FINNISH METEOROLOGICAL INSTITUTE
CONTRIBUTIONS

No. 90

CATEGORICAL METEOROLOGICAL PRODUCTS: EVALUATION AND ANALYSIS

Otto Hyvärinen

Department of Physics
Faculty of Science
University of Helsinki
Helsinki, Finland

ACADEMIC DISSERTATION in meteorology

To be presented, with the permission of the Faculty of Science of the University of Helsinki, for public criticism in Auditorium of Finnish Meteorological Institute (Erik Palménin aukio 1) on 4 November 2011, at 12 o'clock noon.

Finnish Meteorological Institute
Helsinki, 2011

ISBN 978-951-697-751-8 (paperback)
ISSN 0782-6117
Unigrafia
Helsinki, 2011

ISBN 978-951-697-752-5 (PDF)
http://ethesis.helsinki.fi
Helsinki, 2011

FINNISH METEOROLOGICAL INSTITUTE

Published by Finnish Meteorological Institute

P.O. Box 503
FIN-00101 Helsinki, Finland

Series title, number and report code of publication
Finnish Meteorological Institute
Contributions 90, FMI-CONT-90

Date
September 2011

Author
Otto Hyvärinen

Name of project

Commissioned by

Title
Categorical meteorological products: evaluation and analysis

Abstract

In meteorology, observations and forecasts of a wide range of phenomena – for example, snow, clouds, hail, fog, and tornados – can be categorical, that is, they can only have discrete values (e.g., "snow" and "no snow"). Concentrating on satellite-based snow and cloud analyses, this thesis explores methods that have been developed for evaluation of categorical products and analyses.

Different algorithms for satellite products generate different results; sometimes the differences are subtle, sometimes all too visible. In addition to differences between algorithms, the satellite products are influenced by physical processes and conditions, such as diurnal and seasonal variation in solar radiation, topography, and land use. The analysis of satellite-based snow cover analyses from NOAA, NASA, and EUMETSAT, and snow analyses for numerical weather prediction models from FMI and ECMWF was complicated by the fact that we did not have the true knowledge of snow extent, and we were forced simply to measure the agreement between different products. The Sammon mapping, a multidimensional scaling method, was then used to visualize the differences between different products.

The trustworthiness of the results for cloud analyses [EUMETSAT Meteorological Products Extraction Facility cloud mask (MPEF), together with the Nowcasting Satellite Application Facility (SAFNWC) cloud masks provided by Météo-France (SAFNWC/MSG) and the Swedish Meteorological and Hydrological Institute (SAFNWC/PPS)] compared with ceilometers of the Helsinki Testbed was estimated by constructing confidence intervals (CIs). Bootstrapping, a statistical resampling method, was used to construct CIs, especially in the presence of spatial and temporal correlation.

The reference data for validation are constantly in short supply. In general, the needs of a particular project drive the requirements for evaluation, for example, for the accuracy and the timeliness of the particular data and methods. In this vein, we discuss tentatively how data provided by general public, e.g., photos shared on the Internet photo-sharing service Flickr, can be used as a new source for validation. Results show that they are of reasonable quality and their use for case studies can be warmly recommended.

Last, the use of cluster analysis on meteorological in-situ measurements was explored. The Autoclass algorithm was used to construct compact representations of synoptic conditions of fog at Finnish airports.

Publishing unit
Finnish Meteorological Institute, Meteorological Research

Classification (UDC)
551.501.6, 551.501.86,
551.578.46, 551.576,
551.575

Keywords
weather satellites, bootstrap,
snow cover, cloud analysis,
validation, Sammon mapping

ISSN and series title
0782-6117 Finnish Meteorological Institute Contributions

ISBN
978-951-697-751-8 (paperback), 978-951-697-752-5 (pdf)

Language
English

Pages
138

Price

Sold by
Finnish Meteorological Institute / Library
P.O. Box 503, FIN-00101 Helsinki, Finland

Note

ILMATIETEEN LAITOS

Julkaisija
Ilmatieteen laitos

PL 503, 00101 Helsinki

Julkaisun sarja, numero ja raporttikoodi
Finnish Meteorological Institute Contributions 90, FMI-CONT-90

Julkaisuaika
Syyskuu 2011

Tekijä
Otto Hyvärinen

Projektin nimi

Toimeksiantaja

Nimike
Luokkamuotoisten säätuotteiden laadunarviointi ja analyysi

Tiivistelmä

Monia säähän liittyviä ilmiöitä havainnoidaan jakamalla ne kahteen tai useampaan toisensa poissulkevaan luokkaan. Tätä jaottelua käytetään yleisesti sääsatelliittikuvatuotteiden analyysissä, jolloin esimerkiksi kuvan pikseli luokitellaan joko lumiseksi tai lumettomaksi.

Väitöskirjan päätutkimuskohteena olivat sääsatelliittihavaintoihin perustuvat automaattiset luokkamuotoiset pilvisyys- ja lumianalyysit. Osassa artikkeleista tarkasteltiin myös ihmisten tekemiä rae- ja sumuhavaintoja. Tarkoituksena oli soveltaa erilaisia laadunarviointimenetelmiä erityisesti pilvisyys- ja lumituotteisiin.

Osassa väitöskirjaa verrattiin sää- ja avaruusorganisaatioiden, EUMETSAT, NOAA ja NASA, tuottamia satelliittipohjaisia lumituotteita Ilmatieteen laitoksen ja ECMWF:n säänennustusmallien lumianalyyseihin. Koska lumipeitteen todellisesta laajuudesta on hankalaa saada hyvälaatuisia riippumattomia havaintoja, lumituotteita ei voi asettaa paremmuusjärjestykseen. Siksi väitöskirjassa lumianalyysien ja -tuotteiden eroja visualisoitiin Sammonin kuvausta (Sammon mapping) apuna käyttäen. Tämä tilastollinen menetelmä projisoi moniulotteisen aineiston kahteen ulottuvuuteen.

Väitöskirjan toisen vertailuaineiston muodostivat Euroopan sääsatelliittijärjestön EUMETSATin MPEF- ja SAFNWC-ohjelmien kokonaispilvituotteet sekä Helsinki Testbed -alueen ceilometrihavainnot. Pilvituotteiden luotettavuutta arvioitiin laskemalla vertailutuloksille luottamusvälit bootstrap-menetelmällä. Koska pilviaineistolle on ominaista ajallinen ja paikallinen korrelaatio, menetelmän käyttäminen oli perusteltua.

Säätuotteiden laadunarviointia varten tarvitaan aina riippumattomia havaintoja, mutta niiden hankkiminen voi olla hankalaa ja kallista. Siksi työssä tarkasteltiin, onko laadunarviointia mahdollista täydentää tavallisten ihmisten säähavainnoilla, esimerkiksi Internetin kuvapalvelu Flickrin sääaiheisten valokuvien avulla. Nämä GPS-paikannetut kuvat ovat luotettavia ja helposti käytettäviä todisteita jonkin ilmiön esiintymisestä tietyssä paikassa tiettyyn aikaan. Siksi niitä voi suositella etenkin tapaustutkimusaineistoksi.

Väitöskirjan viimeisessä osassa suomalaisilla lentokentillä tehtyjä sumuhavaintoja ryhmiteltiin Autoclass-menetelmällä. Saadulla tilastollisella tiedolla voidaan esimerkiksi auttaa päivystävää meteorologia tunnistamaan sumulle otollisia sääolosuhteita, sen sijaan, että hänen täytyisi opetella sumutietämys kokemustensa kautta.

Julkaisijayksikkö
Meteorologinen tutkimus

Luokitus (UDK)
551.501.6, 551.501.86, 551.578.46, 551.576, 551.575

Asiasanat
sääsatelliitit, bootstrap-menetelmä, kokonaispilvisyys, lumipeite, validointi, Sammonin kuvaus

ISSN ja avainnimike
0782-6117 Finnish Meteorological Institute Contributions

ISBN
978-951-697-751-8 (paperback), 978-951-697-752-5 (pdf)

Kieli
englanti

Sivumäärä
138

Hinta

Myynti
Ilmatieteen laitos / Kirjasto
PL 503, 00101 Helsinki

Lisätietoja

Preface

The work presented in this thesis has been carried out at the Research and Development of the Finnish Meteorological Institute (FMI) during the period 2007-2011. But the seeds were sown in my very first summer job in FMI at Christmas 1994, when I did my first calculations of verification measures using FORTRAN-77. After that my twisty-turny career path lead me to operational satellite work, and then back to work on evaluation.

First and foremost, I want to thank Profs. Sylvain Joffre, David M. Schultz, and Jarkko Koskinen for showing me how scientific papers are written, Pirkko Pylkkö for being my mentor in all things satellite and Pertti Nurmi for introducing me to the world of verification in the beginning and helping me at the very end (and writing all those unpublished and therefore unquotable papers), my Custos, Prof. Hannu Savijärvi, for guiding me safely to my final destination, and my pre-examiners, Drs. Christopher Ferro and Marion Mittermaier, for their encouraging and thoughtful comments.

And of course without my co-authors — Elena, Janne, Jukka, Kalle, Niilo, Sauli, and Vesa — I would not have got to this point at all. The times of a lone scientist in the ivory tower are over.

Many thanks for all the people around the coffee table during coffee breaks, who directly or indirectly helped when life was bleak (e.g., Upper, 1974). The table and people around it change but the spirit of the coffee table does not.

And in the age of Internet, kudos must be given to unsung heroes of obscure newsgroups and mailing lists who have had time to answer questions from perplexed newcomers. Their advice was much needed and appreciated, sometimes years later. Unbeknownst to them, they formed my extended invisible circle of mentors.

And finally, Marianne, you are my ground of being.

Helsinki, September 2011

Otto Hyvärinen

Contents

List of acronyms

AVHRR Advanced Very High Resolution Radiometer

B Bias

CI Confidence interval

CMA China Meteorological Administration

CSI Critical success index

E the reference value for the skill score

EO Earth Observing

EOS Earth Observing System

EPS EUMETSAT Polar System

ESA European Space Agency

EUMETSAT European Organisation for the Exploitation of Meteorological Satellites

F False alarm rate

FAR False alarm ratio

GPS Global Positioning System

H Hit rate

HIRLAM High Resolution Limited Area Model

HRV High-resolution visible

HSS Heidke Skill Score

ICA Independent component analysis

IEEE Institute of Electrical and Electronics Engineers

IGARSS IEEE International Geoscience & Remote Sensing Symposium

IMS Interactive Multisensor Snow and Ice Mapping System

JPSS Joint Polar Satellite System

KSS Hanssen-Kuiper Skill Score

LandSat Satellite for land studies

LSA SAF Land Surface Analysis Satellite Application Facility

MDS Multidimensional scaling

Metop Meteorological Operational

Meteosat Meteorological Satellite

MODIS Moderate Resolution Imaging Spectroradiometer

MPEF Meteorological Products Extraction Facility

MSG Meteosat Second Generation

MTG Meteosat Third Generation

NASA National Aeronautics and Space Administration

NESDIS National Environmental Satellite, Data, and Information Service

NOAA National Oceanic and Atmospheric Administration

NPOESS National Polar-orbiting Operational Environmental Satellite System

NWCSAF Satellite Application Facility on Support to Nowcasting and Very Short Range Forecasting

NWP Numerical weather prediction

PC Proportion Correct

PCA Principal Component Analysis

PCO Principal Coordinates Analysis

PPS Polar Platform System

PSS Peirce Skill Score

SAF Satellite Application Facility

SEVIRI Spinning Enhanced Visible and Infrared Imager

SS Skill Score

UTC Coordinated Universal Time

List of original publications

I Siljamo N., Hyvärinen O. (2011): New geostationary satellite-based snow cover algorithm. *Journal of Applied Meteorology and Climatology*, **50 (6)**, 1275–1290. doi: 10.1175/2010JAMC2568.1

II Joro S., Hyvärinen O., Kotro J. (2010): Comparison of satellite cloud masks with ceilometer sky conditions in southern Finland. *Journal of Applied Meteorology and Climatology*, **49 (12)**, 2508–2526. doi: 10.1175/2010JAMC2442.1

III Hyvärinen O., Eerola K., Siljamo N., Koskinen J. (2009): Comparison of snow cover from satellite and numerical weather prediction models in Northern Hemisphere and northern Europe. *Journal of Applied Meteorology and Climatology*, **48 (6)**, 1199–1216. doi: 10.1175/2008JAMC2069.1

IV Hyvärinen, O., Saltikoff E. (2010): Social media as a source of meteorological observations. *Monthly Weather Review*, **138 (8)**, 3175–3184. doi: 10.1175/2010MWR3270.1

V Hyvärinen O., Julkunen J. and Nietosvaara V. (2007). Climatological tools for low visibility forecasting, *Journal of Pure and Applied Geophysics*, **164**, 1383–1396. doi: 10.1007/s00024-007-0224-5

1 Introduction

Weather satellites provide observations of the state and evolution of the Earth's surface and its atmosphere. Their use as a qualitative tool for forecasters has been essential ever since the first satellites devoted to meteorology. Lately, thanks to advances in data assimilation, they have become indispensable data source for numerical weather prediction (NWP). Between these two extremes, images for humans and raw radiances for computers, there is still room for quantitative products that both humans and computers can utilize. For example, snow analyses can be beneficial for both duty forecasters and NWP models. These products are often categorical, that is, they can only have discrete values (e.g.,"snow" and "no snow"). However, categorical classification of phenomena in meteorology is not restricted to satellite-based products, but observations and forecasts of a wide range of phenomena—for example, hail, fog, and tornados—can be categorical.

But how should categorical analyses be evaluated? What kind of measures or statistics are available? What are the common pitfalls to be avoided? Luckily, these musings do not need to start from scratch, as the same questions have been discussed since the 1880's. This thesis explores some methods that have been developed for categorical forecasts and analyses. The trustworthiness of the results obtained is estimated by constructing confidence intervals within which the true value should be with a certain probability. This is often recommended but is not always practiced as much as it should be. This thesis uses simple and not so simple examples to explore confidence intervals. The analysis is complicated by the fact that we do not always have the full knowledge of how things really are, and we are sometimes forced simply to measure the agreement between different products.

The contents of Papers I-V and the author's contribution are briefly outlined below.

- Paper I compares two satellite-based snow-cover products from the Land Surface Analysis Satellite Application Facility (LSA SAF) and the National Oceanic and Atmospheric Administration/National Environmental Satellite, Data, and Information Service (NOAA/NESDIS). The NOAA/NESDIS product is assumed to be a better representative of the real snow conditions. The performance of the LSA SAF snow cover is then assessed. The author was responsible for the statistical analysis and contributed substantially to the introduction and conclusions.

- Paper II compares the total cloudiness estimated by three different satellite products with ground-based lidar measurements. The effect of temporal and spatial correlation on the results is discussed. The bootstrap method is used to construct the confidence intervals for different measures. The author was

responsible for the statistical analysis and contributed substantially to the introduction and conclusions.

- Paper III compares five different snow cover products. None of them is wholly independent from the others and none of them can be said to be representation of the real snow conditions. The agreement between analyses was explored using multidimensional scaling. The author was responsible for the statistical analysis and contributed substantially to the writing as the lead author.

- Paper IV discusses whether the vast amounts of data available from social network sites on Internet can be used as a source of validation for meteorological applications. The quality of metadata for photos from Flickr, a photo sharing site, was assessed. A case study for validation of hail observations from weather radar was conducted using photos from Flickr. The author was responsible for the statistical analysis and contributed substantially to the writing as the lead author.

- Paper V shows the use of cluster analysis for giving a compact representation of synoptic conditions of fog at airports. The author was responsible for the statistical analysis and contributed substantially to the writing as the lead author.

In this introduction part of the thesis, the satellite instruments used in Papers I, II, and III are introduced first, followed by a brief outline of the nature of the remote-sensed data and methods typically applied for constructing products from the data. Methods of evaluating categorical products are then reviewed and, finally, methods for visualization the differences between these products are examined.

2 Construction of satellite-based products for meteorology

Noordung (1929) was perhaps the first to suggest that a space station positioned in a geostationary orbit would be useful for meteorology. This started to become reality fifty years ago, when the first (polar) satellite dedicated to meteorology was launched (Fritz and Wexler, 1960; Kållberg et al., 2010). Kidder and Vonder Haar (1995) present a history of satellites from meteorological perspective.

Observations from satellites have obvious benefits in terms of spatial and temporal coverage of synoptic and mesoscale phenomena for meteorological applications. However, satellite observations are radically different from traditional meteorological in-situ observations of air temperature, pressure, and humidity. All remote-sensing instruments measure electromagnetic radiation. The instruments are either passive, measuring radiation emitted or reflected by some other body, or the instruments are active, sending a pulse of radiation that other bodies reflect or scatter back to the sensor. Passive instruments can be divided into imagers and sounders. Imagers have better spatial resolution and fewer channels (narrow segments of spectrum measured), usually in the visible, near-infrared and window regions (where the atmosphere is transparent) of infrared. On the other hand, sounders are more concerned in the vertical structure of the atmosphere and have more channels with moderate spatial resolution, especially in the spectral regions where the atmosphere is semi-transparent.

The first widespread use of satellite data was to provide images from the imaging instruments. These images give a qualitative view of the present state of the atmosphere and are still very important for assessing the synoptic situation and nowcasting (e.g., Bader et al., 1995; ZAMG, 2009). The quantitative use of satellite data in NWP developed more slowly, but recent advances in the data assimilation have made them an essential part of the observing system of NWP models (e.g., Kelly and Thépaut, 2007). NWP models mostly utilize data from sounders, but products derived from images — such as atmospheric motion vectors ("winds") and surface temperatures — are also used, along with products from scatterometers and Global Positioning System (GPS) radio occultation measurements.

This chapter introduced the satellites used for this thesis, and their instruments, and describes how products are constructed from the data. This thesis is oriented towards imagers measuring in the visible and infra-red regions of the spectrum. Therefore, active and passive instruments in the microwave region are not discussed. Nor do we discuss satellites and instruments that are mainly designed for land applications (e.g., Landsat).

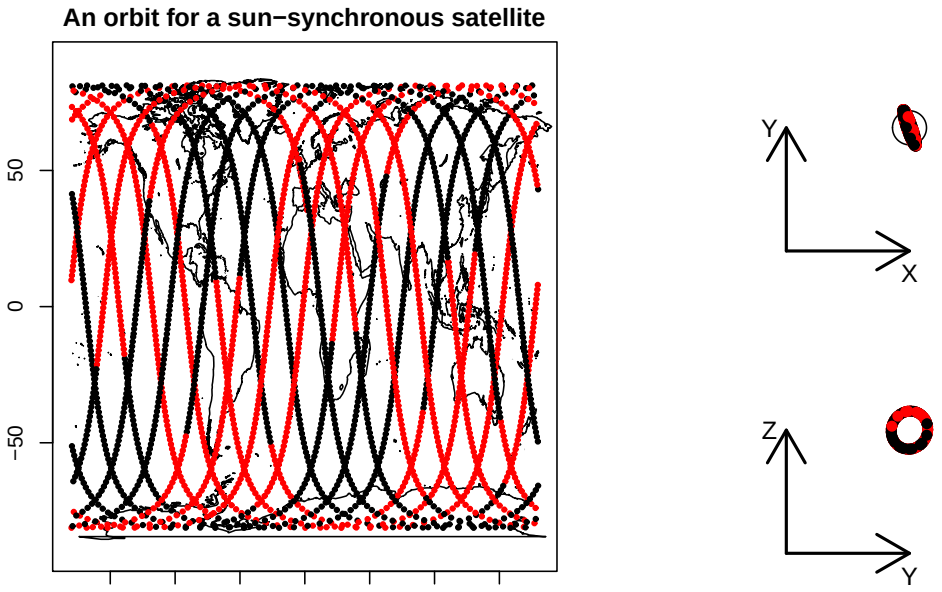

FIGURE 2.1. The 24-hour coverage of a hypothetical satellite on a sun-synchronous orbit relative to the Earth's surface and the right ascension-declination coordinate system, where z is aligned with the Earth's spin and x points to the Sun at vernal equinox. The colors alternate every hour.

2.1 SATELLITES AND THEIR INSTRUMENTS

Satellites can roughly be divided into operational and experimental or research satellites. In theory, operational satellites form a steady source of data that can span decades, while research satellites can come and go without further notice. However, in practice, data from research satellites are used increasingly by the operational services if possible. For example, data from the Moderate Resolution Imaging Spectroradiometer (MODIS) instrument onboard the Earth Observing System (EOS) research satellites are often used in operational setting (e.g., Bormann and Thépaut, 2004). Another useful way of classifying satellites is by their orbits. A thorough introduction to orbits can be found, for example, in Kidder and Vonder Haar (1995). In this thesis, satellites using sun-synchronous near-polar and geostationary orbits are used.

Sun-synchronous orbits have a high inclination angle (the angle between the equatorial plane and the orbital plane) — for example, about 98 degrees for NOAA satellites — so they reach high latitudes and are therefore called near-polar orbits, often shortened, somewhat imprecisely, to polar orbits (Figure 2.1). The satellite series from NOAA has been the main operational meteorological polar satellite series since the 1980's and its Advanced Very High Resolution Radiometer (AVHRR) instrument, with 6 channels and 1.1 kilometer resolution, has been the most-used and best-known imager of polar orbiters. Operating since 2006, the European

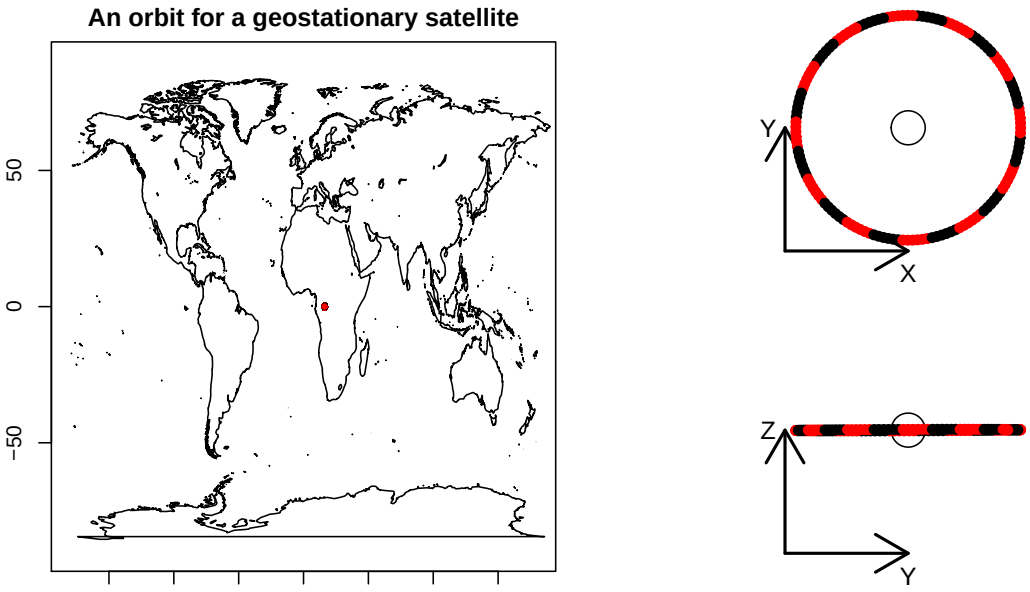

FIGURE 2.2. As in Figure 2.1 but for a hypothetical satellite on a geostationary orbit.

EUMETSAT Polar System (EPS) Meteorological Operational (Metop) satellite from EUMETSAT also has an AVHRR onboard. Products utilizing AVHRR data were used in PAPER II.

Noteworthy polar-orbiting research satellites are three EOS satellites from the National Aeronautics and Space Administration (NASA), of which two, Aqua and Terra, have the MODIS instrument onboard. The MODIS instrument is more advanced than the AVHRR, as it has more channels (36 compared to 6), some of which have better spatial resolution. Products based on MODIS data were used quantitatively in PAPER III, and MODIS images were used qualitatively in PAPER I.

In the geostationary orbit, the satellite orbits the Earth as fast as the Earth rotates, or in other words, they have the same angular velocity, and for the observer on the ground it seems that the satellite is not moving (Figure 2.2). This orbit makes it possible to obtain observations with high temporal resolution, but observations of high latitudes are problematic because the viewing angle from the satellite is very low. A low viewing angle is cumbersome, as the surface area included into one pixel grows as the viewing angle decreases and, compared against the straight view to the nadir, the view is through more atmosphere, which complicates quantitative calculations.

For Europe, geostationary satellites from EUMETSAT are the most relevant geostationary satellites for nowcasting and local NWP (for global NWP models, data from all geostationary satellites are useful). The Spinning Enhanced Visible and Infrared Imager (SEVIRI) instrument onboard Meteosat Second Generation (MSG) has 11 channels at 3 km resolution and the High-resolution visible (HRV)

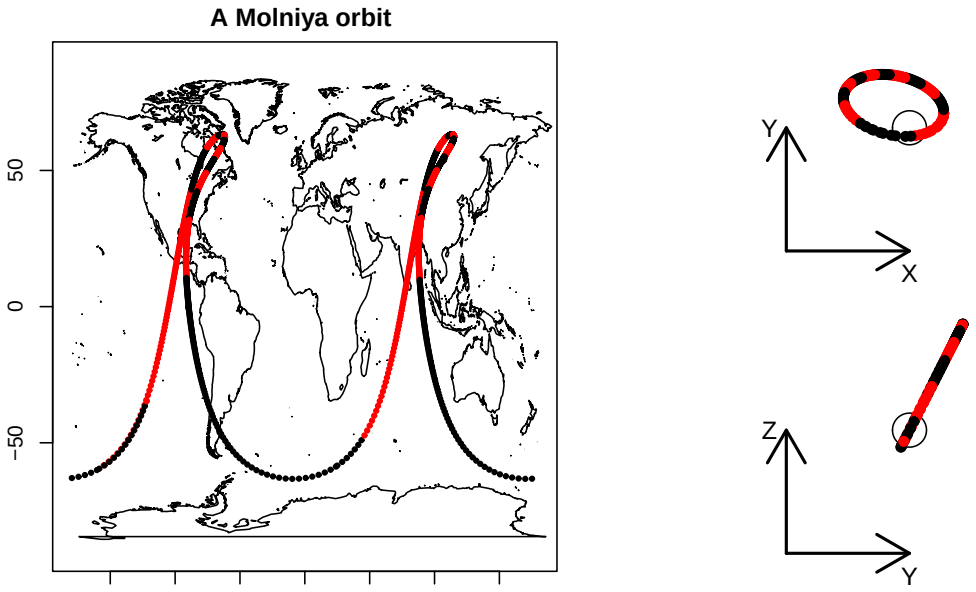

FIGURE 2.3. As in Figure 2.1 but for a hypothetical satellite on a Molniya orbit.

channel that has a spatial resolution comparable to that of the AVHRR. Products based on SEVIRI data were used in PAPERS I, II and III.

Of course, the satellite systems continue evolving. For European users, the most important changes will take place in the late 2010's when EPS satellites will be replaced by EPS Second Generation satellites and MSG by Meteosat Third Generation (MTG), while NOAA series are being replaced by the Joint Polar Satellite System (JPSS) [originally the National Polar-orbiting Operational Environmental Satellite System (NPOESS)]. New polar orbiters from the European Space Agency (ESA) and China Meteorological Administration (CMA) will be also useful. Satellites using new orbits are also planned. Especially for polar regions, Molniya orbits, named after the series of telecommunication satellites from the former Soviet Union, would be most useful (Figure 2.3). The orbit has a large eccentricity so satellites remain near the apogee (the furthest point from the Earth) quite awhile. Such an orbit is therefore often called quasi-geostationary. Satellites on this orbit could make observations of high latitudes with temporal resolution almost as good as that of geostationary satellites, but with a much better viewing angle (Trishchenko and Garand, 2011). Such meteorological and environmental satellites are very much anticipated.

2.2 SPECTRAL CHARACTERISTICS OF SATELLITE DATA

Roughly, there are two sources of radiation measured by passive instruments: the Sun and the Earth. The short-wave (visible and near-infrared) radiation of the Sun

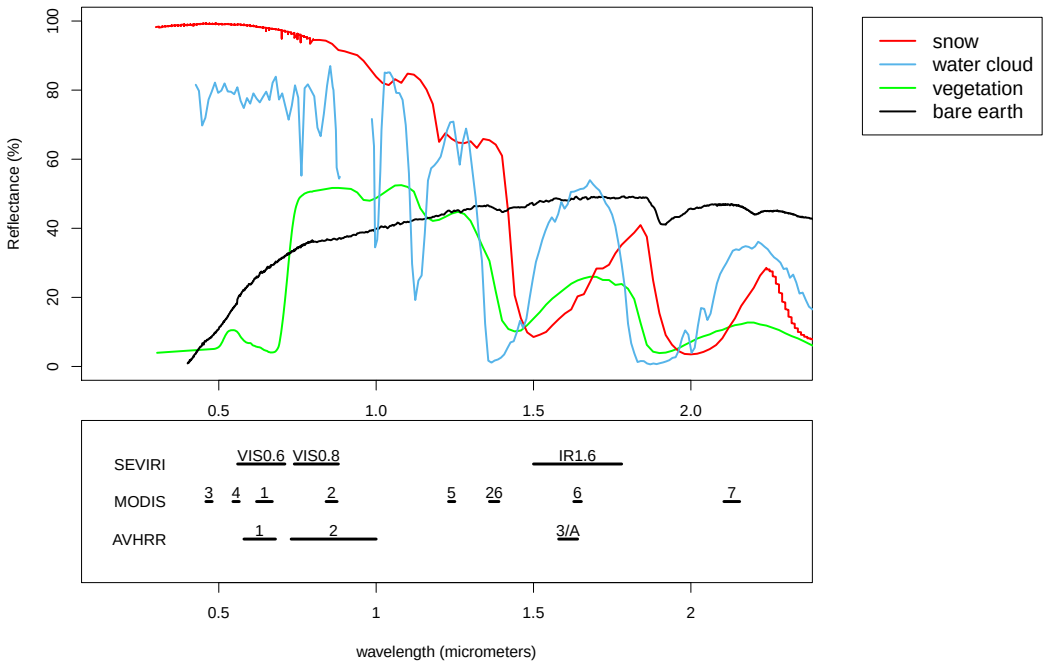

FIGURE 2.4. A rough sketch of the distributions of reflectances of common ground types and a cloud made of liquid water droplets ("water cloud") with the channels of meteorological satellite instruments mentioned in this thesis. The reflectances of ground types are based on Baldridge et al. (2009) and the reflectance of a water cloud is based on an arbitrary image taken on 12 June 2010 02:08 UTC near Barbourville, Kentucky, United States using the hyperspectral Hyperion instrument onboard the Earth Observing-1 (EO-1) satellite.

is reflected from the surface and clouds of the Earth, while the Earth itself emits long-wave (infrared) radiation. The atmosphere is, roughly again, transparent in the short wave and opaque in the long wave, letting the Sun in but not letting the warmth out. However, there are transparent regions in the infrared spectrum ("window regions") that are much utilized by satellite instruments.

The channels of imagers are selected for their physical properties, and thus enable different quantitative products to be constructed from the images. PAPER II takes a closer look at the cloud mask, a basic satellite product for discriminating cloud-filled pixels from others, while PAPER I and PAPER III examine snow analyses from pixels classified as snow-filled for the use of NWP or as a utility for duty forecasters.

The reflectances of common ground types (snow, vegetation, bare soil) and cloud made of liquid water droplets in the visible and near infrared part of the spectrum are sketched in Figure 2.4, along with the channels of satellite instruments used in this thesis. These ground types have distinct spectral signatures

and can be discriminated using combination of different channels. Note that the channels of different instruments are rather similar because of common interests and physical constraints (for example, the atmosphere is opaque in some parts of the shortwave spectrum).

In this example, we used visible and near-infrared channels, mainly because their characteristics are somewhat simpler than the cloud physics of the infrared channels. For example, twilight is the most problematic period for automatic processing of satellite images, as indicated in the results of Paper II. For discriminating between clouds and surface, the 3.6μm channel is useful both during day and night, but is of limited use in twilight. Lately, the 8.7μm channel of SEVIRI has proved to be a viable alternative for products used 24 hours a day (Derrien and LeGléau, 2005).

2.3 From satellite data to categorical products

In this thesis, the products derived from satellite data are classifications, not physical parameters such as cloud top temperature. For example, Paper II deals with cloud masks, which can get a value of 0 (false, no cloud) or 1 (true, cloud present). A classical and practical way to proceed is to collect samples of interesting phenomena and construct some kind of decision-making algorithm based on statistical properties of these samples. A simple example is presented in Figure 2.5, where histograms of snow, snow-free land, sea and water droplet cloud for different AVHRR/3 channels (Figure 2.5 a-c) are shown, along with a decision tree (Figure 2.5 d) generated by the `rpart` algorithm (Therneau and Atkinson, 1997). This decision tree works quite well for this limited data set and these classes. But it might not be adequate for practical purposes; in other words, there is a reason to doubt whether it generalizes well, especially since the difficult class of "clouds made of ice particles" is omitted.

In practice, the algorithm development is complicated by many factors. Most of the time, one pixel contains many different classes, especially at edges of large clouds and large snowy areas or when small clouds and snow patches are roughly same size or smaller than one pixel. Different weather conditions and satellite viewing angles can cause complications, because the characteristics of classes can change as a function of solar illumination or viewing angle, water vapor and aerosol content. Moreover, snow can get dirty, vegetation grows and withers away, soil can dry out or get moist again.

2.4 Decision-making for categorical products

Interestingly, all schemes for cloud masking or snow analysis discussed in Papers I, II and III use a conceptually simple method for decision-making, thresholding.

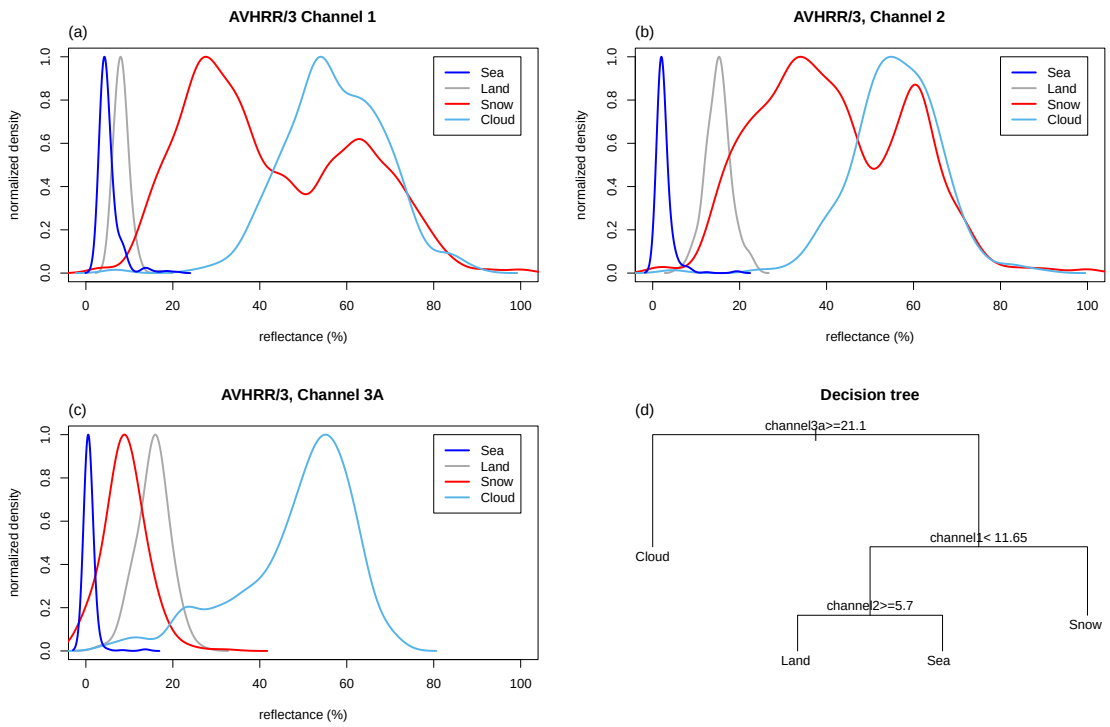

FIGURE 2.5. Histograms of samples of reflectances in AVHRR/3 (a) Channel 1, (b) Channel 2 and (b) Channel 3A for different land and cloud classes, collected by the Finnish Meteorological Institute. (The histograms have been normalized and smoothed. Note that because of smoothing, some reflectances have unphysical values less than zero.) (d) A simple decision tree constructed from the samples using `rpart` (Therneau and Atkinson, 1997). In each branch, the left link is chosen if the result is positive.

In thresholding, a pixel is classified into a certain class if a value constructed from its physical properties exceeds a certain threshold value or in practice many values for many tests. This is conceptually very similar to the decision tree constructed above using the `rpart` algorithm (Figure 2.5), but in practice most developers have shunned statistical methods for constructing their algorithms and have preferred to construct elaborate sequences of grouped threshold tests by hand. Of course, more mathematically advanced methods do not necessarily mean better overall results, but at least this kind of decision-making process might be more transparent and perhaps easier to maintain. More advanced methods were tried in the 1990's (e.g., Welch et al., 1992; Tovinkere et al., 1993), but they were not embraced by the meteorological satellite community. One reason is that those schemes used channels or channel combinations to construct static decision boundaries, but most current schemes use dynamic thresholds. Radiative transfer models are used simulate the satellite signal as it would have been observed in cloud-free but otherwise similar conditions and this is compared to the observed signal (Dybbroe

et al., 2005a). There has thus been great advances in how the thresholds have been constructed, but the decision-making has remained conceptually quite simple.

Even though current thresholding schemes give high-quality results, sometimes something more is needed. For example, from the decision-making perspective, the use of probability gives a very useful indication of confidence in a given classification. Even if no explicit probabilities are presented, there is always a hidden probability threshold, p_{thresh}, implicit in categorical products, as it is not clear what the developer meant when a pixel was classified as cloud. Was he more concerned about too much clouds ($p_{thresh} > .5$) or too little clouds ($p_{thresh} < .5$)? Therefore, if a pixel is classified as cloud, the user does not really know how much confidence can be given to the classification. Of course, all schemes give some indication of the quality of a classification, but these indicators are often quite *ad hoc* and it is difficult to know how to use this information further. The uncertainty of forecasts and the use of probabilities have long been discussed in meteorology (e.g., Murphy, 1998), lately in the context of ensemble forecasts (Richardson, 2000; Zhu et al., 2002). A decision-making process that takes into account the needs of the user has been considered in meteorology in the context of cost-loss model (Thompson, 1950; Savage, 1951; Thompson and Brier, 1955). However, according to Lazo (2010), its economic aspects have been given too much emphasis.

The lack of probabilities is one of the reasons why a separate cloud mask had to be created for Paper I. If the Satellite Application Facility on Support to Nowcasting and Very Short Range Forecasting (NWCSAF) cloud mask had given a reliable, or well-calibrated, probability of cloud, then only pixels with a sufficiently high no-cloud probability (90%, 95%, or even 99%?) could have been selected, depending on how many incorrect snow pixels are tolerated. Then only those pixels would have been processed further. It can be hoped that more operational systems will use probabilistic methods in the coming years.

3 Verification and validation of categorical products

Historically, in meteorology the work on evaluation of forecasts started from the famous tornado forecasts of Finley (1884). Nowadays, his reports of "percentages of verifications" are often shown pooled into one contingency table (Table 3.1) that suggests 96.6% of correct forecasts. However, it is easy to see that a simple constant "No Tornado" forecast would give a success rate of 98.2%, and the publication of Finley's article spurred other authors to publish what they considered better measures. This "Finley affair" is well documented in Murphy (1996). Most of the measures invented then have stood the test of time. Unfortunately, many have been rediscovered and renamed many times. For example, in a response to Finley, Doolittle (1888) published what is now better known as the Heidke Skill Score (HSS) after Heidke (1926).

In meteorology, the terminology used can be somewhat ambiguous. Verification can be defined as the process of assessing the quality of forecasts (Wilks, 2006), but the term validation is used as well. However, there has been no authoritative definition of differences between verification and validation in meteorology. In the context of software engineering, Boehm (1979) defined verification as activities to establish the truth of the correspondence between a software product and its specification ("Am I building the product right?") and validation as activities to establish the fitness or worth of a software product for its operational mission ("Am I building the right product?"). If we replace "a software product" with "forecasts" and "its specification" with "observations", this is close to what most meteorologists would agree, if pressed. On the other hand, the use of "verification", instead of "evaluation" or "assessment", may be a historical accident [following the example of Finley (1884)], so perhaps the difference of terms should not be emphasized much[1]. Outside meteorology, the term "verification" can have different meanings in different fields and contexts. For example, in epistemology, verification is the limiting case of confirmation: a piece of evidence verifies a hypothesis just in case it conclusively establishes that hypothesis as true (Kelly, 2008). This aim maybe be too ambitious for our purposes.

In this introduction, terms "validation" and "verification" are used interchangeably. In Papers I, II, III and IV, the term "validation" rather than "verification" was used, mainly because these articles did not consider forecasts but different types of satellite- and radar-based products.

This thesis deals with categorical variables, and even then we only consider

[1]See the informal discussion on this and the previous sentence in the *vx-discuss – Verification discussion group* mail list in January 2011 (http://mail.rap.ucar.edu/mailman/listinfo/vx-discuss). Hopefully the next edition of Jolliffe and Stephenson (2003) will have something to say about this matter, too.

Table 3.1 Results of Finley (1884) pooled into one contingency table (Murphy, 1996).

Tornados	Tornados observed	
forecasted	Tornado	No tornado
Tornado	28	72
No tornado	23	2680

Table 3.2 Contingency table of the comparison between forecasts and observations or any two analyses or products. The symbols a-d represent the different number of cases (or, for example, pixels) observed to occur in each category.

	Observation x	
Forecast $\hat{x}$	1	0
1	a (Hit)	b (False Alarm)
0	c (Miss)	d (Correct Rejection)

binary or dichotomous cases, that is, whether snow (Paper I and Paper III) or cloud (Paper II) was present or not. Thus, no methods for continuous or probabilistic forecasts (or products) are presented here. This chapter presents the verification measures used in this thesis and shows how to calculate confidence intervals for them. Last, a new source of observations for evaluation is discussed.

3.1 Verification measures for binary data

The difference between the forecast or the analysis, $\hat{x}$, and the observation, the ground truth or the baseline, x, can be shown in a contingency table, as in Table 3.2. Here we only consider 2×2 contingency tables.

Paraphrasing Tolstoy, all products are correct in the same way, but every incorrect product is incorrect in its own way, and one measure alone cannot represent all information of the contingency table. Numerous measures have been proposed with slightly different names. This thesis follows the nomenclature of Jolliffe and Stephenson (2003).

Perhaps the most intuitive measures are the Proportion Correct

$$\mathrm{PC} = \frac{a+d}{a+b+c+d} = \frac{a+d}{n}, \tag{3.1}$$

and the bias

$$\mathrm{B} = \frac{a+b}{a+c}. \tag{3.2}$$

PC measures the accuracy, i.e., the fraction of items classified correctly. The best value for PC is one and the worst zero. B is the ratio of the number of the forecasts to the number of the observations. The best value for B is one; less than one means underforecasting and more than one overforecasting.

From the contingency table, the following conditional probabilities are intuitively obvious:

Hit rate: the probability of the event was forecasted when it did occur

$$H = P(\hat{x} = 1|x = 1) = \frac{a}{a + c} \tag{3.3}$$

False alarm rate: the probability of the event was forecasted when it did *not* occur

$$F = P(\hat{x} = 1|x = 0) = \frac{b}{b + d} \tag{3.4}$$

False Alarm Ratio: the probability of the event did *not* occur when it was forecasted

$$\mathrm{FAR} = P(x = 0|\hat{x} = 1) = \frac{b}{a + b} \tag{3.5}$$

For the sake of symmetry, we could also define the probability of the event did *not* occur when it was *not* forecasted, $P(x = 0|\hat{x} = 0) = \frac{d}{c+d}$, but in practice it is not widely used, especially in the case of forecasts, where the non-existence of non-forecasted event is perhaps not very interesting. In other situations it has its uses: For example, when comparing cloud-masks, it is useful to know the number of correct cloud-free cases (Derrien and LeGléau, 2005; Dybbroe et al., 2005b). However, Stephenson (2000) shows how this (and FAR) can be expressed as a function of H, F and B, so its use (and the use of FAR) is somewhat redundant, but it can still be useful.

Similarly, Jolliffe and Stephenson (2003) show how all measures can be expressed as a function of H, F and the base rate

$$s = P(x = 1) = \frac{a + c}{n}. \tag{3.6}$$

For example, PC is then

$$\mathrm{PC} = \mathrm{H}s + (1 - \mathrm{F})(1 - s). \tag{3.7}$$

This might make us think that if s tends to unity, PC can give us overoptimistic results. A constant "yes" forecast has a perfect H and bad (high) F, but the effect of F may be obscured if s is high enough. The same happens if $1-s$ tends to unity, as with Finley's tornado forecasts where the no-tornado cases dominated, and a constant forecast of no tornado would have given even better results than those reported by Finley. Here a useful term is *equitability* (Gandin and Murphy, 1992).

A measure is not equitable if a forecaster can *hedge* his forecasts by favoring some events at the expense of others. PC is clearly not equitable; think "No tornado" forecasts in Finley's case. An often used refinement of PC is the Critical Success Index

$$\mathrm{CSI} = \frac{a}{a+b+c}. \tag{3.8}$$

However, if the probability of an event is p_1, then a constant "yes" forecast will give an expected score of p_1, while a constant "no" forecast will give an expected score of 0. But according to Gandin and Murphy (1992), the first requirement [and for Hogan et al. (2010), a sufficient requirement] for all equitable verification measures is that they should award the same expected score for both random and constant forecasts. So CSI is unfortunately still not equitable.

A way forward is to use skill scores that are equitable. A general skill score (SS) for a measure S is

$$\mathrm{SS} = \frac{S-E}{S_{perf}-E}, \tag{3.9}$$

where S_{perf} is the perfect value for S and E is the reference value; for example, a climatological value or a persistence, that is, the same value as at the present. For categorical forecasts, the reference is often chance agreement or the proportion correct by chance. The much-used HSS is based on PC, so $S_{perf} = 1$. The assumption for E is that forecast and observation probabilities are independent, so

$$\begin{aligned} E &= P(\{x=1 \text{ and } \hat{x}=1\} \text{ or } \{x=0 \text{ and } \hat{x}=0\}) && (3.10)\\ &= P(x=1)P(\hat{x}=1) + P(x=0)P(\hat{x}=0) && (3.11)\\ &= \left(\frac{a+c}{n}\right)\left(\frac{a+b}{n}\right) + \left(\frac{b+d}{n}\right)\left(\frac{c+d}{n}\right). && (3.12) \end{aligned}$$

Actually, not all the commonly used skill scores in meteorology fit comfortably into the framework of (3.9). For example, Wilks (2006) shows that E for the Peirce Skill Score (PSS) [which is called the Hanssen-Kuiper Skill Score (KSS) in Paper II] in the nominator is the same as (3.12), but in the denominator it is

$$E = \left(\frac{a+c}{n}\right)\left(\frac{a+c}{n}\right) + \left(\frac{b+d}{n}\right)\left(\frac{b+d}{n}\right). \tag{3.13}$$

Here PSS assumes that the (sample) probabilities are equal [i.e., $P(x=1) = P(\hat{x}=1)$ and $P(x=0) = P(\hat{x}=0)$]. In the original derivation of PSS (Peirce, 1884), the framework of (3.9) was not used.

PSS can be calculated simply as

$$\mathrm{PSS} = H - F, \tag{3.14}$$

while for HSS the useful formula for the 2×2 contingency table is

$$\mathrm{HSS} = \frac{2(ad-bc)}{(a+c)(c+d)+(a+b)(b+d)}. \tag{3.15}$$

In addition to HSS and PSS, the meteorological community has derived other scores and will probably deliver still others in the future. Useful lists with discussion can be found in, e.g., Wilks (2006) and Hogan et al. (2010). In recent years, much of this research has considered verification of rare events (e.g., Ferro and Stephenson, 2011). These new scores could be applicable for snow studies in summer where the amount of snow is very low, and should be considered in future studies continuing Paper I and Paper III. Note also that equitability is just one of the properties for evaluating the verification measures in the process often called *metaverification* (Murphy, 1996). These properties are discussed in, e.g., Stephenson and Jolliffe (2003).

The general assumption for the measures presented above is that the data consist of points that are independent and have no spatial or temporal correlation. So, for grid data that break these assumptions, the simple measures presented above may not be enough. Especially for precipitation forecasts, the methods of taking into account the spatial and temporal displacement of precipitation are in vogue (e.g., Zingerle and Nurmi, 2008; Gilleland et al., 2010). For our data, these methods might be applicable to snow data in Paper I and Paper III. However, snow is much more conservative in its movements, snow accumulates as the winter progresses and the snow edge moves roughly from north to south or up the mountain slopes. So it is rare that sizable separate regions of snow would be classified in different locations by different algorithms and therefore such methods were not used (but might be reconsidered in future studies). However, how to take spatial and temporal correlation into account is discussed later in this chapter, when the calculation of confidence intervals is discussed.

3.2 Measures used outside the meteorological community

The task of evaluation is, of course, important also outside meteorology, and similar problems have been discussed using different terminology. For example, the contingency table in Table 3.2 is often called an error matrix (e.g., Stehman, 1997) or a confusion matrix (e.g., Ripley, 1996) in pattern recognition and in remote sensing.

Stephenson and Jolliffe (2003) discuss the work in statistics, economics, environmental sciences, and medical studies [for medical studies, see also Pepe (2003)]. A viewpoint from the remote-sensing perspective is given, for example, by Stehman (1997), Congalton and Green (1998), and Liu et al. (2007). Interestingly, in the commonly used textbooks of pattern recognition (Ripley, 1996; Bishop, 2007; Hastie et al., 2009) only the error rate (1-PC) is used explicitly. Only Duda et al. (2001) discuss some other measures in the context of signal detection theory.

Another example of an independent discovery is kappa (Cohen, 1960), which is much used in psychometry and, for example, in the remote-sensing community [being introduced there by Congalton et al. (1983)]. This measure is identical to HSS. In addition, outside meteorology other definitions for E that would fit in the framework of (3.9) are used. A very simple alternative (Bennett et al., 1954; Foody, 1992) is $E = 1/k$, where k is the number of categories ($k = 2$ in this thesis). This can be improved; for example, Scott (1955) suggested using the average (sample) probabilities of x and $\hat{x}$ in calculating E (this is known as Scott's π).

In meteorological literature, HSS has been criticized from time to time, especially for its dependence on s (e.g., Hogan et al., 2009) and if more than two categories are used (Livezey, 2003). The use of kappa has also been criticized (e.g., Brennan and Prediger, 1981; Zwick, 1988; Foody, 1992; Krippendorff, 2003) because it gives better values if the marginal distributions differ, i.e., $P(x = 1) \neq P(\hat{x} = 1)$ and $P(x = 0) \neq P(\hat{x} = 0)$. Krippendorff's α (Krippendorff, 1970), based on Scott's π, is one measure devised to address these shortcomings. This problem is not much discussed in the meteorological community, perhaps because even if kappa and HSS are mathematically identical, their use and interpretation can be different. What are usually called *accuracy* and *precision* in the physical sciences are often called *validity* and *reliability* in the social and behavioral sciences. In the social sciences, kappa is often used for measuring the reliability/precision of the agreement between different raters. This is different from the use of HSS in meteorology, where it is used to measure the validity/accuracy of forecasts. In measuring reliability, it is not prudent to penalize for the same marginal distributions, but in measuring validity the fact that marginal distributions are equal is not enough[2]. Furthermore, both Scott's π and Krippendorff's α give different results for different constant forecasts (like CSI above), and therefore are not equitable and not useful as meteorological verification measures.

3.3 Quantifying uncertainty of measures

Having calculated some measures, for example, skill scores for two algorithms, the next question is whether we can say that one of them is really better than the other (or at least gets better scores). In other words, is the difference between scores really meaningful or, even more specifically, is the difference between scores different from zero.

However, there is more than one way to do this. There are two major paradigms over the interpretation of probability, the frequentist and Bayesian

[2]Peirce (1884) writes "The second witness may know *how often* he ought to answer 'yes'; but I give him no credit for that, because he is ignorant *when* he ought to answer 'yes.' " (original emphasis)

approaches. The frequentist approach is usually the "default", the one that is most often given in textbooks of statistics, but the Bayesian approach has gained ground in many fields, and it may be the main paradigm in some fields [in pattern recognition, see Bishop (2007), in other fields, see, e.g, Russell and Norvig (2010) or Jeffrey (2004)]. An informal, succinct definition is that frequentists want to know what happens if an experiment is repeated many times, while Bayesians want to know what new information a new experiment will bring (D'Agostini, 2003). Hacking (2001) gives a clear introduction to these different interpretations. De Elía and Laprise (2005) discuss them in the meteorological framework.

Still, almost all of the work in this thesis (except parts of Paper V) was done under the frequentist approach. The aim was to determine whether the conclusions are not merely random noise and for that purpose the frequentist approach is sufficient. In this case, when confronted with a low p-value, it is enough to conclude that the null hypothesis was wrong or something extraordinary happened.

Confidence intervals (CIs) are perhaps a more intuitive way to assess uncertainty than p-values [see Jolliffe (2007) in meteorology and Foody (2009) in remote sensing]. The CI gives the interval within which the true value of the parameter falls $p\%$ of time (Usually p is 95%, but this is just a convention and other values can be used, though they would need more explaining). This is often implied meaning "when the experiment is repeated over and over", but it is not necessary to repeat the same experiment, only to follow the method of constructing the CI while the data sets and parameters vary over the time (Hacking, 2001; Wasserman, 2003).

For proportions or sample estimates of probabilities $\hat{p}$, there is the coarse but well-known Wald CI

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \tag{3.16}$$

where $z_{\alpha/2}$ is a quantile of the standard Gaussian distribution. For a 95% CI, $z_{\alpha/2} = 1.96$. This is based on the normal approximation, and is accurate only if a reasonable amount of data is available. Agresti and Coull (1998) suggest a more accurate formula that works even for small data sets

$$\left(\hat{p} + \frac{z^2_{\alpha/2}}{2n} \pm z_{\alpha/2}\sqrt{[\hat{p}(1-\hat{p}) + /n + z^2_{\alpha/2}/4n]/n}\right)/(1 + z^2_{\alpha/2}/n). \tag{3.17}$$

As expected, these methods give different results when the number of observations is small, depending on $\hat{p}$ (Figure 3.1 a), but the difference diminishes as n grows, diminishing slower for the more extreme values of $\hat{p}$.

Contrary to proportions, the distribution of HSS is not known, and there is no analytical formula for CI. However, different approximations can be derived. Cohen (1960) gives a simple approximation for the standard error

$$\sqrt{\frac{\mathrm{PC}(1-\mathrm{PC})}{n(1-E)^2}}, \tag{3.18}$$

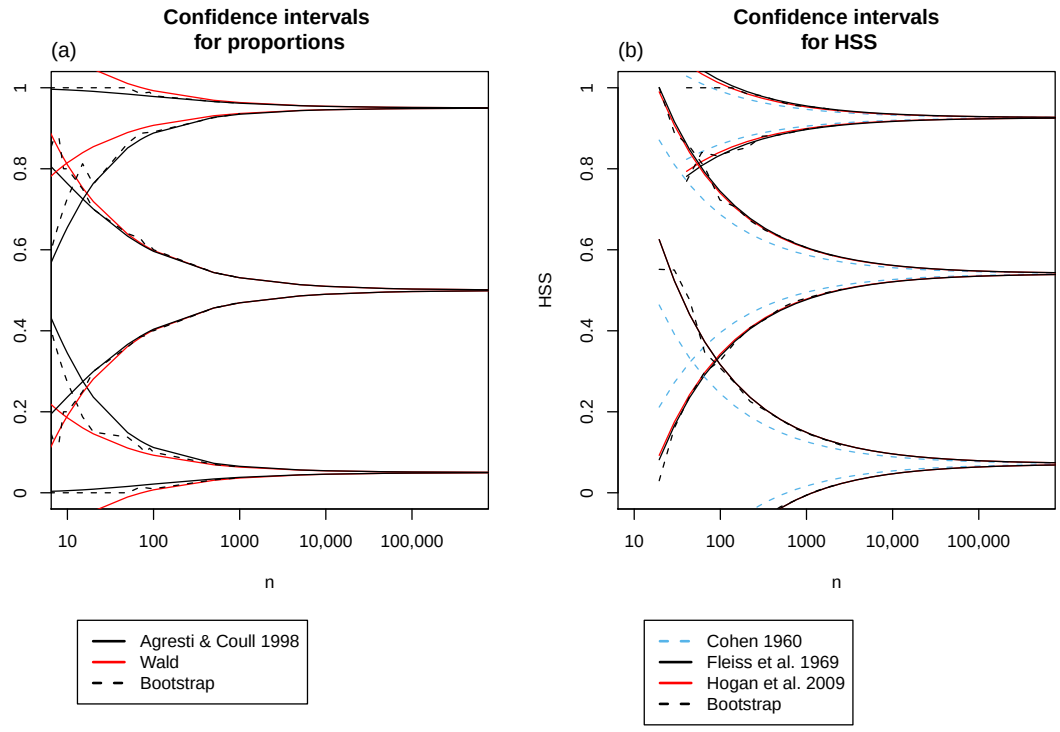

FIGURE 3.1. Confidence intervals for fictive data by different methods for (a) proportions and (b) HSS as a function of n.

shown here in the nomenclature of this thesis. Fleiss et al. (1969) give a much more complex formula for an approximate large sample variance using the delta method (e.g., Wasserman, 2003). The delta method is also used, for example, by Congalton and Green (1998), but without acknowledging Fleiss et al. (1969). Hogan et al. (2009) show another approximation with somewhat different assumptions. The CI obtained with (3.18) is much narrower than that obtained with the other two formulas, while the results from these two formulas are rather identical (Figure 3.1 b). Like (3.16), all three approximations give impossible values above unity for small n.

Another way to construct the CI is to use resampling methods, for example, bootstrap.

3.4 Bootstrap for quantifying uncertainty

The bootstrap is a resampling method, where the sample is resampled with replacement. A standard reference is Efron and Tibshirani (1994), which can be supplemented, for instance, with Chernick (2007). An overview of meteorological applications can be found in Wilks (2006). A practical example of how CIs are constructed is shown in Figure 3.2. The lure of bootstrapping is that the

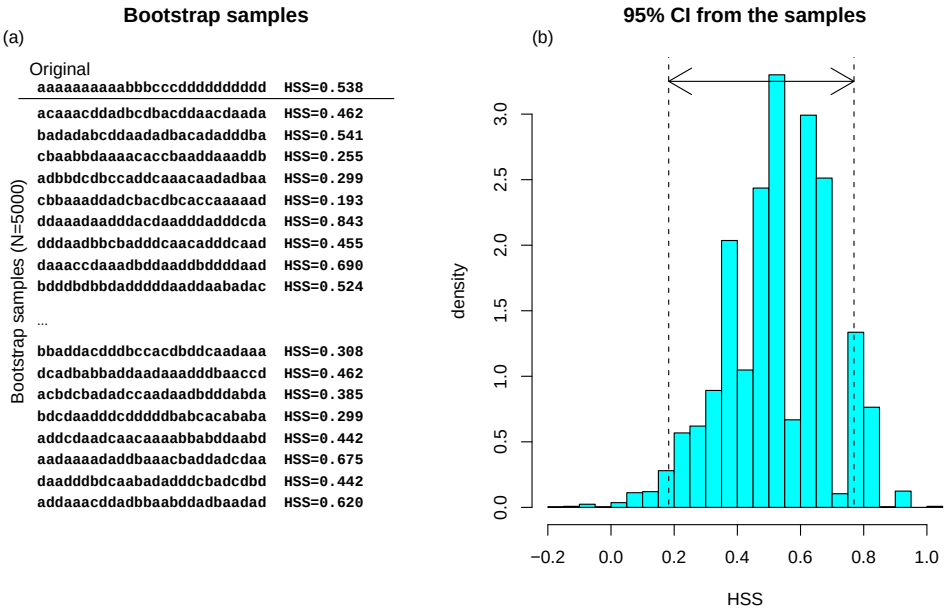

FIGURE 3.2. CI of the HSS for fictive data when $a = 10, b = 3, c = 3, d = 10$ (a) Bootstrap samples. Note that the order of items is arbitrary. (b) the histogram of samples and CI from quantiles of 2.5% and 92.5%.

user can employ a relative simple computational algorithm to solve problems that would otherwise require quite advanced knowledge of mathematics and statistics. Bootstrapping is compared to other methods in Figure 3.1, which shows that the results are comparable to others. Especially interesting is Figure 3.1 b, where, compared to two approximation methods discussed above, the bootstrap CI for the HSS is the only method that does not give values exceeding unity for small data sets. Of course, bootstrapping is computationally much more demanding than any formulas, but often those formulas are not available.

In this thesis, bootstrapping was used for some extent in PAPER I to construct the CI and in PAPER V to investigate whether synoptic situation between clusters of fog cases was significantly different from each other, but the most extensive use of bootstrapping was made in PAPER II.

In PAPER II, bootstrapping was used to calculate CIs under spatial and temporal correlation between data points. Then simple bootstrap (as in Figure 3.2) would give overly narrows CIs. The method of Hamill (1999) was used for spatial correlation, while for temporal correlation, the moving-block bootstrap was used. The method of Wilks (2006) was applied for determining the block length L. Some parts would require more detailed study: Can we really separate the spatial and temporal correlations and treat them individually? Moreover, the method of Wilks (2006) is an empirical one for first-order autoregressive processes. This assumption was accepted in PAPER II, but in future studies this should be con-

sidered more carefully along with other methods for calculating L (e.g, Lahiri, 2003).

Furthermore, in all papers, the percentile method, a very simple method for constructing the CI from bootstrap samples, was used. Other methods are presented in the literature, but the percentile method was deemed to be adequate for our purposes. In future studies this should be assessed better.

3.5 A somewhat surprising source for validation

Especially in Papers I and III, and to some extent in Paper II, a recurrent problem was the lack of independent observations. Paper IV suggests the use of photos from Internet users — a somewhat surprising source — for validation purposes. This was motivated by the dramatic increase in the social use of the Internet that has occurred in the past few years. An obvious way to use social networks is to engage stakeholders directly. This has been done to great effect with Finnish storm spotters (Tuovinen et al., 2009), and the Internet helps to make this feasible. A recent IGARSS conference (GRS-S Newsletter, 2010) had a special session for this kind of work, under the heading "Community remote sensing". Other buzzwords for this kind of work include collective intelligence, crowdsourcing, folksonomy, group intelligence, and social information processing. However, it remains to be seen how many of these words are still in use ten years from now.

The approach in Paper IV was different. The idea was to use data that are available, but were not originally meant to be used for scientific purposes. Therefore, there is a certain stalking aspect in this work. This approach provides more data, but those data are of variable quality. Of course, the usefulness of data is not only a function of quantity but also of interpretability. Metadata for photos from Flickr, a widely used web service for sharing photos and videos, were reasonably interpretable. Paper IV showed that they are also of reasonable quality and their use for case studies can be warmly recommended. These photos, especially with GPS-based location and temporal information, can readily be used to find evidence of the existence of some event, which is often enough for case studies. For more quantitative studies, the problem of finding also evidence of non-existence of an event can be a problem. Of course, this could be avoided if we use the photos themselves and not the tags from metadata, but this approach is much more labor-intensive or needs advanced computer vision methods. An interesting future research topic, perhaps?

It can be said that the jury is still out. It is possible that the amount of freely available data will increase as the Internet grows, but it is equally possible that access to most personal data will restricted in future, and thus will not be available for independent researchers. Only time will tell.

4 Analyzing differences between products

Paper I and Paper II used one of the data sets as the truth. This is feasible if one of the data sets is deemed better than, or at least independent of, the others. However, in Paper III no single snow product was independent as no product was constructed using data not used by others (Figure 4.1) and nor could any of the data sources be used as the ground truth. It was then necessary to conclude that none of the data sets represents the truth and that only consistency or agreement between products can be assessed.

One way forward is to consider the verification measure as the distance from a data point (a product, a forecast, etc.) to the ground truth. And if the independent truth is not available, as in Paper III, it is still possible to measure the distances between data points and to collect these distances into a "similarity matrix". All the information is in this similarity matrix, but visualizing the matrix would be useful. For this purpose, this chapter explores the Sammon mapping, a multidimensional scaling (MDS) method. Related to MDS is clustering, where similar data points are divided into groups or clusters. Clustering was used in Paper III and Paper V. For both Sammon mapping and clustering, data from Paper II are used as an example.

4.1 Visualizing with multidimensional scaling

MDS methods (e.g., Ripley, 1996; Duda et al., 2001; Venables and Ripley, 2002; Borg and Groenen, 2005) can be used to visualize high-dimensional data in two dimensions, so that the original distances are more or less retained in the new dimensions.

The distances between all data points are required for MDS. Technically, these do not have to be real distances, only mathematically less strict dissimilarities,

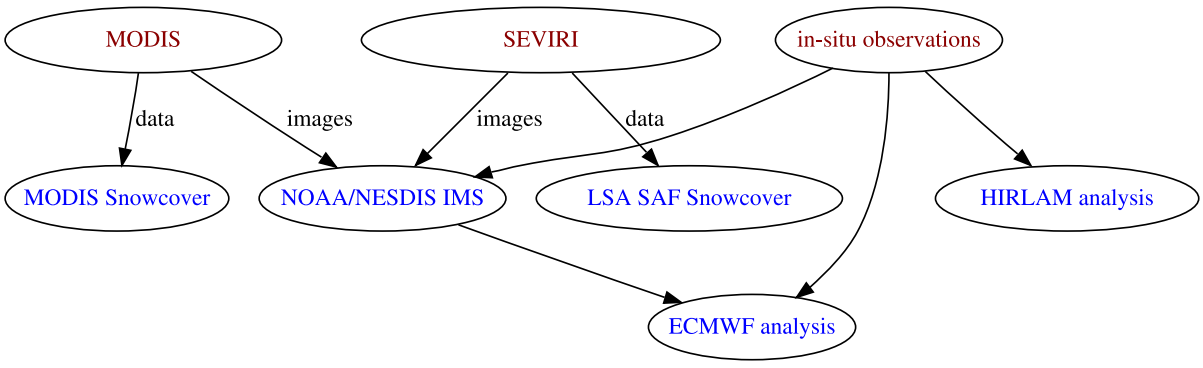

Figure 4.1. Data sources (red) and final products (blue) in Paper III.

nonnegative symmetric numbers. Many verification measures, such as PC, CSI and HSS, if subtracted from unity, can be used as dissimilarities. For example, 1-HSS was used as the dissimilarity in Paper III. However, other measures, such as PSS, H, F and FAR, do not give the same values if forecasts and observations change places in Table 3.2 and cannot be used. Kaufman and Rousseeuw (1990) and Borg and Groenen (2005) discuss distances and dissimilarities in more detail. This thesis has considered only categorical verification measures, but continuous measures could also be used in future studies.

MDS has been seldom used in meteorological literature. A more familiar method is principal component analysis (PCA) (e.g., Jolliffe, 2002), which is sometimes used to reduce the dimensionality of data or to compress the data (Huang and Antonelli, 2001) and reduce the noise (Antonelli et al., 2004). However, MDS and PCA are closely related. For example, principal coordinates analysis (PCO), the simplest MDS, gives results identical to those of PCA if the Euclidean distance is used. However, PCO is more general and not restricted to the Euclidean distance. The results of PCO can still be improved by minimizing a cost or stress function. For Sammon's non-linear mapping (Sammon, 1969), or Sammon mapping for short, the stress function to be minimized is

$$\frac{1}{\sum_{i<j} d_{ij}} \sum_{i<j} \frac{(d_{ij} - \delta_{ij})^2}{d_{ij}}, \tag{4.1}$$

where d_{ij} is the original dissimilarity and δ_{ij} the new distance between data points. Other possible stress functions are discussed in Ripley (1996), Duda et al. (2001), and Borg and Groenen (2005).

In addition to Paper III, where different snow analyses were compared, we could have used Sammon mapping to assess cloudmasks in Paper II (Figure 4.2). The relative position of cloudmasks in Sammon mapping is more easily interpreted in August than in February. In February, NWCSAF/MSG is nearer to ceilometers, while the other two schemes are much further and seem to disagree with others. On the other hand, in August, masks based on SEVIRI [i.e., NWCSAF/MSG and Meteorological Products Extraction Facility (MPEF)] are close to each other, which is understandable as they are based on the same instrument. Between SEVIRI-based instruments and ceilometers, there is the AVHRR-based mask from the NWCSAF/Polar Platform System (PPS), which received better scores than SEVIRI-based masks when compared to ceilometers. Still, all satellite-based products are nearer to each other than to ceilometers; that is, scores between satellite products are higher than scores between ceilometers and satellites. This may indicate that satellite cloud-masks and ceilometer detect clouds in two distinctive ways. An interesting question is how much of this difference stems from the inability of ceilometers to detect high clouds. Had we got access to cloud-type and cloud-top-height information from NWCSAF/MSG and MPEF, we could have assessed this. Unfortunately, the archiving of products is not at the level one would

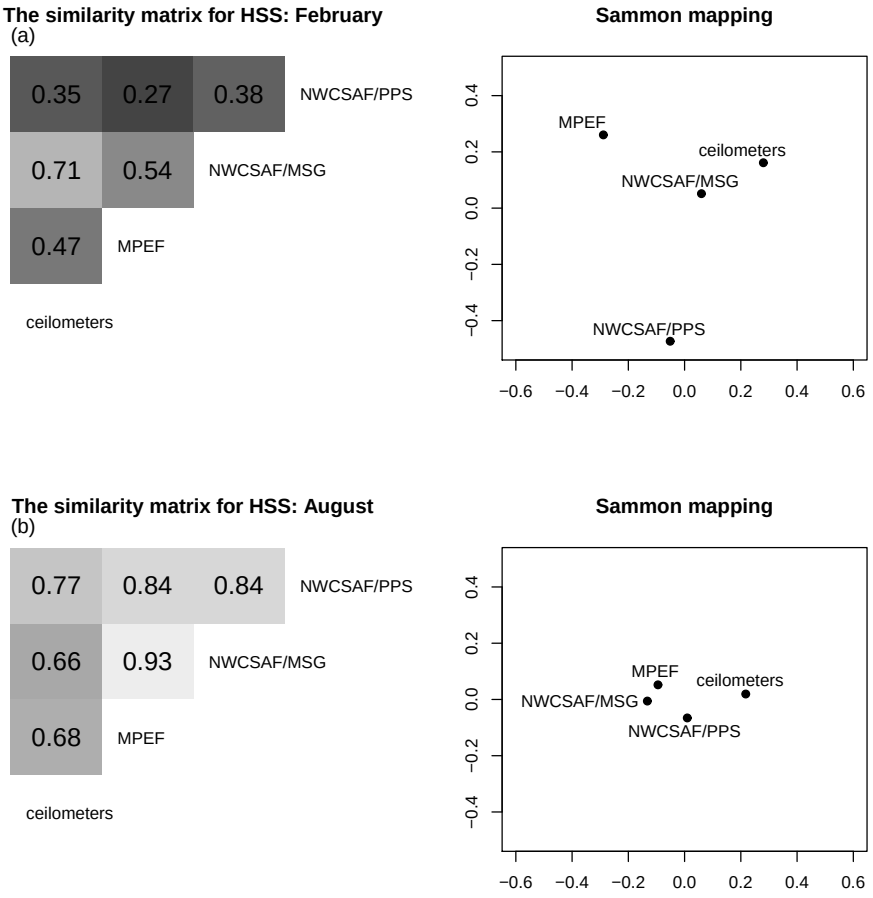

FIGURE 4.2. Sammon mapping of relations of different data sources from PAPER II in (a) February (b) August.

like it to be. Reprocessing of MODIS products provided by NASA is ongoing work, but has not been performed by EUMETSAT's SAF framework. Hopefully, this will change in the not-too-distant future. (Reprocessing of old data, when new algorithms become available, would be also useful for the work of PAPER I and PAPER III.)

MDS is not perfect, however. For one, the CIs can be difficult to compute. This is discussed by, for example, Meulman and Heiser (1983) and Kiers and Groenen (2006), but apparently their ideas have not attracted much attention. For example, it seems that no widely-used standard implementations are available for their methods. In addition, MDS does not give us much indication as to where we should try to find the truth, or at least the best solution.

For future studies, it would useful once more to reconsider the purpose of MDS. Are we assessing the accuracy or just comparing different items, and if so, are we working in the reliability/precision framework and should we then take the

critique of kappa seriously? However, at least for our example in Figure 4.2, use of Scott's π or Krippendorff's α makes no visible difference (not shown).

In Paper III, and in Figure 4.2, MDS was used for a quite restricted number of data points, when it is easy to find a good solution that converges to near zero stress. Sammon mapping becomes impractical when the number of data points is large. The complexity of calculation is $O(n^2)$ [see, e.g., Duda et al. (2001) for the notation], so the memory and time needs grow fast[1]. Different methods are therefore necessary when the number of data points becomes much larger.

4.2 Finding groupings with clustering

Often we want to know if the data can be divided into groupings or clusters. MDS can be used to visually assess the clusters, but quantitative methods of clustering should be used in order to find these clusters explicitly. These methods are discussed, for example, by Ripley (1996), Wilks (2006), Bishop (2007) and Hastie et al. (2009).

Clustering was performed on a small scale in Paper III, where the data set was divided into two using the simple k-means algorithm, and on a larger scale in Paper V, where Autoclass, an unsupervised Bayesian classification system (Cheeseman and Stutz, 1996), was used to obtain insight into different fog situations. Clustering was also considered for Paper I, but in the end, the results were largely determined by latitude and land use, so no new insights were obtained by tentative cluster analysis.

While Sammon mapping is a simple mapping from a high-dimensional coordinate system to a lower-dimensional one, advanced methods for clustering, such as Autoclass, explicitly model the data. In most clustering, the task is to find discrete latent variables (the clusters). In PCA and independent component analysis (ICA) (Hyvärinen et al., 2001), these latent variables are continuous (Bishop, 2007).

MDS and clustering are closely related, as both strive to answer similar questions about the interesting features of data. On a practical level, the same dissimilarity matrix constructed for MDS can often be used as input for clustering. For example, in Figure 4.3, dendrograms are constructed from the similarity matrixes of Figure 4.2. Conclusions similar to those for Figure 4.2 can be drawn, but the more divergent conditions in February are not as evident in Figure 4.3 (a). In fact, dendrograms can be considered as a one-dimensional display of similarity

[1]On a typical desktop PC, a few thousand points is manageable, but more than ten thousand points is not feasible. And more informally speaking, it seems that if there are more than about 1000 data points, it can be hard for Sammon mapping to improve on the starting point calculated by PSO. However, this can depend on the implementation of the algorithm. Here, as in Paper III, the implementation of Venables and Ripley (2002) was used.

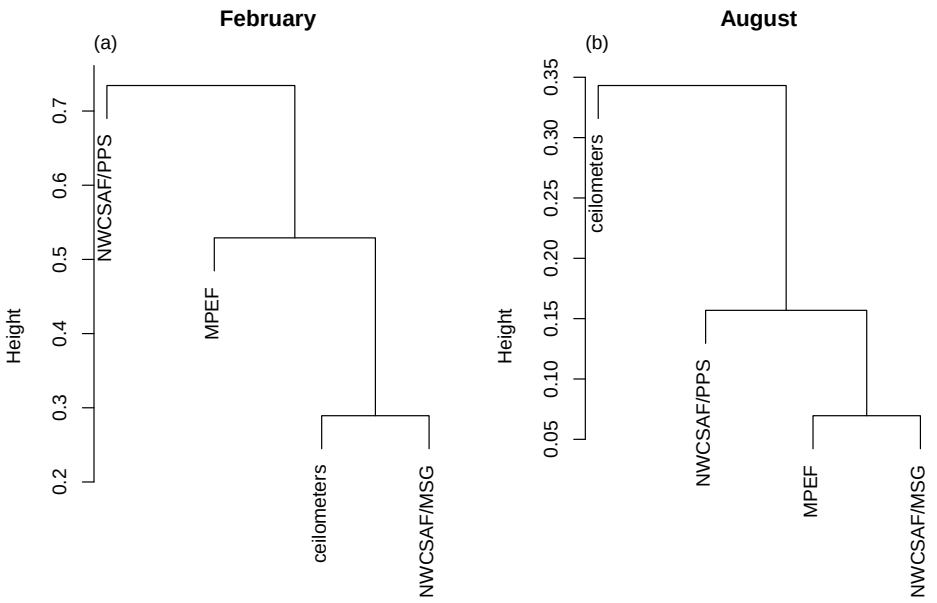

FIGURE 4.3. A hierarchical clustering based on the similarity matrix of Figure 4.2 for (a) February and (b) August. The complete-linkage method (e.g., Ripley, 1996) was used.

(Ripley, 1996), but two dimensions are required in order to get a better look at the data here.

One open question is whether the clusters represent real physical phenomena or are just useful tools for making sense of complicated data. For the former, we would continue to search for more sophisticated, and perhaps computationally more expensive, algorithms; while for the latter, we might be content with simpler, computationally less demanding algorithms (such as k-means). This was briefly discussed in PAPER V, and would be an important part of further studies.

5 Conclusions

In this chapter, the main results are reiterated and possible future directions are outlined.

5.1 Main results

Different algorithms for satellite products generate different results; sometimes the differences are subtle, some times all too visible. In addition to differences between algorithms, the satellite products are influenced by physical and meteorological processes and conditions, such as diurnal and seasonal variation in solar radiation (especially in Paper II and Paper III), topography (Paper III), and land use (Paper I).

In Papers I and II, bootstrapping, a statistical resampling method, was used to construct confidence intervals that can be used to assess the uncertainty in the evaluation results. Especially in the presence of spatial and temporal correlation, bootstrapping provides a useful tool for constructing the confidence intervals. However, sometimes no ground truth is available for evaluation. In Paper III, the Sammon mapping, a multidimensional scaling method, was then used to visualize the differences between different products.

Paper IV discussed how data provided by general public can be used as an interesting new source for validation. In general, the needs of a particular research project drive the requirements, for example, for the accuracy and the timeliness of the particular data and methods.

The results of Paper V suggest that a combination of the subjective analysis and a clustering algorithm (e.g., the AutoClass system) could be used to construct climatological guidelines for forecasters.

5.2 Future directions

In this thesis, we have tried to assess how the different conditions influence the results mainly by dividing the data into small segments and calculating the confidence interval from bootstrap samples. But, for example, in Paper II, the fact that lidars could not see high clouds may be too complicated to be assessed in this way. It might be interesting to construct statistical models that describe the contingency table so that it would be easier to take other, extra parameters into account. This is the way forward suggested, for example, by Agresti (2007), who frowns upon the use of summary indexes such as HSS, and encourages the use of statistical models that describe the structure of agreement and disagreement. (Here again we should consider the suitability of these methods for the assessment of accuracy.) Possible approaches might be Bayesian nets (e.g., Wasserman,

2003; Bishop, 2007) and hierarchical modelling (e.g., Wasserman, 2003; Gelman and Hill, 2007) of loglinear models (e.g., Wasserman, 2003; Agresti, 2007). The Bayesian framework for these approaches and for verification in general would be interesting. Evaluation with Bayesian methodology is discussed in Broemeling (2009), with many examples from medical studies. In addition, our problem of no ground truth has been discussed in medical studies as the absence of "a gold standard" (e.g., Pepe, 2003; Rutjes et al., 2007) and might merit a closer look if this part of thesis is pursued further.

Finally, another obvious possibility for future studies is to continue the work started in Paper V, which considered the occurrences of fog at airports as independent points in time, disconnected from their spatial and temporal extent and evolution. How different statistical methods, such as clustering, can help forecasters or scientists to understand the different weather phenomena remains an interesting problem. In addition, lately the amount of manual observations of professional observers has decreased as human observations are being replaced by automatic observations. Automatic in-situ observations complemented with remote-sensed observations open up new possibilities, but what kind of subtle atmospheric phenomena, only observable by humans at least now and for coming decades, would then be lost? Digitizing of old observation records might help us to find some hidden gems.

References

Agresti, A., 2007: *An Introduction to Categorical Data Analysis*. 2d ed., Wiley-Interscience.

Agresti, A. and B. A. Coull, 1998: Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, **52 (2)**, 119–126, doi:10.2307/2685469.

Antonelli, P., et al., 2004: A principal component noise filter for high spectral resolution infrared measurements. *Journal of Geophysical Research*, **109 (D23102)**, doi:10.1029/2004JD004862.

Bader, M., J. Forbes, J. Grant, R. Lilley, and A. Waters, (Eds.) , 1995: *Images in Weather Forecasting: A Practical Guide for Interpreting Satellite and Radar Imagery*. Cambridge University Press.

Baldridge, A. M., S. J. Hook, C. I. Grove, and G. Rivera, 2009: The ASTER spectral library version 2.0. *Remote Sensing of Environment*, **113 (4)**, 711–715.

Bennett, E. M., R. Alpert, and A. C. Goldstein, 1954: Communications through limited response questioning. *Public Opinion Quarterly*, **19 (3)**, 303–308.

Bishop, C. M., 2007: *Pattern Recognition and Machine Learning*. Springer.

Boehm, B. W., 1979: Guidelines for verifying and validating software requirements and design specifications. *Euro IFIP 79*, P. A. Samet, Ed., North Holland, 711–719.

Borg, I. and P. Groenen, 2005: *Modern Multidimensional Scaling: Theory and Applications*. Springer.

Bormann, N. and J.-N. Thépaut, 2004: Impact of MODIS polar winds in ECMWF's 4DVAR data assimilation system. *Monthly Weather Review*, **132 (4)**, 929–940, doi:10.1175/1520-0493(2004)132<0929:IOMPWI>2.0.CO;2.

Brennan, R. L. and D. J. Prediger, 1981: Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement*, **41 (3)**, 687–699, doi:10.1177/001316448104100307.

Broemeling, L. D., 2009: *Bayesian Methods for Measures of Agreement*. Chapman and Hall/CRC.

Cheeseman, P. and J. Stutz, 1996: Bayesian classification (AutoClass): Theory and results. *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad,

G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds., AAAI Press/MIT Press, chap. 6, 62–83.

Chernick, M. R., 2007: *Bootstrap Methods: A Guide for Practitioners and Researchers*. 2d ed., Wiley-Interscience.

Cohen, J., 1960: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20 (1)**, 37.

Congalton, R. G. and K. Green, 1998: *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices*. CRC, 160 pp.

Congalton, R. G., R. G. Oderwald, and R. A. Mead, 1983: Assessing Landsat classification accuracy using discrete multivariate analysis statistical techniques. *Photogrammetric Engineering and Remote Sensing*, **49 (12)**, 1671–1678.

D'Agostini, G., 2003: *Bayesian Reasoning in Data Analysis: A Critical Introduction*. World Scientific Publishing Company.

De Elía, R. and R. Laprise, 2005: Diversity in interpretations of probability: Implications for weather forecasting. *Monthly Weather Review*, **133 (5)**, 1129–1143, doi:10.1175/MWR2913.1.

Derrien, M. and H. LeGléau, 2005: MSG/SEVIRI cloud mask and type from SAFNWC. *International Journal of Remote Sensing*, **26 (21)**, 4707–4732.

Doolittle, M. H., 1888: Association ratios. *Bulletin of the Philosophical Society of Washington*, **10**, 83–96.

Duda, R. O., P. E. Hart, and D. G. Stork, 2001: *Pattern Classification*. 2d ed., Wiley, New York.

Dybbroe, A., K. Karlsson, and A. Thoss, 2005a: NWCSAF AVHRR cloud detection and analysis using dynamic thresholds and radiative transfer modeling. Part I: Algorithm description. *Journal of Applied Meteorology*, **44 (1)**, 39–54.

Dybbroe, A., K. Karlsson, and A. Thoss, 2005b: NWCSAF AVHRR cloud detection and analysis using dynamic thresholds and radiative transfer modeling. Part II: Tuning and validation. *Journal of Applied Meteorology*, **44 (1)**, 55–71.

Efron, B. and R. Tibshirani, 1994: *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Ferro, C. A. T. and D. B. Stephenson, 2011: Extremal Dependence Indices: improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting*, doi:10.1175/WAF-D-10-05030.1.

Finley, J. P., 1884: Tornado predictions. *American Meteorological Journal*, **1**, 85–88.

Fleiss, J. L., J. Cohen, and B. S. Everitt, 1969: Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, **72 (5)**, 323–327, doi: 10.1037/h0028106.

Foody, G., 1992: On the compensation for chance agreement in image classification accuracy assessment. *Photogrammetric Engineering and Remote Sensing*, **58**, 1459–1460.

Foody, G. M., 2009: Classification accuracy comparison: Hypothesis tests and the use of confidence intervals in evaluations of difference, equivalence and non-inferiority. *Remote Sensing of Environment*, **113 (8)**, 1658 – 1663, doi: 10.1016/j.rse.2009.03.014.

Fritz, S. and H. Wexler, 1960: Cloud pictures from satellite TIROS I. *Monthly Weather Review*, **88 (3)**, 79–87.

Gandin, L. S. and A. H. Murphy, 1992: Equitable skill scores for categorical forecasts. *Monthly Weather Review*, **120 (2)**, 361–370, doi:10.1175/1520-0493(1992)120<0361:ESSFCF>2.0.CO;2.

Gelman, A. and J. Hill, 2007: *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Gilleland, E., D. A. Ahijevych, B. G. Brown, and E. E. Ebert, 2010: Verifying forecasts spatially. *Bulletin of the American Meteorological Society*, **91 (10)**, 1365–1373, doi:10.1175/2010BAMS2819.1.

GRS-S Newsletter, 2010: Spotlight on community remote sensing. *IEEE Geoscience and Remote Sensing Society Newsletter*, **(155)**, 10–11, http://www.grss-ieee.org/spotlight-on-community-remote-sensing/.

Hacking, I., 2001: *An Introduction to Probability and Inductive Logic*. Cambridge University Press.

Hamill, T., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Weather and Forecasting*, **14**, 155–167.

Hastie, T., R. Tibshirani, and J. H. Friedman, 2009: *The Elements of Statistical Learning*. 2d ed., Springer.

Heidke, P., 1926: Berechnung des Erfolges und der Gute der Windstarkevorhersagen im Sturmwarnungsdienst. *Geografiska Annaler*, **8**, 301–349.

Hogan, R. J., C. A. T. Ferro, I. T. Jolliffe, and D. B. Stephenson, 2010: Equitability revisited: Why the "equitable threat score" is not equitable. *Weather and Forecasting*, **25 (2)**, 710–726, doi:10.1175/2009WAF2222350.1.

Hogan, R. J., E. J. O'Connor, and A. J. Illingworth, 2009: Verification of cloud-fraction forecasts. *Quarterly Journal of the Royal Meteorological Society*, **135 (643)**, 1494–1511, doi:10.1002/qj.481.

Huang, H.-L. and P. Antonelli, 2001: Application of principal component analysis to high-resolution infrared measurement compression and retrieval. *Journal of Applied Meteorology*, **40 (3)**, 365–388, doi:10.1175/1520-0450(2001)040<0365:AOPCAT>2.0.CO;2.

Hyvärinen, A., J. Karhunen, and E. Oja, 2001: *Independent Component Analysis*. Wiley-Interscience.

Jeffrey, R., 2004: *Subjective Probability: The Real Thing*. Cambridge University Press.

Jolliffe, I., 2002: *Principal Component Analysis*. Springer.

Jolliffe, I. T., 2007: Uncertainty and inference for verification measures. *Weather and Forecasting*, **22 (3)**, 637–650, doi:10.1175/WAF989.1.

Jolliffe, I. T. and D. B. Stephenson, (Eds.) , 2003: *Forecast Verification: A Practitioners Guide in Atmospheric Science*. Wiley, 240 pp.

Kållberg, P., S. Uppala, and A. Simmons, 2010: The real first weather satellite picture. *Weather*, **65 (8)**, 1477–8696, doi:10.1002/wea.652.

Kaufman, L. and P. Rousseeuw, 1990: *Finding Groups in Data — An Introduction to Cluster Analysis*. Wiley Interscience, New York.

Kelly, G. and J.-N. Thépaut, 2007: Evaluation of the impact of the space component of the Global Observing System through Observing System Experiments. *Seminar on Recent developments in the use of satellite observations in Numerical Weather Prediction, 3 – 7 September 2007*, ECMWF, 327–348.

Kelly, T., 2008: Evidence. *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., Stanford University, Fall 2008 edition, available at http://plato.stanford.edu/archives/fall2008/entries/evidence/.

Kidder, S. Q. and T. H. Vonder Haar, 1995: *Satellite meteorology: an introduction*. Academic Press.

Kiers, H. and P. Groenen, 2006: Visualizing dependence of bootstrap confidence intervals for methods yielding spatial configurations. *Data Analysis, Classification and the Forward Search*, S. Zani, A. Cerioli, M. Riani, and M. Vichi, Eds., Springer Berlin Heidelberg, 119–126, doi:10.1007/3-540-35978-8_14.

Krippendorff, K., 1970: Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement*, **30 (1)**, 61–70, doi:10.1177/001316447003000105.

Krippendorff, K. H., 2003: *Content Analysis: An Introduction to Its Methodology*. 2d ed., Sage Publications, Inc.

Lahiri, S. N., 2003: *Resampling Methods for Dependent Data*. Springer.

Lazo, J. K., 2010: The costs and losses of integrating social sciences and meteorology. *Weather, Climate, and Society*, **2 (3)**, 171–173, doi: 10.1175/2010WCAS1086.1.

Liu, C., P. Frazier, and L. Kumar, 2007: Comparative assessment of the measures of thematic classification accuracy. *Remote Sensing of Environment*, **107 (4)**, 606 – 616, doi:10.1016/j.rse.2006.10.010.

Livezey, R. E., 2003: Categorical events. *Forecast Verification: A Practitioners Guide in Atmospheric Science*, I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, chap. 4.

Meulman, J. and W. Heiser, 1983: The display of bootstrap solutions in multidimensional scaling, unpublished manuscript.University of Leiden, Netherlands.

Murphy, A. H., 1996: The Finley affair: A signal event in forecast verification. *Weather and Forecasting*, **11 (1)**, 3–20.

Murphy, A. H., 1998: The early history of probability forecasts: Some extensions and clarifications. *Weather and Forecasting*, **13 (1)**, 5–15, doi:10.1175/1520-0434(1998)013<0005:TEHOPF>2.0.CO;2.

Noordung, H., 1929: *Das Problem der Befahrung des Weltraumes – Der Raketen-Motor*. Richard Carl Schmidt, Berlin.

Peirce, C. S., 1884: The numerical measure of the success of predictions. *Science*, **4**, 453–454, doi:10.1126/science.ns-4.93.453-a.

Pepe, M. S., 2003: *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **126 (565)**, 649–667, doi:10.1002/qj.49712656313.

Ripley, B. D., 1996: *Pattern Recognition and Neural Networks*. Cambridge University Press.

Russell, S. and P. Norvig, 2010: *Artificial Intelligence: A Modern Approach*. 3d ed., Prentice-Hall.

Rutjes, A., J. Reitsma, A. Coomarasamy, K. Khan, and P. Bossuyt, 2007: Evaluation of diagnostic tests when there is no gold standard. A review of methods. *Health Technology Assessment*, **11 (50)**, 1–72, doi:10.3310/hta11500.

Sammon, J. W., 1969: A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, **C-18**, 401–409, doi:10.1109/T-C.1969.222678.

Savage, L. J., 1951: The theory of statistical decision. *Journal of the American Statistical Association*, **46 (253)**, 55–67.

Scott, W., 1955: Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, **19 (3)**, 321–325.

Stehman, S. V., 1997: Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, **62 (1)**, 77 – 89, doi:10.1016/S0034-4257(97)00083-7.

Stephenson, D., 2000: Use of the "odds ratio" for diagnosing forecast skill. *Weather and Forecasting*, **15 (2)**, 221–232.

Stephenson, D. B. and I. T. Jolliffe, 2003: Forecast evaluation in other disciplines. *Forecast Verification: A Practitioners Guide in Atmospheric Science*, I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, chap. 9.3.

Therneau, T. M. and E. J. Atkinson, 1997: An introduction to recursive partitioning using the RPART routine. Tech. Rep. 61, Section of Biostatistics, Mayo Clinic, Rochester. http://www.mayo.edu/hsr/techrpt/61.pdf.

Thompson, J. C., 1950: A numerical method for forecasting rainfall in the Los Angeles area. *Monthly Weather Review*, **78 (7)**, 113–124, doi:10.1175/1520-0493(1950)078<0113:ANMFFR>2.0.CO;2.

Thompson, J. C. and G. W. Brier, 1955: The economic utility of weather forecasts. *Monthly Weather Review*, **83 (11)**, 249–253, doi:10.1175/1520-0493(1955)083<0249:TEUOWF>2.0.CO;2.

Tovinkere, V. R., M. Penaloza, A. Logar, J. Lee, R. C. Weger, T. A. Berendes, and R. M. Welch, 1993: An intercomparison of artificial intelligence approaches for polar scene identification. *Journal of Geophysical Research*, **98 (D3)**, 5001–5016.

Trishchenko, A. P. and L. Garand, 2011: Spatial and temporal sampling of polar regions from two-satellite system on Molniya orbit. *Journal of Atmospheric and Oceanic Technology*, doi:10.1175/JTECH-D-10-05013.

Tuovinen, J.-P., A.-J. Punkka, J. Rauhala, H. Hohti, and D. M. Schultz, 2009: Climatology of severe hail in finland: 1930–2006. *Monthly Weather Review*, **137 (7)**, 2238–2249, doi:10.1175/2008MWR2707.1.

Upper, D., 1974: The unsuccessful self-treatment of a case of "writer's block". *Journal of Applied Behavior Analysis*, **7 (3)**, 497, doi: 10.1901/jaba.1974.7-497a.

Venables, W. N. and B. D. Ripley, 2002: *Modern Applied Statistics with S. Fourth Edition*. Springer, New York.

Wasserman, L., 2003: *All of Statistics: A Concise Course in Statistical Inference*. Springer.

Welch, R. M., S. K. Sengupta, A. K. Goroch, P. Rabindra, N. Rangaraj, and M. S. Navar, 1992: Polar cloud and surface classification using AVHRR imagery: An intercomparison of methods. *Journal of Applied Meteorology*, **31**, 405–420.

Wilks, D. S., 2006: *Statistical methods in the atmospheric sciences*. 2d ed., Academic Press.

ZAMG, 2009: *Manual of synoptic satellite meteorology – conseptual models and case studies, Version 6.8*. Zentralanstalt für Meteorologie und Geodynamik (Central Institute for Meteorology and Geodynamics), Koninklijk Nederlands Meteorologisch Instituut (Royal Netherlands Meteorological Institute), FMI, Drzavni Hidrometeoroloski Zavod (Croatian Meteorological and Hydrological Service), and EUMETSAT, http://www.zamg.ac.at/docu/Manual/.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bulletin of the American Meteorological Society*, **83 (1)**, 73–83, doi:10.1175/1520-0477(2002)083<0073:TEVOEB>2.3.CO;2.

Zingerle, C. and P. Nurmi, 2008: Monitoring and verifying cloud forecasts originating from operational numerical models. *Meteorological Applications*, **15 (3)**, 325–330, doi:10.1002/met.73.

Zwick, R., 1988: Another look at interrater agreement. *Psychological Bulletin*, **103**, 374–378.

Finnish Meteorological Institute Contributions

1. Joffre, Sylvain M., 1988. Parameterization and assessment of processes affecting the long-range transport of airborne pollutants over the sea. 49 p.

2. Solantie, Reijo, 1990. The climate of Finland in relation to its hydrology, ecology and culture. 130 p.

3. Joffre, Sylvain M. and Lindfors, Virpi, 1990. Observations of airborne pollutants over the Baltic Sea and assessment of their transport, chemistry and deposition. 41 p.

4. Lindfors, Virpi, Joffre, Sylvain M. and Damski, Juhani, 1991. Determination of the wet and dry deposition of sulphur and nitrogen compounds over the Baltic Sea using actual meteorological data. 111 p.

5. Pulkkinen, Tuija, 1992. Magnetic field modelling during dynamic magnetospheric processes. 150 p.

6. Lönnberg, Peter, 1992. Optimization of statistical interpolation. 157 p.

7. Viljanen, Ari, 1992. Geomagnetic induction in a one- or two-dimensional earth due to horizontal ionospheric currents. 136 p.

8. Taalas, Petteri, 1992. On the behaviour of tropospheric and stratospheric ozone in Northern Europe and in Antarctica 1987-90. 88 p.

9. Hongisto, Marke, 1992. A simulation model for the transport, transformation and deposition of oxidized nitrogen compounds in Finland — 1985 and 1988 simulation results. 114 p.

10. Taalas, Petteri, 1993. Factors affecting the behaviour of tropospheric and stratospheric ozone in the European Arctic and Antarctica. 138 s.

11. Mälkki, Anssi, 1993. Studies on linear and non-linear ion waves in the auroral acceleration region. 109 p.

12. Heino, Raino, 1994. Climate in Finland during the period of meteorological observations. 209 p.

13. Janhunen, Pekka, 1994. Numerical simulations of E-region irregularities and ionosphere-magnetosphere coupling. 122 p.

14. Hillamo, Risto E., 1994. Development of inertial impactor size spectroscopy for atmospheric aerosols. 148 p.

15. Pakkanen, Tuomo A., 1995. Size distribution measurements and chemical analysis of aerosol components. 157 p.

16. Kerminen, Veli-Matti, 1995. On the sulfuric acid-water particles via homogeneous nucleation in the lower troposphere. 101 p.

17. Kallio, Esa, 1996. Mars-solar wind interaction: Ion observations and their interpretation. 111 p.

18. Summanen, Tuula, 1996. Interplanetary Lyman alpha measurements as a tool to study solar wind properties. 114 p.

19. Rummukainen, Markku, 1996. Modeling stratospheric chemistry in a global three-dimensional chemical transport model, SCTM-1. Model development. 206 p.

20. Kauristie, Kirsti, 1997. Arc and oval scale studies of auroral precipitation and electrojets during magnetospheric substorms. 134 p.

21. Hongisto, Marke, 1998. Hilatar, A regional scale grid model for the transport of sulphur and nitrogen compounds. 152 p.

22. Lange, Antti A.I., 1999. Statistical calibration of observing systems. 134 p.

23. Pulkkinen, Pentti, 1998. Solar differential rotation and its generators: computational and statistical studies. 108 p.

24. Toivanen, Petri, 1998. Large-scale electromagnetic fields and particle drifts in time-dependent Earth's magnetosphere. 145 p.

25. Venäläinen, Ari, 1998. Aspects of the surface energy balance in the boreal zone. 111 p.

26. Virkkula, Aki, 1999. Field and laboratory studies on the physical and chemical properties of natural and anthropogenic tropospheric aerosol. 178 p.

27. Siili, Tero, 1999. Two-dimensional modelling of thermal terrain-induced mesoscale circulations in Mars' atmosphere. 160 p.

28. Paatero, Jussi, 2000. Deposition of Chernobyl-derived transuranium nuclides and short-lived radon-222 progeny in Finland. 128 p.

29. Jalkanen, Liisa, 2000. Atmospheric inorganic trace contaminants in Finland, especially in the Gulf of Finland area. 106 p.

30. Mäkinen, J. Teemu, T. 2001. SWAN Lyman alpha imager cometary hydrogen coma observations. 134 p.

31. Rinne, Janne, 2001. Application and development of surface layer flux techniques for measurements of volatile organic compound emissions from vegetation. 136 p.

32. Syrjäsuo, Mikko T., 2001. Auroral monitoring system: from all-sky camera system to automated image analysis. 155 p.

33. Karppinen, Ari, 2001. Meteorological pre-processing and atmospheric dispersion modelling of urban air quality and applications in the Helsinki metropolitan area. 94 p.

34. Hakola, Hannele, 2001. Biogenic volatile organic compound (VOC) emissions from boreal deciduous trees and their atmospheric chemistry. 125 p.

35. Merenti-Välimäki, Hanna-Leena, 2002. Study of automated present weather codes.153 p.

36. Tanskanen, Eija I., 2002. Terrestrial substorms as a part of global energy flow. 138 p.

37. Nousiainen, Timo, 2002. Light scattering by nonspherical atmospheric particles. 180 p.

38. Härkönen, Jari, 2002. Regulatory dispersion modelling of traffic-originated pollution. 103 p.

39. Oikarinen, Liisa, 2002. Modeling and data inversion of atmospheric limb scattering measurements. 111 p.

40. Hongisto, Marke, 2003. Modelling of the transport of nitrogen and sulphur contaminants to the Baltic Sea Region. 188 p.

41. Palmroth, Minna, 2003. Solar wind – magnetosphere interaction as determined by observations and a global MHD simulation. 147 p.

42. Pulkkinen, Antti, 2003. Geomagnetic induction during highly disturbed space weather conditions: Studies of ground effects 164 p.

43. Tuomenvirta, Heikki, 2004. Reliable estimation of climatic variations in Finland. 158 p.

44. Ruoho-Airola, Tuija, 2004. Temporal and regional patterns of atmospheric components affecting acidification in Finland. 115 p.

45. Partamies, Noora, 2004. Meso-scale auroral physics from groundbased observations. 122 p.

46. Teinilä, Kimmo, 2004. Size resolved chemistry of particulate ionic compounds at high latitudes. 138 p.

47. Tamminen, Johanna, 2004. Adaptive Markov chain Monte Carlo algorithms with geophysical applications. 156 p.

48. Huttunen, Emilia, 2005. Interplanetary shocks, magnetic clouds, and magnetospheric storms. 142 p.

49. Sofieva, Viktoria, 2005. Inverse problems in stellar occultation. 110 p.

50. Harri, Ari-Matti, 2005. In situ observations of the atmospheres of terrestrial planetary bodies. 246 p.

51. Aurela, Mika, 2005. Carbon dioxide exchange in subarctic ecosystems measured by a micrometeorological technique. 132 p.

52. Damski, Juhani, 2005. A Chemistry-transport model simulation of the stratospheric ozone for 1980 to 2019. 147 p.

53. Tisler, Priit, 2006. Aspects of weather simulation by numerical process. 110 p.

54. Arola, Antti, 2006. On the factors affecting short- and long-term UV variability. 82 p.

55. Verronen, Pekka T., 2006. Ionosphere-atmosphere interaction during solar proton events. 146 p.

56. Hellén, Heidi, 2006. Sources and concentrations of volatile organic compounds in urban air. 134 p.

57. Pohjola, Mia, 2006. Evaluation and modelling of the spatial and temporal variability of particulate matter in urban areas. 143 p.

58. Sillanpää, Markus, 2006. Chemical and source characterisation of size-segregated urban air particulate matter. 184 p.

59. Niemelä, Sami, 2006. On the behaviour of some physical parameterization methods in high resolution numerical weather prediction models. 136 p.

60. Karpechko, Alexey, 2007. Dynamical processes in the stratosphere and upper troposphere and their influence on the distribution of trace gases in the polar atmosphere. 116 p.

61. Eresmaa, Reima, 2007. Exploiting ground-based measurements of Global Positioning System for numerical weather prediction. 95 p.

62. Seppälä, Annika, 2007. Observations of production and transport of NOx formed by energetic particle precipitation in the polar night atmosphere. 100 p.

63. Rontu, Laura, 2007. Studies on orographic effects in a numerical weather prediction model. 151 p.

64. Vajda, Andrea, 2007. Spatial variations of climate and the impact of disturbances on local climate and forest recovery in northern Finland. 139 p.

65. Laitinen, Tiera, 2007. Rekonnektio Maan magnetosfäärissä – Reconnection in Earth's magnetosphere. 226 s.

66. Vanhamäki, Heikki, 2007. Theoretical modeling of ionospheric electrodynamics including induction effects. 170 p.

67. Lindfors. Anders, 2007. Reconstruction of past UV radiation. 123 p.

68. Sillanpää, Ilkka, 2008. Hybrid modelling of Titan's interaction with the magnetosphere of Saturn. 200 p.

69. Laine, Marko, 2008. Adaptive MCMC methods with applications in environmental and geophysical models. 146 p.

70. Tanskanen, Aapo, 2008. Modeling of surface UV radiation using satellite data. 109 p.

71. Leskinen, Ari, 2008. Experimental studies on aerosol physical properties and transformation in environmental chambers. 116 p.

72. Tarvainen, Virpi, 2008. Development of biogenetic VOC emission inventories for the boreal forest. 137 p.

73. Lohila, Annalea, 2008. Carbon dioxide exchange on cultivated and afforested boreal peatlands. 110 p.

74. Saarikoski, Sanna, 2008. Chemical mass closure and source-specific composition of atmospheric particles. 182 p.

75. Pirazzini, Roberta, 2008. Factors controlling the surface energy budget over snow and ice. 141 p.

76. Salonen, Kirsti, 2008. Towards the use of radar winds in numerical weather prediction. 87 p.

77. Luojus, Kari, 2009. Remote sensing of snow-cover for the boreal forest zone using microwave radar. 178 p.

78. Juusola, Liisa, 2009. Observations of the solar wind-magnetosphere-ionosphere coupling. 167 p.

79. Waldén, Jari, 2009. Meteorology of gaseous air pollutants. 177 p.

80. Mäkelä, Jakke, 2009. Electromagnetic signatures of lightning near the HF frequency band. 152 p.

81. Thum, Tea, 2009. Modelling boreal forest CO_2 exchange and seasonality. 140 p.

82. Lallo, Marko, 2010. Hydrogen soil deposition and atmospheric variations in the boreal zone. 91 p.

83. Sandroos, Arto, 2010. Shock acceleration in the solar corona. 116 p.

84. Lappalainen, Hanna, 2010. Role of temperature in the biological activity of a boreal forest. 107 p.

85. Mielonen, Tero, 2010. Evaluation and application of passive and active optical remote sensing methods for the measurement of atmospheric aerosol properties. 125 p.

86. Lakkala, Kaisa, 2010. High quality polar UV measurements : scientific analyses and transfer of the irradiance scale. 156 p.

87. Järvinen, Riku, 2011. On ion escape from Venus. 150 p.

88. Saltikoff, Elena, 2011. On the use of weather radar for mesoscale applications in northern conditions. 120 p.

89. Timonen, Hilkka, 2011. (Valmisteilla – In preparation)

90. Hyvärinen, Otto, 2011. Categorical meteorological products: evaluation and analysis. 138 p.