sinus (D) Daily mean wind speed (D) Daily mean sea level pressure (D-1) Daily wind fraction from 330° (D-1)
May	Median SST in buffer SR (D) Daily mean sea level pressure (D-2) Daily wind fraction from 60° (D) Sea ice fraction in buffer LR (D) Daily mean sea level pressure (D-1)	day of year sinus (D) day of year cosin (D) Max SST in buffer SR (D) Daily mean wind speed (D) Median SST in buffer LR (D)	day of year cosin (D) day of year sinus (D) Median SST in buffer LR (D) Median SST in buffer SR (D) Daily wind fraction from 40° (D)
Jun	Median SST in buffer SR (D) Daily wind fraction from 230° (D) Daily wind fraction from 250° (D) Sea ice fraction in buffer SR (D-1) Median SST in buffer SR (D-1)	day of year cosin (D) Daily wind fraction from 300° (D) day of year sinus (D) Daily wind fraction from 290° (D) Daily wind fraction from 110° (D)	day of year sinus (D) day of year cosin (D) Sea ice fraction in buffer SR (D) Daily mean wind speed (D-1) Median SST in buffer SR (D)
Jul	Daily mean sea level pressure (D-1) Median SST in buffer LR (D) Daily wind fraction from 30° (D) Daily wind fraction from 270° (D-1) Daily mean sea level pressure (D-2)	Daily wind fraction from 280° (D) Daily wind fraction from 300° (D) Daily wind fraction from 290° (D) Daily wind fraction from 270° (D) Daily wind fraction from 120° (D)	Median SST in buffer LR (D) Median SST in buffer SR (D) Median SST in buffer SR (D-1) Daily mean sea level pressure (D-2) Daily wind fraction from 100° (D)
Aug	day of year cosin (D) day of year sinus (D) Daily wind fraction from 70° (D-1) Daily wind fraction from 60° (D) Median SST in buffer LR (D)	day of year cosin (D) day of year sinus (D) Daily wind fraction from 100° (D) Median SST in buffer SR (D) Daily wind fraction from 300° (D)	day of year cosin (D) day of year sinus (D) Median SST in buffer SR (D) Daily wind fraction from 160° (D-1) Median SST in buffer LR (D)
Sep	day of year cosin (D) day of year sinus (D) Max SST in buffer LR (D) Daily wind fraction from 60° (D-1) Daily wind fraction from 50° (D-2)	day of year sinus (D) day of year cosin (D) Daily wind fraction from 320° (D-2) Max SST in buffer LR (D) Max SST in buffer LR (D-1)	day of year cosin (D) day of year sinus (D) Daily wind fraction from 30° (D) Median SST in buffer SR (D) Max SST in buffer LR (D)
Oct	day of year sinus (D) Daily mean sea level pressure (D-2) Daily wind fraction from 320° (D-2) Daily mean sea level pressure (D) day of year cosin (D)	day of year sinus (D) Daily mean wind speed (D) day of year cosin (D) Sea ice fraction in buffer LR (D) Daily mean wind speed (D-1)	Daily wind fraction from 20° (D) Daily wind fraction from 340° (D) day of year sinus (D) Daily wind fraction from 280° (D) Median SST in buffer SR (D)
Nov	Sea ice fraction in buffer LR (D) Daily wind fraction from 130° (D) Daily wind fraction from 160° (D) Sea ice fraction in buffer LR (D-1) Daily wind fraction from 160° (D-1)	Daily mean wind speed (D) Sea ice fraction in buffer LR (D) Daily mean wind speed (D-1) Sea ice fraction in buffer SR (D-2) Daily mean sea level pressure (D-1)	Daily mean wind speed (D) Daily wind fraction from 20° (D-1) Median SST in buffer SR (D) Daily mean sea level pressure (D) Daily wind fraction from 20° (D)
Dec	Sea ice fraction in buffer LR (D) Sea ice fraction in buffer SR (D) Sea ice fraction in buffer LR (D-1) Sea ice fraction in buffer SR (D-1) Daily wind fraction from 40° (D-1)	Daily mean wind speed (D) Daily mean wind speed (D-1) Sea ice fraction in buffer LR (D) Sea ice fraction in buffer LR (D-1) Daily mean sea level pressure (D)	Daily mean wind speed (D) Daily wind fraction from 30° (D-1) Daily wind fraction from 310° (D) Daily wind fraction from 80° (D-1) Sea ice fraction in buffer SR (D)

Table 10: The top 5 most influent features as a function of station and month. The color scheme indicates the feature type: wind, ice, water, sea level pressure and time

	Pituffik (Thule Airbase)	Prins Christian Sund	Tasiilaq / Ammassali
Jan	Daily wind fraction from 80° (D) Daily wind fraction from 90° (D) Daily wind fraction from 110° (D) Max SST in buffer LR (D) Daily wind fraction from 120° (D)	Daily wind fraction from 260° (D) Daily wind fraction from 260° (D-1) Daily wind fraction from 270° (D) Daily mean sea level pressure (D) Daily mean sea level pressure (D-1)	Daily wind fraction from 260° (D) Daily wind fraction from 60° (D) Daily wind fraction from 170° (D-1) Daily wind fraction from 30° (D) Daily wind fraction from 170° (D)
Feb	Median SST in buffer SR (D) Daily wind fraction from 100° (D) Median SST in buffer SR (D-1) Daily wind fraction from 300° (D) Daily wind fraction from 90° (D)	Daily wind fraction from 20° (D-2) Daily mean sea level pressure (D) Daily wind fraction from 270° (D-1) Daily wind fraction from 20° (D-1) Daily wind fraction from 270° (D)	Daily wind fraction from 50° (D) Daily wind fraction from 60° (D) Sea ice fraction in buffer LR (D) Daily wind fraction from 70° (D) Daily wind fraction from 50° (D-1)
Mar	Daily wind fraction from 90° (D) Daily mean wind speed (D) Daily wind fraction from 80° (D) Max SST in buffer LR (D) Daily wind fraction from 100° (D)	Daily wind fraction from 270° (D) Sea ice fraction in buffer LR (D-2) Daily wind fraction from 260° (D) Daily mean sea level pressure (D) Daily mean sea level pressure (D-1)	Sea ice fraction in buffer LR (D) Daily wind fraction from 60° (D) Sea ice fraction in buffer LR (D-2) Sea ice fraction in buffer LR (D-1) Daily wind fraction from 90° (D)
Apr	Daily wind fraction from 100° (D) day of year sinus (D) day of year cosin (D) Max SST in buffer LR (D) Daily wind fraction from 100° (D-1)	day of year cosin (D) day of year sinus (D) Daily mean sea level pressure (D) Max SST in buffer LR (D) Daily wind fraction from 230° (D-2)	day of year cosin (D) day of year sinus (D) Daily wind fraction from 30° (D) Daily wind fraction from 30° (D-1) Daily wind fraction from 250° (D)
May	day of year cosin (D) day of year sinus (D) Daily wind fraction from 90° (D) Median SST in buffer LR (D) Daily wind fraction from 160° (D-1)	Daily mean sea level pressure (D) Daily mean sea level pressure (D-1) day of year sinus (D) Daily wind fraction from 230° (D) Daily wind fraction from 30° (D)	day of year sinus (D) day of year cosin (D) Daily mean wind speed (D) Daily wind fraction from 10° (D-1) Daily wind fraction from 360° (D-2)
Jun	day of year sinus (D) day of year cosin (D) Daily wind fraction from 280° (D) Daily wind fraction from 110° (D) Daily wind fraction from 100° (D)	Daily wind fraction from 30° (D) Median SST in buffer LR (D) Daily wind fraction from 90° (D) Median SST in buffer LR (D-1) Daily wind fraction from 120° (D)	day of year sinus (D) day of year cosin (D) Daily wind fraction from 320° (D) Daily wind fraction from 270° (D-1) Daily wind fraction from 270° (D)
Jul	Daily wind fraction from 280° (D-1) Median SST in buffer SR (D) Daily wind fraction from 280° (D) Daily wind fraction from 90° (D) Daily mean sea level pressure (D-2)	Daily wind fraction from 280° (D) Daily wind fraction from 270° (D) Daily wind fraction from 30° (D) Daily wind fraction from 90° (D) Daily wind fraction from 80° (D)	Daily wind fraction from 170° (D) Max SST in buffer SR (D) Daily wind fraction from 170° (D-1) Daily wind fraction from 270° (D) Median SST in buffer LR (D)
Aug	day of year sinus (D) day of year cosin (D) Max SST in buffer SR (D) Max SST in buffer SR (D-1) Median SST in buffer SR (D)	Daily wind fraction from 270° (D) Daily wind fraction from 290° (D) Daily wind fraction from 140° (D) Max SST in buffer LR (D) Daily wind fraction from 90° (D)	day of year cosin (D) day of year sinus (D) Median SST in buffer LR (D) Daily wind fraction from 340° (D-2) Median SST in buffer SR (D)
Sep	Daily wind fraction from 100° (D) day of year cosin (D) Daily wind fraction from 90° (D) Daily wind fraction from 110° (D) day of year sinus (D)	Daily wind fraction from 250° (D) Daily wind fraction from 270° (D) Daily wind fraction from 280° (D) Daily wind fraction from 20° (D) day of year sinus (D)	Max SST in buffer LR (D) day of year sinus (D) Max SST in buffer LR (D-1) day of year cosin (D) Max SST in buffer LR (D-2)
Oct	Daily wind fraction from 110° (D) Daily wind fraction from 100° (D) Sea ice fraction in buffer LR (D) Daily wind fraction from 90° (D) Daily wind fraction from 120° (D)	day of year cosin (D) day of year sinus (D) Max SST in buffer LR (D-1) Max SST in buffer LR (D-2) Median SST in buffer LR (D)	day of year sinus (D) day of year cosin (D) Max SST in buffer LR (D) Daily wind fraction from 160° (D) Daily wind fraction from 80° (D)
Nov	Daily wind fraction from 90° (D) Daily wind fraction from 110° (D) Daily mean wind speed (D) Daily wind fraction from 100° (D) Max SST in buffer LR (D)	Daily mean wind speed (D) Daily wind fraction from 280° (D) Daily mean wind speed (D-1) Daily wind fraction from 270° (D) Daily wind fraction from 10° (D-1)	Daily wind fraction from 280° (D) Daily wind fraction from 260° (D) Daily wind fraction from 60° (D) Daily wind fraction from 350° (D) Daily mean sea level pressure (D-2)
Dec	Daily wind fraction from 90° (D) Daily wind fraction from 100° (D) Daily wind fraction from 110° (D) Daily wind fraction from 110° (D-1) Daily mean wind speed (D)	Daily wind fraction from 280° (D) Daily wind fraction from 30° (D-1) Daily wind fraction from 180° (D) Daily wind fraction from 280° (D-1) Daily wind fraction from 20° (D-1)	Daily wind fraction from 70° (D) Daily wind fraction from 60° (D) Sea ice fraction in buffer LR (D) Daily wind fraction from 60° (D-1) Daily wind fraction from 90° (D)

Focusing on the lagged features in table 9 and 10, it seems that D-1 lagged features do exert a limited influence on the prediction outcome. D-2 lagged features hardly ever appear in the top 5 ranks and do not seem to be influent.

Overall a few remarkable characteristics are observable:

- Wind features are influent at all locations. Notice how important they are for stations Pituffik (especially winds from the inland East, thus likely to be cold katabatic), Prins Christian Sund (subject to tip jets) and Tasillaq (known for its katabatic winds, the Piteraq). Wind from the West also has a large impact at Danmarkshavn in July.
- Generally speaking, the sea surface temperature features are important at different times of the year depending on the stations (with some prevailance in summer).
- The sea level pressure feature appears regularly in the top ranks for Eastern stations. They are particulary influent at Danmarkshavn, Ittoqqortoormiit and Prins Christian Sund during January and February. The fact that these stations are located near a semi-permanent winter low pressure system (called the Icelandic Low) could be an explanation. Curiously, Tasiilaq is apparently not influenced by this feature. Pituffik, shows no special influence from this feature.
- The sea ice fraction features tend to be more important between November and the end of April, than during the rest of the year. It plays almost no role for stations Pittufik (iced in most of the year) and Prins Christian Sund (ice free all year).
- The day of year features generally have large weights in April, August and September for all stations, likely due to the change in day light duration.

The more accurate summer predictions may be attributed to the fact that in winter the water and ice features tend to bring little information (i.e. the sea surface temperature is -1.8°C and the sea ice concentration is close to 100%). Indeed, the feature ranking suggests that predictions in winter months seem to rely more on wind features.
Finally, no pattern was found differentiating the influence of long and short range features.

Aasiaat:  | Danmarkshavn:  | Ittoqqortoormiit:  | Pituffik:  | Prins Christian Sund:  | Tasiilaq: 

4.3 Predictions using a multi-kernel

4.3.1 Performance

As the MKL algorithm is time costly, only four additional predictions were made in an attempt to see if the method provided better results than linear SVR. The feature sets corresponding to the two worst and the two best predictions with linear SVR (see fig. 37) were employed. The performance over 15 runs is provided in table 11.

Table 11: MKL performance over 15 runs for selected stations and months

Station name	**Month**	**SVR RMSE** $[°C]$	**MKL RMSE** $[°C]$
Danmarkshavn	January	5.632	5.305
Ittoqqortoormiit	February	5.398	5.211
Aasiaat / Egedesmind	August	1.282	1.386
Prins Christian Sund	September	1.682	1.753

MKL slightly outperforms linear SVR for the worst prediction cases. For the best prediction cases, MKL has slighty worse performances. Although this remark is based on a very limited number of cases, it appears that the use of MKL does not bring a significant advantage over linear SVR.

4.3.2 Feature ranking

Table 12 shows that the wind direction is the most prominent feature in the best prediction cases. They are followed closely by the sea surface temperature features (as previously when using linear SVR). The sea ice fraction seems to be determinant for the two Northeastern stations of the chosen cases, Danmarkshavn and Ittoqqortoormiit. The rankings obtained here are coherent with those found previously with linear SVR, although there is some variation in the order of the ranks and feature type (attributable to the different types of kernels used).

Table 12: The top 5 most influent features for chosen cases. The color scheme indicates the feature type: wind, ice, water, sea level pressure and time

	Ranking
Danmarkshavn January	Max SST in buffer SR (D), sigma=0.001 Sea ice fraction in buffer LR (D-1), sigma=0.001 Daily mean wind speed (D), sigma=0.2 Max SST in buffer LR (D-2), sigma=0.001 Max SST in buffer SR (D-2), sigma=0.001
Ittoqqortoormiit February	Daily mean wind speed (D), sigma=0.05 Daily mean wind speed (D-1), sigma=0.001 Sea ice fraction in buffer LR (D), sigma=0.05 Sea ice fraction in buffer SR (D-1), sigma=0.001 Median SST in buffer SR (D-2), sigma=0.001
Aasiaat / Egedesmind August	Daily wind directions (D) Daily wind directions (D-1) Max SST in buffer LR (D-2), sigma=0.001 Max SST in buffer LR (D-1), sigma=0.001 day of year cosin (D), sigma=0.2
Prins Christian Sund September	Daily wind directions (D) Daily mean sea level pressure (D-2), sigma=0.001 Daily mean wind speed (D), sigma=0.001 Median SST in buffer SR (D-1), sigma=0.001 Max SST in buffer LR (D-2), sigma=0.001

5 Conclusion

The approach proposed to study atmospheric temperature in this study consisted of two sucessive steps. A preliminary exploratory analysis followed by support vector regression. In terms of comprehension benefits, the first step allowed getting acquainted with the data through different structural and comparative visualizations. It did not however help confirm any specific hypothesis about the nature of the coastal atmospheric temperature regime. In particular, at the conclusion of this first step the influence of wind direction on atmospheric temperature remained unclear. It is important to keep in mind that this part of the study involved the most exploration, with the purpose of observing the data's characteristics before using it with SVR. So no specific results were expected from this phase.
The application of support vector regression produced two helpful results: the spatiotemporal variation of prediction errors and the ranking of features. It was found that prediction accuracy is best for summer months and/or lower latitudes. Root mean squared errors below $5°C$ are routinely achievable all year long and RMSE below $2.5°C$ are obtained in summer months. From the methodological point of view, it was found that the use of SVR with a linear kernel was well suited most of the time. This method has the advantage of being less computationally expensive and faster. The multiple kernel SVR with Gaussian and histogram intersection kernels yielded similar predicition accuracies to those obtained with linear SVR. MKL does however produce slightly better results for winter months, giving a hint that non-linear relations are important during this period. No particular short term lag was detected. Altough previous day features do have some influence on the predictions, synchronous features appear to have the most influence.

Using a statistical approach, SVR, it was shown that a correlation link can be found between offshore variables and coastal atmospheric temperatures. The nature of this correlation link seems to be highly dependant on location and time of year. The feature rankings obtained with SVR indicate that wind features play are prominant role for most stations, with sea surface temperatures and sea ice concentration also having an important influence.

Although the initial objectives were attained, some elements require further investigation:

- the possibility of employing additional in situ and remotely sensed data to build more features (e.g. remotely sensed wind velocity, ocean surface currents, chlorophyll concentration, cloud cover, melt pond concentration, etc).
- the elaboration of more complex features (perhaps using synthetic data combinations)
- the effect of the sample size on prediction performances.
- the suitability of kernels other than linear, Gaussian or histogram intersection.
- the performance of MKL using Gaussian or histogram intersection on more months.
- the spatialization of predictions (beyond weather station locations) combined with the integration of topography linked features.

Altough machine learning algorithms are still scarcely applied to the study of Polar environments, the large availability of archival and current Polar data coupled with accessibility to free ready-to-use machine learning implementations is promising. Overall, the results obtained here seem to indicate the adequacy of this approach for temperature prediction and feature ranking tasks, both of which can help in understanding Polar environments.

References

Abdalati, W., Steffen, K. (2001). "Greenland ice sheet melt extent: 1979-1999". In: *J. Geophys. Res.* 106, pp. 33983–33988.

Brockwell, P., Davis, R. (2002). *Introduction to Time Series and Forecasting*. Springer. ISBN: 9780387953519.

Canu, S. et al. (2005). *SVM and Kernel Methods Matlab Toolbox*.

Chang, C.-C., Lin, C.-J. (2011). "LIBSVM: A library for support vector machines". In: *ACM Transactions on Intelligent Systems and Technology* 2 (3), 27:1–27:27.

Cherkassky, V. S., Mulier, F. (Aug. 2007). *Learning from Data: Concepts, Theory, and Methods*. John Wiley & Sons. ISBN: 9780471681823.

Chevalier, R. F. (2008). "Air Temperature Prediction using Support Vector Regression and GENIE: The Georgia Extreme-weather Neural-network Informed Expert". MA thesis. University of Georgia, Athens.

Comiso, J. C. (2002). "Correlation and trend studies of the sea-ice cover and surface temperatures in the Arctic". In: *Annals of Glaciology* 34, pp. 420–428.

Comiso, J. C. et al. (Jan. 2008). "Accelerated decline in the Arctic sea ice cover". In: *Geophys. Res. Lett.* 35, p. L01703.

Comiso, J. C. (1999, updated 2012). "Bootstrap Sea Ice Concentrations from Nimbus-7 SMMR and DMSP SSM/I". In: *Boulder, Colorado USA: National Snow and Ice Data Center. Digital media.*

Cornuéjols, A., Miclet, L. (June 2010). *Apprentissage artificiel*. Editions Eyrolles. ISBN: 9782212124712.

Drange, H. (2005). *The Nordic Seas: An Integrated Perspective : Oceanography, Climatology, Biogeochemistry, and Modeling*. American Geophysical Union. ISBN: 9780875904238.

Foresti, L. et al. (2009). "Multiple Kernel Learning of Environmental Data. Case Study: Analysis and Mapping of Wind Fields". In: *Artificial Neural Networks - ICANN 2009*. Vol. 5769. Springer Berlin / Heidelberg, pp. 933–943. ISBN: 978-3-642-04276-8.

Guyon, I., Elisseeff, A. (2006). "An Introduction to Feature Extraction". In: *Feature Extraction*. Vol. 207. Springer Berlin / Heidelberg, pp. 1–25. ISBN: 978-3-540-35487-1.

Hamel, L. H. (Sept. 2011). *Knowledge Discovery with Support Vector Machines*. John Wiley & Sons. ISBN: 9781118211038.

Hanna, E., Cappelen, J. (Feb. 2003a). "Recent cooling in coastal southern Greenland and relation with the North Atlantic Oscillation". In: *Geophysical Research Letters* 30, 3 PP.

— (Feb. 2003b). "Recent cooling in coastal southern Greenland and relation with the North Atlantic Oscillation". In: *Geophysical Research Letters* 30, 3 PP.

Heinemann, G. (1999). "The KABEG-97 field experiment: An aircraft-based study of katabatic wind dynamics over the Greenland ice sheet". In: *Boundary-Layer Meteorology* 93, pp. 75–116.

Heinemann*, G. (2003). "Forcing and feedback mechanisms between the katabatic wind and sea ice in the coastal areas of polar ice sheets". In: *Journal of Atmospheric & Ocean Science* 9, pp. 169–201.

Holland, D. M. et al. (Oct. 2008). "Acceleration of Jakobshavn Isbrae triggered by warm subsurface ocean waters". In: *Nature Geosci* 1, pp. 659–664.

Kanevski, M., Pozdnoukhov, A., Timonin, V. (Apr. 2009). *Machine Learning for Spatial Environmental Data: Theory, Applications and Software*. EPFL Press. ISBN: 9780849382376.

Klein, T. et al. (2001). "Mesoscale modeling of katabatic winds over Greenland and comparisons with AWS and aircraft data". In: *Meteorology and Atmospheric Physics* 78, pp. 115–132.

Moore, G. W. K., Pickart, R. S., Renfrew, I. A. (2008). "Buoy observations from the windiest location in the world ocean, Cape Farewell, Greenland". In: *Geophys. Res. Lett.* 35, p. L18802.

Morales Maqueda, M. A., Willmott, A. J., Biggs, N. R. T. (Mar. 2004). "Polynya Dynamics: a Review of Observations and Modeling". In: *Rev. Geophys.* 42, RG1004.

Mote, T. L. (Nov. 2007). "Greenland surface melt trends 1973-2007: Evidence of a large increase in 2007". In: *Geophysical Research Letters* 34, 5 PP.

NIMA (2002). *Sailing directions (planning guide).: Arctic Ocean*. National Imagery and Mapping Agency.

Ortiz-Garcia, E. et al. (2012). "Accurate local very short-term temperature prediction based on synoptic situation Support Vector Regression banks". In: *Atmospheric Research* 107, pp. 1–8.

Paniagua-Tineo, A. et al. (Nov. 2011). "Prediction of daily maximum temperature using a support vector regression algorithm". In: *Renewable Energy* 36, pp. 3054–3060.

Pozdnoukhov, A., Foresti, L., Kanevski, M. (2009). "Data-driven topo-climatic mapping with machine learning methods". In: *Natural Hazards* 50, pp. 497–518.

Prostar Sailing Directions 2005 Greenland and Iceland Enroute (Jan. 2005). ProStar Publications. ISBN: 9781577857532.

Qiu, S., Lane, T. (2005). "Multiple Kernel Learning for Support Vector Regression". In: *Science*, pp. 1–17.

Radhika, Y., Shashi, M. (2009). "Atmospheric T emperature Prediction using Support Vector Machines". In: *International Journal of Computer Theoryand Engineering* 1, pp. 1793–8201.

Rakotomamonjy, A. et al. (2008). *SimpleMKL*.

Rees, G. (2006). *Remote Sensing Of Snow And Ice*. Taylor & Francis. ISBN: 9780415298315.

Rennermalm, A. K. et al. (2009). "Does sea ice influence Greenland ice sheet surface-melt?" In: *Environmental Research Letters* 4, p. 024011.

Reynolds, R. W. et al. (Nov. 2007). "Daily High-Resolution-Blended Analyses for Sea Surface Temperature". In: *Journal of Climate* 20, pp. 5473–5496.

Rignot, E., Koppes, M., Velicogna, I. (2010). "Rapid submarine melting of the calving faces of West Greenland glaciers". In: *Nature Geoscience* 3, pp. 187–191.

Sapankevych, N., Sankar, R. (May 2009). "Time Series Prediction Using Support Vector Machines: A Survey". In: *Computational Intelligence Magazine, IEEE* 4, pp. 24 –38.

Screen, J. A., Simmonds, I. (Apr. 2010). "The central role of diminishing sea ice in recent Arctic temperature amplification". In: *Nature* 464, pp. 1334–1337.

Serreze, M. C. et al. (2009). "The emergence of surface-based Arctic amplification". In: *Cryosphere* 3, pp. 11–19.

Serreze, M. C., Barry, R. G. (Oct. 2005). *The Arctic Climate System*. Cambridge University Press. ISBN: 9780521814188.

Shepherd, A. et al. (Jan. 2009). "Greenland ice sheet motion coupled with daily melting in late summer". In: *Geophys. Res. Lett.* 36, p. L01501.

Smith, W., Barber, D. (2007). *Polynyas: Windows to the World*. Elsevier. ISBN: 9780444529527.

Straneo, F. et al. (Mar. 2010). "Rapid circulation of warm subtropical waters in a major glacial fjord in East Greenland". In: *Nature Geosci* 3, pp. 182–186.

Thissen, U et al. (Nov. 2003). "Using support vector machines for time series prediction". In: *Chemometrics and Intelligent Laboratory Systems* 69, pp. 35–49.

Vage, K. et al. (2009). "Multi-event analysis of the westerly Greenland tip jet based upon 45 winters in ERA-40". In: *Quarterly Journal of the Royal Meteorological Society* 135, pp. 1999–2011.

Ward, J. (1963). "Hierarchical grouping to optimize an objective function". In: *J. Am. Stat. Assoc* (58), pp. 236–244.

Wikipedia contributors (June 2012). *East Greenland Current*.

Xue, S. et al. (Dec. 2009). “Meteorological Prediction Using Support Vector Regression with Genetic Algorithms”. In: *Information Science and Engineering (ICISE), 2009 1st International Conference on*, pp. 4931 –4935.

A Monthly wind azimuth distributions

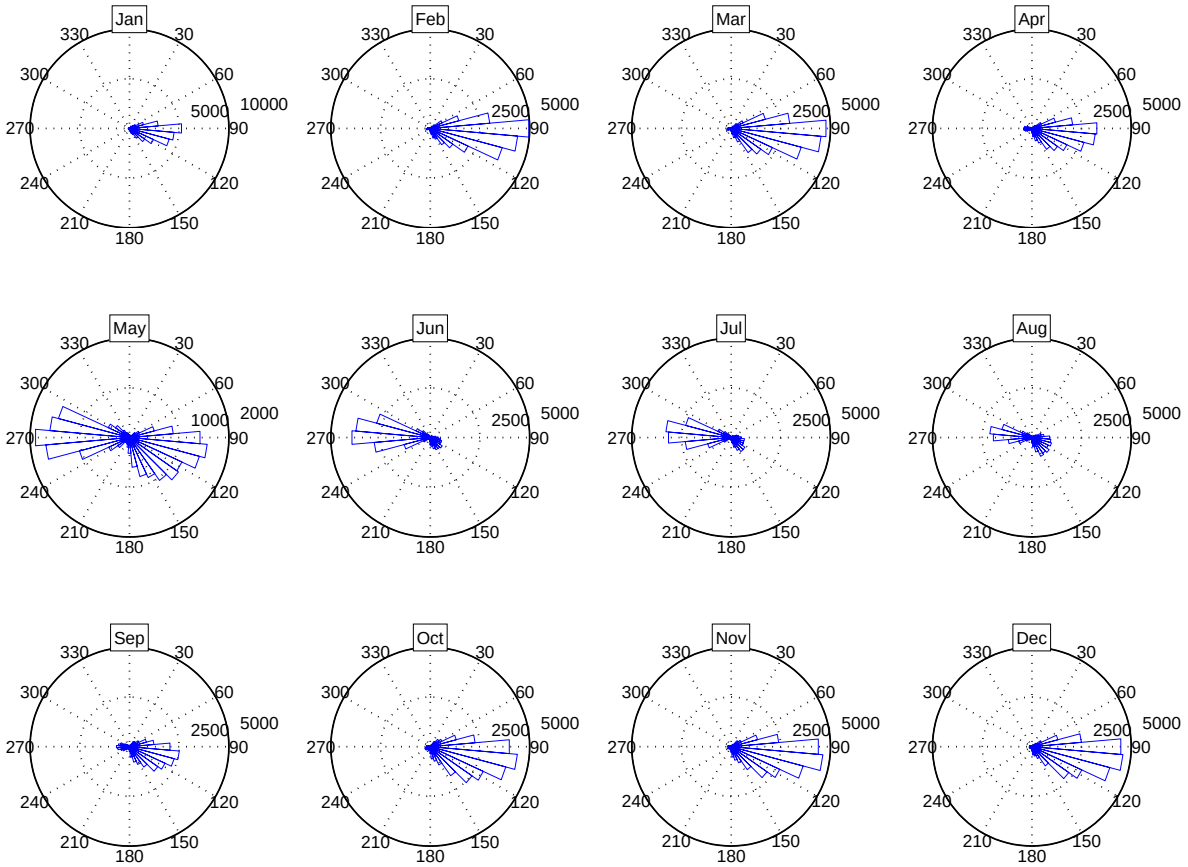

Figure 31: Pituffik - monthly wind azimuth histograms (excluding calm conditions)

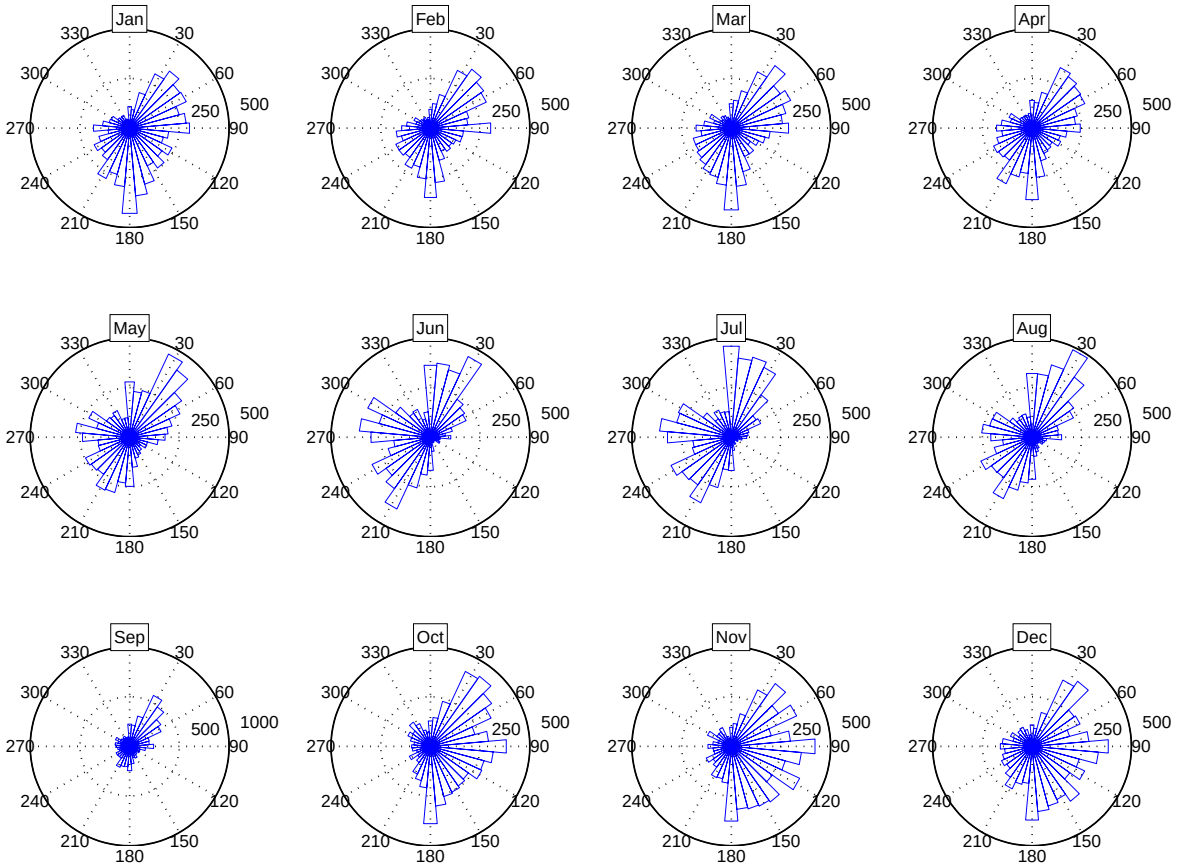

Figure 32: Aasiaat - monthly wind azimuth histograms (excluding calm conditions)

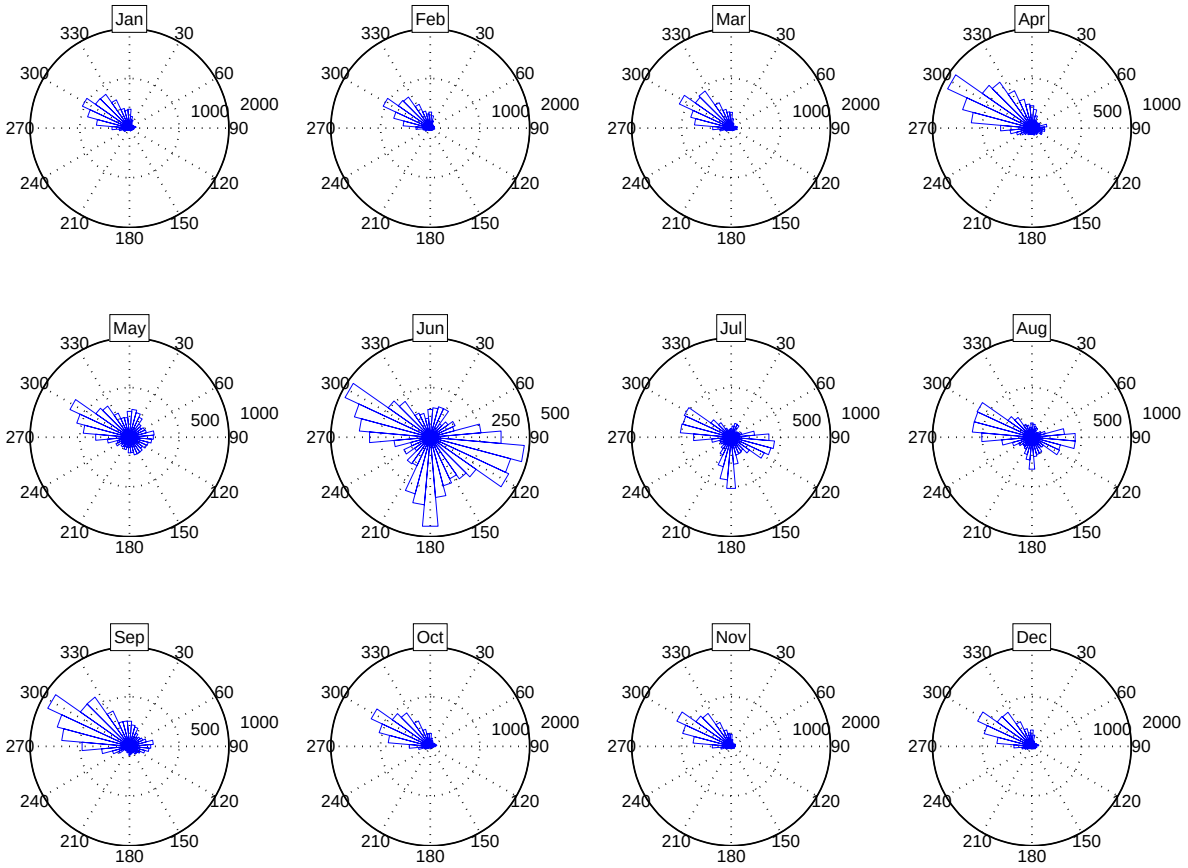

Figure 33: Danmarkshavn - monthly wind azimuth histograms (excluding calm conditions)

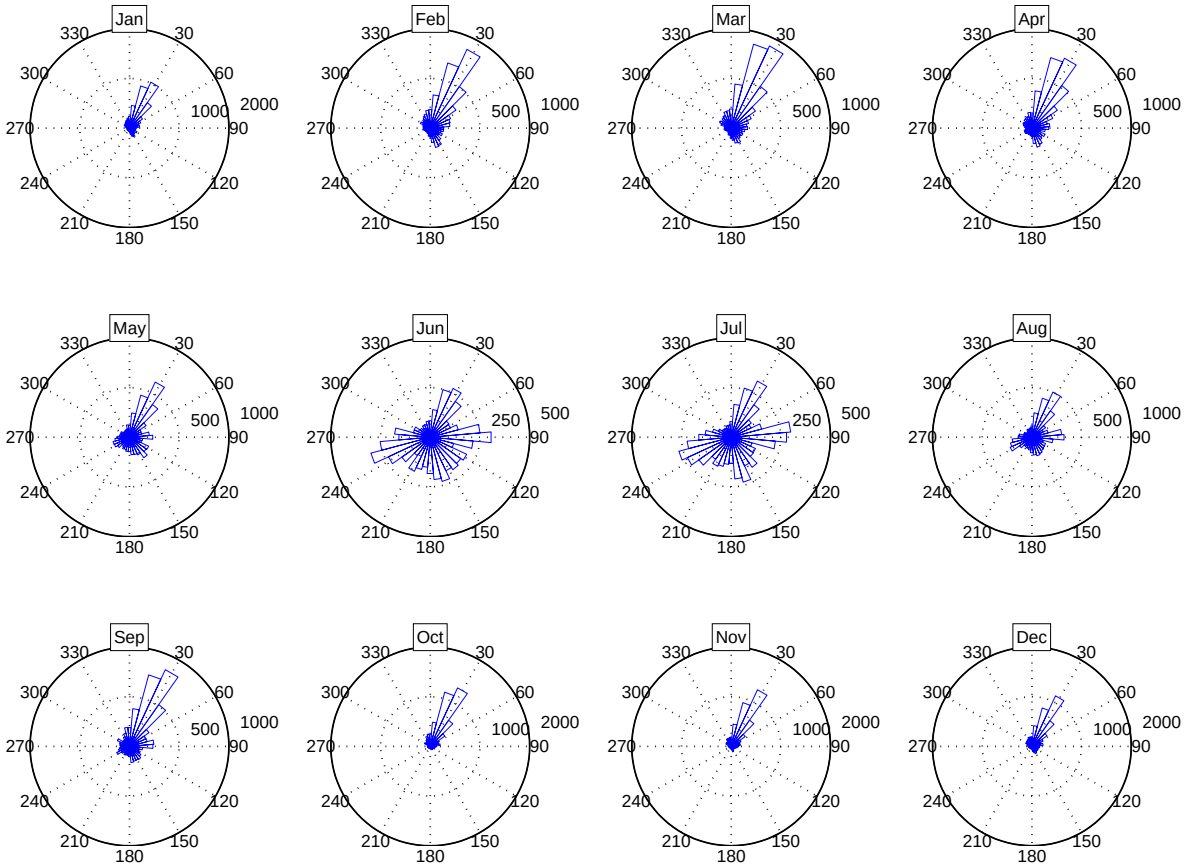

Figure 34: Ittoqqortoormiit - monthly wind azimuth histograms (excluding calm conditions)

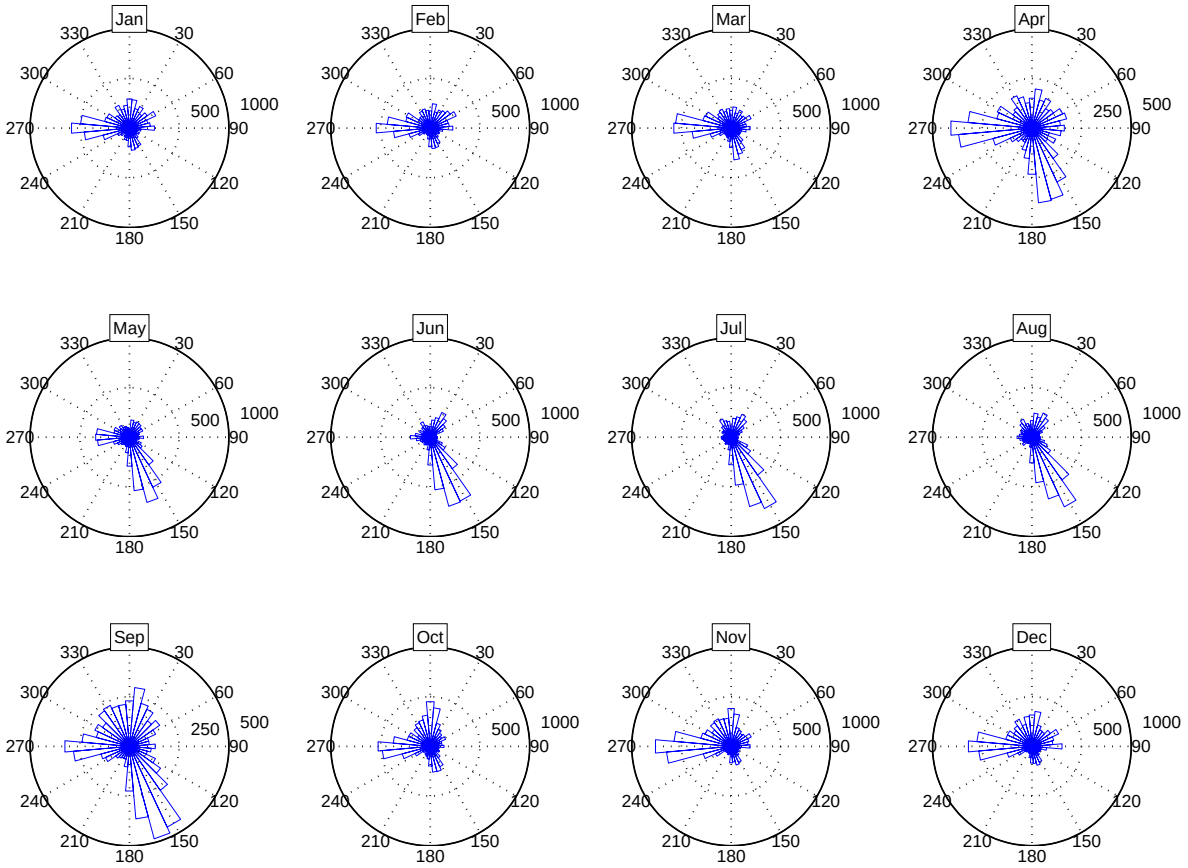

Figure 35: Tasiilaq - monthly wind azimuth histograms (excluding calm conditions)

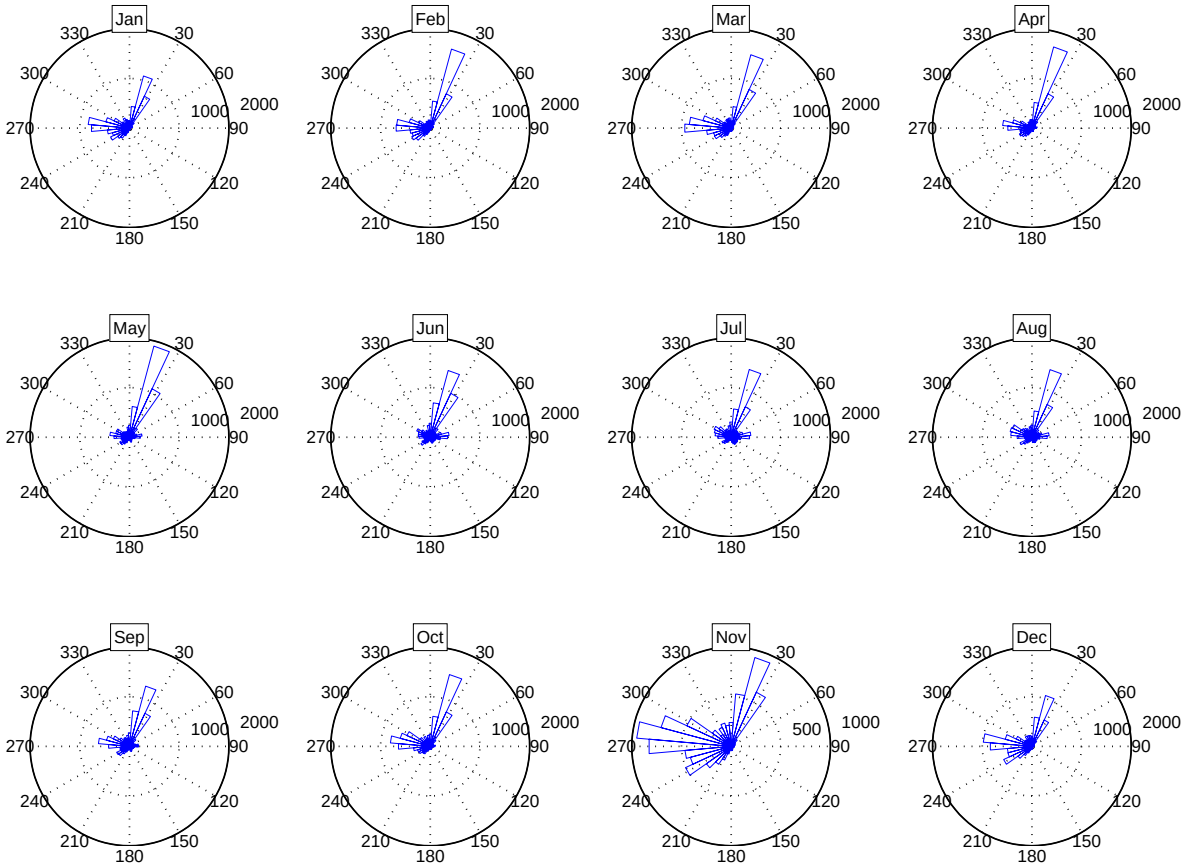

Figure 36: Prins Christian Sund - monthly wind azimuth histograms (excluding calm conditions)

B Detailed SVR performance

Aasiaat: | Danmarkshavn: | Ittoqqortoormiit: | Pituffik: | Prins Christian Sund: | Tasiilaq:

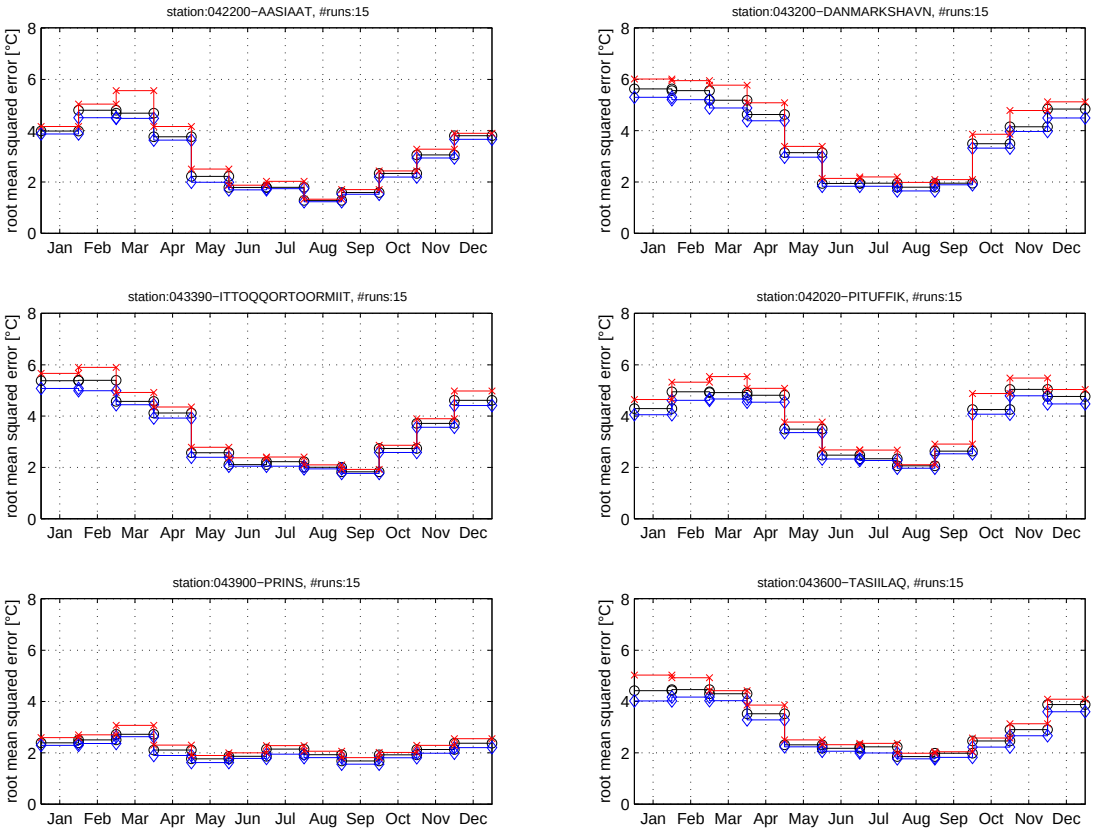

Figure 37: SVR prediction performance as a function of the station and the month. The performance metric is the median of the root mean squared error (RMSE) over 15 runs.

C Arctic ocean currents

Excerpt from the National Geospatial Intelligence Agency Sailing Directions (planning guide) - Pub. 180

The general surface water circulation are the predominant currents of the area throughout the year. However, currents are bound to vary significantly in direction and rate, since they are dependent on oceanographic and meteorological factors with variables. In addition, currents flowing close into the fjords and other fresh water outlets will considerably strengthen runoffs during the summer months, and the rates may exceed 3 knots. The flow of water is mainly determined by two major currents. The Norwegian Atlantic Current sets NNE off the coast of Norway; in about 70°N it divides into the West Spitsbergen Current and the North Cape Current. The East Greenland Current, the major outlet for cold water from the Arctic basin, sets in a generally SW direction along the coast of East Greenland with constancy.

North of 80°N

The flow of water in this area is directed towards the SW, outwards from the Arctic basin. It constitutes the downstream extremity of the wide Transpolar Drift, which commences in East Siberian Sea and mainly emerges from the Arctic through the broad channel between Greenland and Svalbard, where it becomes the East Greenland Current. Two other minor outlets for Arctic water are the East Spitsbergen Current, which sets SSW on the E side of Svalbard, and the Bear Island Current.

South of 80°N and E of 5°E

The West Spitsbergen Current, the W branch of the Norwegian Atlantic Current, sets NNW off Svalbard and at about 80°N it submerges beneath the ice-bearing Arctic water because of its higher salinity. The Spitsbergen Current is the main supplier of warm water into the Arctic basin. The North Cape Current, the E branch of the Norwegian Atlantic Current, sets E into the Barents Sea. The Bear Island Current emerges from the Arctic basin through the channel between Zemlya Frantsa Iosifa and Severnaya Zemlya. The main part merges with the remainder of the East Spitsbergen Current, and then converges with a branch of the North Cape Current to form an eddy in the W part of the Barents Sea. The remainder of the Bear Island Current converges with another branch of the North Cape Current to form an eddy in the E part of the Barents Sea. The rates of the currents in this area are mostly low, probably less than 0.5 knot, but close to the coasts of Svalbard the current may sometimes run up to 2 knots.

South of 80°N and between 5°E and 20°W

The Jan Mayen Current leaves the East Greenland Current at about 76°N,and converges with a recurved branch of the West Spitsbergen Current to form a large permanent eddy centered at about 76°N on the prime meridian. The East Iceland Current leaves the East Greenland Current at about 71°N, and mixes with the Irminger Current off Langanes. Branches from the E flank of the East Iceland Current, together with branches from the W flank of the Norwegian Atlantic Current, form a series of semi-permanent eddies on the meridian of 5°W. The Irminger Current is a warm current, derived from the North Atlantic Current, which is again derived from the Gulf Stream. Off the S coast of Iceland, the Irminger Current sets N to NW; a part encircles the island in a clockwise direction, and mixes with the East Iceland Current as previously described. The combined flow then sets S to converge with the N sets off the SE coast of Iceland. This convergence often being marked by a sharp discontinuity of sea surface temperature. The mean rates of the currents in this area are also low, probably 0.25 to 0.5 knot; S and SE of Iceland the current is variable in direction, with rates from 0.25 to 1 knot. The East Greenland Current, of which the axis of strongest flow lies just seaward of the 200m contour, and the E set off the N coast of Iceland, may sometimes run at 1 to 2 knots.

South of 80°N and W of 20°W

The main current in this area, the East Greenland Current, occupies the NW half of Denmark Strait, and then hugs the SE coast of Greenland. Ice edge movements suggest that a minor branch from the main current sets SE across Denmark Strait to join that part of the Irminger Current which encircles Iceland. The remainder of the Irminger Current, S of about 66°N turns W across the S approaches to the Denmark Strait, and then turns SW to run alongside the East Greenland Current. The positions of the ice edge S of the Denmark Strait suggest that little mixing occurs between these two currents. Instead, the Irminger Current forms a fairly sharp E boundary to the East Greenland Current and to the drift ice which it bears for most months of the year. Farther S this flow of water forms the NE sector of an elongated counterclockwise eddy centered off Kap Farvel. The East Greenland Current and the Irminger Current probably run on average about 0.5 knot, though both currents may attain rates of 1 to 2 knots. Elsewhere, the mean rate is probably less than 0.5 knot.

D NOAA weather station attributes

```
IDENTIFICATION: CDS header
Length:0

USAF: Identification FIXED-WEATHER-STATION USAF MASTER STATION CATALOG identifier

The identifier that represents a FIXED-WEATHER-STATION.
Length:6

NCDC: Identification FIXED-WEATHER-STATION NCDC WBAN identifier

The identifier that represents a FIXED-WEATHER-STATION.
Length:5

DATE: Identification GEOPHYSICAL-POINT-OBSERVATION date

The date of a GEOPHYSICAL-POINT-OBSERVATION.
Length:8

HRMN: Identification GEOPHYSICAL-POINT-OBSERVATION time

The time of a GEOPHYSICAL-POINT-OBSERVATION based on Coordinated Universal
Time Code (UTC).
Length:4

I: Identification GEOPHYSICAL-POINT-OBSERVATION data source flag

The flag of a GEOPHYSICAL-POINT-OBSERVATION showing the source or combination
of sources used in creating the observation.
Length:1
Default Value:9
Table of Values:

1: DATSAV3 observation, candidate for merge with DSI-3280
(not yet merged, failed element cross-checks)
2: DSI-3280 observation, candidate for merge with DATSAV3
(not yet merged, failed element cross-checks)
3: DATSAV3/DSI-3280 merged observation
4: DATSAV3 observation
5: DSI-3280 observation
6: ASOS/AWOS observation from NCDC
7: ASOS/AWOS observation merged with DATSAV3 observation
8: MAPSO observation (NCDC)
A: DATSAV3/DSI-3240 merged observation, candidate for merge with DSI-3280
(not yet merged, failed element cross-checks)
B: DSI-3280/DSI-3240 merged observation, candidate for merge with DATSAV3
(not yet merged, failed element cross-checks)
C: DATSAV3/DSI-3280/DSI-3240 merged observation
D: DATSAV3/DSI-3240 merged observation
E: DSI-3280/DSI-3240 merged observation
F: Form OMR/1001 - Weather Bureau city office (keyed data)
G: SAO surface airways observation, pre-1949 (keyed data)
H: SAO surface airways observation, 1965-1981 format/period (keyed data)
I: Climate Reference Network observation
J: Cooperative Network observation
K: Radiation Network observation
L: Data from Climate Data Modernization Program (CDMP) data source
N: NCAR / NCDC cooperative effort (various national datasets)
```

TYPE: Identification GEOPHYSICAL-REPORT-TYPE code

The code that denotes the type of geophysical surface observation.
Length:5
Default Value:99999
Table of Values:

AERO: Aerological report
AUST: Dataset from Australia
AUTO: Report from an automatic station
BOGUS: Bogus report
BRAZ: Dataset from Brazil
COOPD: US Cooperative Network summary of day report
COOPS: US Cooperative Network soil temperature report
CRB: Climate Reference Book data from CDMP
CRN05: Climate Reference Network report, with 5-minute reporting interval
CRN15: Climate Reference Network report, with 15-minute reporting interval
FM-12: SYNOP Report of surface observation form a fixed land station
FM-13: SHIP Report of surface observation from a sea station
FM-14: SYNOP MOBIL Report of surface observation from a mobile land station
FM-15: METAR Aviation routine weather report
FM-16: SPECI Aviation selected special weather report
FM-18: BUOY Report of a buoy observation
GREEN: Dataset from Greenland
MEXIC: Dataset from Mexico
PCP15: US 15-minute precipitation network report
PCP60: US 60-minute precipitation network report
S-S-A: Synoptic, airways, and auto merged report
SA-AU: Airways and auto merged report
SAO: Airways report (includes record specials)
SAOSP: Airways special report (excluding record specials)
SMARS: Supplementary airways station report
SOD: Summary of day report from U.S. ASOS or AWOS station
SOM: Summary of month report from U.S. ASOS or AWOS station
SURF: Surface Radiation Network report
SY-AE: Synoptic and aero merged report
SY-AU: Synoptic and auto merged report
SY-MT: Synoptic and METAR merged report
SY-SA: Synoptic and airways merged report
WBO: Weather Bureau Office
WNO: Washington Naval Observatory

WIND: WIND-OBSERVATION header
Length:0

DIR: WIND-OBSERVATION direction angle

The angle, measured in a clockwise direction, between true north and the
direction from which the wind is blowing.
Length:3
Scale:1
Unit:Angular Degrees
Default Value:999
Table of Values:

999: Missing. If type code (below) = V, then 999 indicates variable wind direction.

Q: WIND-OBSERVATION direction quality code

The code that denotes a quality status of a reported WIND-OBSERVATION direction
angle.
Length:1
Default Value:9
Table of Values:

0: Passed gross limits check
1: Passed all quality control checks
2: Suspect
3: Erroneous
4: Passed gross limits check , data originate from an NCDC data source
5: Passed all quality control checks, data originate from an NCDC data source
6: Suspect, data originate from an NCDC data source
7: Erroneous, data originate from an NCDC data source
9: Passed gross limits check if element is present

I: WIND-OBSERVATION type code

The code that denotes the character of the WIND-OBSERVATION.
Length:1
Default Value:9
Table of Values:

A: Abridged Beaufort
B: Beaufort
C: Calm
H: 5-Minute Average Speed
N: Normal
Q: Squall
R: 60-Minute Average Speed
T: 180 Minute Average Speed
V: Variable

SPD: WIND-OBSERVATION speed rate

The rate of horizontal travel of air past a fixed point.
Length:4
Scale:10
Unit:Meters per Second
Default Value:9999

Q: WIND-OBSERVATION speed quality code

The code that denotes a quality status of a reported WIND-OBSERVATION speed
rate.
Length:1
Default Value:9
Table of Values:

0: Passed gross limits check
1: Passed all quality control checks
2: Suspect
3: Erroneous
4: Passed gross limits check , data originate from an NCDC data source
5: Passed all quality control checks, data originate from an NCDC data source
6: Suspect, data originate from an NCDC data source
7: Erroneous, data originate from an NCDC data source
9: Passed gross limits check if element is present

TEMP: AIR-TEMPERATURE-OBSERVATION header
Length:0

TEMP: AIR-TEMPERATURE-OBSERVATION air temperature

The temperature of the air.
Length:5
Scale:10
Unit:Degrees Celsius
Default Value:+9999

Q: AIR-TEMPERATURE-OBSERVATION air temperature quality code

The code that denotes a quality status of an AIR-TEMPERATURE-OBSERVATION.

Length:1
Default Value:9
Table of Values:

0: Passed gross limits check
1: Passed all quality control checks
2: Suspect
3: Erroneous
4: Passed gross limits check , data originate from an NCDC data source
5: Passed all quality control checks, data originate from an NCDC data source
6: Suspect, data originate from an NCDC data source
7: Erroneous, data originate from an NCDC data source
9: Passed gross limits check if element is present

DEWPT: AIR-TEMPERATURE-OBSERVATION-DEWPOINT header
Length:0

DEWPT: AIR-TEMPERATURE-OBSERVATION-DEWPOINT temperature

The temperature to which a given parcel of air must be cooled at constant
pressure and water vapor content in order for saturation to occur.
Length:5
Scale:10
Unit:Degrees Celsius
Default Value:+9999

Q: AIR-TEMPERATURE-OBSERVATION-DEWPOINT quality code

The code that denotes a quality status of the reported dew point temperature.

Length:1
Default Value:9
Table of Values:

0: Passed gross limits check
1: Passed all quality control checks
2: Suspect
3: Erroneous
4: Passed gross limits check, data originate from an NCDC data source
5: Passed all quality control checks, data originate from an NCDC data source
6: Suspect, data originate from an NCDC data source
7: Erroneous, data originate from an NCDC data source
9: Passed gross limits check if element is present

SLP: ATMOSPHERIC-PRESSURE-OBSERVATION header
Length:0

SLP: ATMOSPHERIC-PRESSURE-OBSERVATION sea level pressure

The air pressure relative to Mean Sea Level (MSL).
Length:5
Scale:10
Unit:Hectopascals
Default Value:99999

Q: ATMOSPHERIC-PRESSURE-OBSERVATION sea level pressure quality code

The code that denotes a quality status of the sea level pressure of an
ATMOSPHERIC-PRESSURE-OBSERVATION.
Length:1
Default Value:9
Table of Values:

0: Passed gross limits check
1: Passed all quality control checks
2: Suspect
3: Erroneous
4: Passed gross limits check , data originate from an NCDC data source
5: Passed all quality control checks, data originate from an NCDC data source
6: Suspect, data originate from an NCDC data source
7: Erroneous, data originate from an NCDC data source
9: Passed gross limits check if element is present