Verification of Global Radiation Forecasts from the Ensemble Prediction System at DMI

SISSE CAMILLA LUNDHOLM

Centre for Meteorological Modelling
The Danish Meteorological Institute
DMI

The PhD School of Science
Faculty of Science
University of Copenhagen

Denmark

February 18, 2013
Preface

This thesis was prepared at the Danish Meteorological Institute as part of acquiring a Ph.D. degree in science.

The Ph.D. project is part of the project "Solar/Electric heating systems in the future energy system" guided by the Technical University of Denmark and funded by The Danish Council for Strategic Research.

The thesis consists of two parts separated in time by a year of leave of absence.

København, February 2013

Sisse Camilla Lundholm

Supervisors: Aksel Walløe Hansen and Kristian Pagh Nielsen

Cover image from:
Acknowledgements

I would like to thank associate professor Simon Furbo at the Technical University of Denmark and head of the project “Solar/electric heating systems in the future energy system”, of which my PhD project is a part, and I would also like to thank “The Danish Council for Strategic Research” under the “Danish Agency for Science Technology and Innovation” for funding the project. I would like to thank all partners within the project especially the Department of Civil Engineering and Informatics at The Technical University in Denmark, Innogie Aps, ENFOR A/S, and COWI A/S for their dedication at our follow-up meetings. I thank DMI for help and support during my work.

I am thankful for being given the opportunity to work on this highly prospective subject.
Abstract

To comply with an increasing demand for sustainable energy sources, a solar heating unit is being developed at the Technical University of Denmark. To make optimal use — environmentally and economically —, this heating unit is equipped with an intelligent control system using forecasts of the heat consumption of the house and the amount of available solar energy. In order to make the most of this solar heating unit, accurate forecasts of the available solar radiation are essential. However, because of its sensitivity to local meteorological conditions, the solar radiation received at the surface of the Earth can be highly fluctuating and challenging to forecast accurately.

To comply with the accuracy requirements to forecasts of both global, direct, and diffuse radiation, the uncertainty of these forecasts is of interest. Forecast uncertainties can become accessible by running an ensemble of forecasts, and to this end, these three meteorological quantities have since August 2011 been output parameters from the high-resolution ensemble prediction system at the Danish Meteorological Institute.

The appropriateness of complementing forecast values with uncertainty estimates derived from the ensemble forecasts has been assessed by investigating the degree to which the ensemble members and the truth — here materialised by the verifying observation — are statistically indistinguishable. A degree of under-dispersion of the ensemble members is evident concerning global radiation, and the ensemble forecasts will therefore tend to express too little uncertainty in the forecast values. Under-dispersiveness is a well-known problem in ensemble prediction. Uncertainties on observations may cause some of the under-dispersiveness of the ensemble forecasts of global radiation.
Referat (danish)

For at imødekomme en stigende efterspørgsel efter vedvarende energikilder er en særlig solvarmeenhed — udstyret med et intelligent styresystem — under udvikling på Danmarks Tekniske Universitet.

For at få optimalt udnytte af solvarmeenheden — klimaemæssigt og økonomisk — anvender det intelligente styresystem prognoser af varmeforbruget i huset og af den solenergi, der er til rådighed. Præcise prognoser af solstrålingen er afgørende for bedst mulig brug af varmeenheden, men på grund af solstrålingenens sensitivitet over for lokale meteorologisk forhold kan den være stærkt varierende og vanskelig at forudsige præcist.

I forbindelse med udviklingen af solvarmeenheden er det Danmarks Meteorologiske Instituts (DMI’s) — og dermed mit Ph.d.-projekts — rolle at levere præcise prognoser af solstrålingen. En evaluering af prognoser af globalstrålingen og den direkte og den diffuse stråling har afsløret problemer i strålingskemaet i DMI’s HIRLAM modeller, og strålingskemaet er efterfølgende blevet revideret.

En måde at efterkomme kravene til præcision af prognoser af globalstrålingen, den direkte og den diffuse strålingen, er ved at beskæftige sig med usikkerhederne for prognoserne. Disse usikkerheder kan blive tilgængelige ved at kører et ensemble af simultane prognoser. DMI’s ensemble-prognose-system baseret på en højopløsningsmodel beregnet til at opfange lokale detaljer i vejret har siden august 2011 leveret prognoser af globalstrålingen, den direkte og den diffuse stråling.

Egnetheden af de estimater — udeladt af ensemblemedlemmerne — af prognoseusikkerhederne er blevet vurderet ud fra en undersøgelse af i hvor høj grad ensemblemedlemmerne og den sande værdi — her observationen — er statistisk uafhængige. Spredningen blandt ensemblemedlemmerne er fundet for lille til at indeholde observationerne med en forventet andel, og ensembleprognoserne vil derfor være tilbøjelige til at give en for beskeden usikkerhed på de forudsagte værdier. For lille spredning blandt ensemblemedlemmerne er et velkendt problem i ensemble-prognose-systemer. En del af den begrænsede ensemblespredning kan muligvis skyldes usikkerheder på observationerne.

Hvad angår RMSE, skill score og diskrimination præsterer ensemble-
midlet bedre ens kontrolprognosen, og det kan være en ide at supplere værdien af en operationel prognose af globalstrålingen, men måske også relevant for andre meteorologiske parametre, med værdien af ensemble-midlet.
# Contents

1. **Introduction**  
   1.1 Ensemble Prediction .......................... 2  
   1.2 Organisation of the Thesis ..................... 3  

2. **Prelude: The Smart Solar Heating Unit**  

3. **Global Radiation**  
   3.1 Definitions .................................... 9  
   3.2 Shortwave Radiation in DMI's HIRLAM models .... 11  

4. **Data**  
   4.1 Observations .................................... 15  
   4.2 Matching Observations with Model Calculations ... 15  
   4.3 Model calculations ............................... 17  
   4.4 Multidimensional Data ............................ 18  

5. **Performance of Global Radiation Forecasts**  
   5.1 Data ............................................. 20  
   5.1.1 Observations .................................. 20  
   5.1.2 Model Calculations of Global Radiation ...... 20  
   5.2 Comparison of Model Calculations with Observations .... 21  
   5.3 Comparing Model Calculations with DTU Observations ... 22  
   5.4 A Revision of the Radiation Scheme in DMI's HIRLAM models 24  

6. **DMI's Ensemble Prediction System**  
   6.1 Ensemble Prediction .............................. 28  
   6.1.1 Initial Condition Perturbations ............... 30  
   6.1.2 Model Perturbations ........................... 31  
   6.2 Construction of DMI-EPS .......................... 32  
   6.2.1 Initial Condition Perturbations ............... 32  
   6.2.2 Model perturbations ............................ 33
7  A Framework for Forecast Verification 35
  7.1 Aspects of Forecast Quality 35
    7.1.1 Traditional Measures of Forecast Quality 35
    7.1.2 General approach to Forecast Verification 37
  7.2 Quality of Ensemble Forecasts 42
    7.2.1 Consistency 42
    7.2.2 Multidimensional Consistency 43
    7.2.3 Discrimination 47
    7.2.4 Summary of Aspects of Forecast Quality 48

8 Verification of Global Radiation from DMI-EPS 50
  8.1 The Data 50
    8.1.1 Ensemble Calculations of Global Radiation 50
    8.1.2 The Observations 51
    8.1.3 The Clearness Index 51
  8.2 Assessing the Performance of the Ensemble Mean 52
    8.2.1 Sharpness 53
    8.2.2 Bias, MSE, RMSE, and Skill Score 53
    8.2.3 Reliability 61
    8.2.4 Discrimination 64
  8.3 Assessing the Performance of DMI-EPS 68
    8.3.1 Application of quality measures of single forecasts 68
    8.3.2 Application of quality measures of ensemble forecasts 73

9 Discussion 80
  9.1 Verification of Global Radiation from DMI-EPS 80
    9.1.1 Assessing the Performance of DMI-EPS 80
    9.1.2 Assessing the Performance of the Ensemble Mean 83
  9.2 Performance of Global Radiation Forecasts 84
  9.3 High-Resolution, Limited-Area Models 85
  9.4 The Solar Heating Unit 85

10 Conclusion & Outlook 86

Appendices I

A Derivation of Beer’s Law II

B Discrimination VI

C Generalized Discrimination Score VIII

Summary X

Bibliography XII
1 Introduction

Under the direction of the Department of Civil Engineering at the Technical University of Denmark, a project with the purpose of elucidating how to best design an individual heating unit based on solar energy was initiated in 2008. This solar heating unit is a so-called smart solar heating unit in the sense that it adjusts its heated water volume according to forecasts of the heat consumption of the household and of the available solar energy.

The solar radiation available at the ground (and at the solar collectors) is highly variable due to the dependence on meteorological conditions (Lorenz et al., 2011), and in order to make optimal use — environmentally and economically — of this smart solar heating unit, accurate predictions of the available solar energy are essential. Within this project, the task of the Danish Meteorological Institute (DMI) — and thus my Ph.D. work — is to deliver accurate forecasts of the solar radiation reaching the surface of the Earth.

It is, however, highly challenging to forecast solar radiation at the surface of the Earth accurately even though the position of the Sun in the sky can be determined accurately from astronomical formulas and the transfer of solar radiation through an absorbing and scattering medium like the atmosphere is well established (Paltridge and Platt, 1976; Thomas and Stamnes, 1999). On the way through the atmosphere, the solar radiation interacts with atmospheric constituents such as molecules, ozone, water vapour, aerosols, and cloud particles (water droplets and/or ice particles) (Paltridge and Platt, 1976; Savijärvi, 1990; Stephens, 1984; Wyser et al., 1999), which all attenuate the solar radiation by absorption or scattering. The distribution of cloud particles and aerosols can be difficult to forecast to a high degree of accuracy, and especially cloud particles may reflect a large fraction of the incoming solar radiation (Juan and Da-Rem, 2012; Kasten and Czeplak, 1980).

Different approaches have been used to predict the solar surface radiation depending on the time scale of the forecast. Hammer et al. (1999) used cloud motion vectors derived from consecutive satellite pictures to make very short-range (30 minutes to 2 hours) forecasts of solar radiation reaching the surface. At forecast ranges beyond about 6 hours, numerical weather prediction (NWP) models have been found superior to other alternatives (Perez...
et al., 2011). Apparently, the full coherent system of equations describing the evolution of the atmosphere is needed for these longer forecast ranges.

1.1 Ensemble Prediction

An inherent instability of an atmospheric state to small-scale perturbations (Lorenz, 1963) imposes a challenge to NWP forecasting of any meteorological parameter. This atmospheric instability — caused by non-linearity of the atmosphere — introduces a high degree of sensitivity to the initial conditions of an NWP forecast. The observations of the atmosphere, on which the initial conditions of an NWP forecast is based, are typically neither sufficiently exact nor sufficiently dense to consider the initial conditions as known with certainty (Epstein, 1969). Due to the imperfections in the observations, a multitude of nearly identical initial states will all be consistent with the observations, and each of these initial states may in time evolve into considerably different atmospheric states (Epstein, 1969; Lorenz, 1982). Because of the non-linearity of the atmosphere, the fate of NWP forecasting is that small errors in determining the initial state of the atmosphere, may with time lead to a forecasts that diverges from the true evolution of the atmosphere. Epstein (1969) concluded that the atmosphere could not be completely described with a single forecast run due to this uncertainty in the initial state, and in the following decades, the idea of running a collection, or an ensemble, of forecasts initiated from slightly perturbed states emerged (see e.g. Molteni et al. (1996); Toth and Kalnay (1993); Tracton and Kalnay (1993)).

Since the 1990s, ensemble prediction forecasts have become increasingly important as a mean of addressing forecast uncertainty, and ensemble predictions are made at most of the major operational weather prediction centres worldwide including the National Centers for Environmental Prediction (NCEP), the European Centre for Medium-Range Weather Forecasts (ECMWF), the United Kingdom Met Office and Meteo-France. Until recently, the main focus has been on global medium range (typically 10 days) forecasts, but with increased computer resources it has become relevant to pay attention to high resolution ensemble forecasts for a limited area (referred to as LAM models) (Bowler et al., 2008; Díaz et al., 2012; Feddersen, 2009). These model systems have the potential of addressing the uncertainty in both high impact weather and more detailed features of cloud fields and radiation up to about two days ahead (Díaz et al., 2012).

To access the uncertainty of solar radiation forecasts, DMI’s high-resolution, limited area (LAM) ensemble prediction system (DMI-EPS) has since August 2011 made forecasts of the incoming solar radiation. DMI-EPS is aimed at capturing small-scale weather features of which a solar heating unit can be expected to be sensitive. Making forecasts of the future and providing
these with estimates of the uncertainty associated with them is fundamental in ensemble forecasting and is in line with the on-going transition within the field of meteorology from point prediction toward distributional prediction (Gneiting et al., 2008).

1.2 Organisation of the Thesis

In addressing the performance of forecasts of solar radiation, this thesis consists of two parts separated in time:

1. A part that describes the identification of shortcomings in the radiation scheme of DMI’s NWP models that followed from evaluating model calculations of solar radiation against observations (a work performed during 2009). A short description of the subsequent revision of the radiation scheme finishes this part.

2. A part that describes a detailed verification of ensemble forecasts of solar radiation with the aim of investigating the prospects of complementing the forecasts with uncertainty estimates as derived from ensemble forecasts from DMI-EPS (a work performed during 2011–2012).

The organization of my thesis is as follows:

Chapter 2 describes the smart solar heating unit being the foundation of this work, and the role of the involved partners in developing this solar heating unit.

Chapter 3 defines global, direct, and diffuse radiation and describes the parametrisations of these quantities in the radiation scheme of DMI’s NWP models.

Chapter 4 gives a general description of observations and model calculations of global radiation.

Chapter 5 describes the identification of problems of the radiation scheme of DMI’s NWP models and the subsequent revision of the scheme.

Chapter 6 describes the configuration of DMI’s ensemble prediction system, DMI-EPS.

Chapter 7 describes aspects of forecasts quality used to evaluate the ensemble forecasts.

Chapter 8 verifies by different means the forecasts of global radiation from DMI’s ensemble prediction system.
Chapter 9 contains an assessment of the potential of DMI’s ensemble forecasts of solar radiation and of the possibility of using the distribution of the ensemble members as estimates of the forecast uncertainties.

Chapter 10 concludes this work and provides some outlooks.

Most data analysis and plotting have been performed with the aid of the language for statistical computing R Development Core Team (2011) and associated packages to R: NCAR - Research Application Program (2010) and Weigel (2010).
2 Prelude: The Smart Solar Heating Unit

The project on developing a smart solar heating unit using forecasts of the heat consumption and the available solar energy is a collaboration between a number of partners. Beside the Department of Civil Engineering at the Technical University of Denmark (DTU) and DMI, other contributors include DTU Informatics (DTU), COWI A/S (Consultancy Within Engineering, Environmental Science and Economics), ENFOR A/S (Forecasting and Optimization for the Energy Sector), Ohmatex Aps, Ajva Aps, and Innogie Aps (Perers et al., 2009). In this brief description of the project and this smart solar heating unit, the role of the different partners will be outlined.

A solar heating unit illustrated in figure 2.1 includes a solar thermal collector plate and a water tank, where heat can be stored. If the solar energy cannot cover the full heat demand, which in the Danish climate probably will be the case on most days, an electricity based auxiliary energy source, such as an electric heating element or a heat pump, should provide the heat deficiency.

In Denmark, it is expected that an increasing part of the energy consumption in the future will be covered by wind farms (Perers et al., 2009). From the newly negotiated energy agreement (http://www.kemin.dk/en-US/Climate_energy_and_building_policy_Denmark/energy_agreements/Sider/Forside.aspx), it is evident that by 2020, just below 50% of the Danish electricity consumption should be covered by wind energy. This will give an increased number of windy periods with a surplus of electricity and hence low electricity prices. These variable electricity prices, which contains a daily variation, are not yet of benefit to the customer. With the introduction in large scale of wind energy into the Danish electricity grid, fluctuating electricity prices at the customer level are expected to be a reality (Perers et al., 2009). The solar heating unit developed at DTU is optimized to make use of this occasional excess electricity produced in windy periods either through an electric heating element or via a heat pump. The smart solar heating unit can, if used in large numbers, facilitate the introduction of wind energy into the Danish energy system (Perers et al., 2009). In the future electricity...
Figure 2.1: Illustration of the principle in a solar heating unit; cold water from the bottom of the tank gets heated by the solar collectors and re-enters in the top of the tank, which acts as a storage of energy. The hot water is used for domestic use and space heating (from the two top outlets). Typically, solar thermal collectors make use of solar radiation with wavelengths within the interval $300 - 4000 \text{ nm}$ [Badescu, 2008, Chap. 1], which is referred to as shortwave radiation. Note that here, no auxiliary heating of the water is shown.


In a traditional solar heating tank seen in figure 2.2(a) [left], a fixed volume in the top of the water tank is kept at a certain temperature either by solar energy or by the auxiliary heating. If the solar energy suffices, it also heats the remaining volume of the water. In the smart solar heating unit developed at DTU illustrated in figure 2.2(a) [right], the water is heated from above and has a flexible volume determined by the actual heat demand. This configuration lowers the heat loss of the tank and hence increases the thermal performance of the water tank compared to ordinary solar heating tanks [Furbo and Andersen, 2009]. To lower the heat loss further, a strong thermal stratification within the tank is desirable. At the Department of Civil Engineering at DTU, simulations of solar heating units to identify the best design of the water tank, the optimal area of the solar collector, and the financial reduction for the customer by introducing variable electricity prices have been performed and are described in [Perers et al., 2009, 2010].

At time of writing, a solar collector area of $9 \text{ m}^2$ and a water tank of $750 \text{ l}$ produced by Ajva Aps (as the one seen in figure 2.2(b)) are chosen for the smart solar heating unit. Tests of stratification inlet pipes within the solar water tank resulted in fabric pipes from Ohmatex Aps, which are able to support a thermal stratification. To find the most favorable solution of the
auxiliary heating concerning the degree of thermal stratification within the tank and thereby the heat balance of the tank and price of purchase and endurance, three different auxiliary systems are being tested at the test site at the Department of Civil Engineering at DTU.

At COWI A/S, a socio-economic analysis evaluating the advantages to society in terms of cost-savings and reductions in CO$_2$ emissions of implementing in large scale the smart solar heating unit into the Danish energy system has been performed for different types of houses with respect to the heat demand and for both types of auxiliary heating (electric heating element and heat pump). Without considering expenses to the solar heating unit, a unit using a heat pump as the auxiliary heating performs better with respect to reductions in socio-economic costs and to reductions in CO$_2$ emissions. The smart solar heating unit has the potential of being an appealing alternative to oil burners and natural gas boilers both from an environmental and an economical point of view (Perers et al. 2009).

Adjusting the hot water volume in the tank to always cover the heat demand requires knowledge of the future heat consumption. In co-operation, DTU Informatics and ENFOR A/S have developed a model described in Bacher (2012) to forecast the demand for space heating in a single-family house. The forecast model is based on observations of the space heat consumption in 16 single-family houses in Sønderborg in Denmark, knowledge of the heat dynamics of buildings, and weather forecasts of the ambient, i.e. outdoor, temperature (a low ambient temperature typically increases the heat demand), the global radiation (solar radiation entering through the
windows typically lowers the heat demand), and the wind speed (high wind speeds typically increase the ventilation of the house and thereby increase the heat demand).

If the solar energy available cannot cover the full heat demand of the house, the auxiliary heating will be activated by an advanced control system within the smart solar heating unit. To increase the fraction of solar energy usage in this smart solar heating unit and thereby reduce the use of the electricity based auxiliary heating, the advanced control system should activate the auxiliary heating only when solar energy is insufficient in covering the expected heat consumption and, if possible, only when electricity prices are favorable e.g. in windy periods or at night (Perers et al., 2009).

For optimal use environmentally and economically of the smart solar heating unit, the advanced control system needs forecasts of the expected heat production in the solar collectors and forecasts of the electricity prices. Based on weather forecasts of the ambient temperature and of the direct and diffuse components of the solar radiation (described in chapter 3) provided by DMI, DTU Informatics and ENFOR A/S have provided an approach described in Bacher et al. (2011) to obtain forecasts of the heat production in the solar thermal collectors. The solar thermal collectors are sensitive to the characteristics of the solar radiation, that is, the amount and incidence angle of the direct radiation and the amount of diffuse radiation received at the collectors (illustrated in figure 3.1). Using variable electricity prices in an advanced control system, might imply considerable savings at the customer level. Perers et al. (2009) estimate annual reductions in the costs of the auxiliary heating of 300 € at Nord Pool Spot prices. Nord Pool Spot is the Nordic electricity stock exchange, which sell electricity at prices that vary with fluctuating availability. Each day at 13:00 (local time), hourly electricity prices for the next 35 hours are determined.

From measured water temperatures within the tank, forecasts of the future heat demand of the house, forecasts of the disposable solar heat, and future electricity prices, the advanced control system developed by Innogie Aps should decide the optimal solution environmentally and economically of the water volume to heat, by which energy supply, and when. Central to the smart solar heating unit is the delivery of accurate forecasts of the solar radiation provided by DMI. These accurate predictions is the basis of my thesis.

\[http://www.nordpoolspot.com\]
3 Global Radiation

As solar shortwave radiation traverses the atmosphere of the Earth, interactions between the solar beam and the atmospheric constituents attenuates the solar radiation by either absorption or scattering of the radiation. Radiation that has been subjected to absorption is typically “lost” with respect to the amount of solar radiation penetrating the atmosphere. Radiation that has been scattered one or several times is referred to as diffuse. The remaining unabsorbed and unscattered radiation is called the direct radiation. The total downward solar radiation from the celestial sphere impinging upon a horizontal surface is called the global radiation.

The contributions to the diffuse radiation may vary with the position in the sky (Paltridge and Platt, 1976, Chap. 6), but is approximately isotropic (Paltridge and Platt, 1976, Chap. 6). The direct radiation varies likewise with the position in the sky, but its direction is definite and given by the zenith angle.

For a solar thermal collector, the direction of the direct beam relative to the normal of the solar collector plate is decisive for the amount of energy received at the solar collector. This is not the case for the diffuse radiation, which is approximately isotropic and therefore not sensitive to the slope of the solar collector. This is illustrated in figure 3.1. This difference in sensitivity to the slope of the solar collector between the direct and the diffuse radiation makes a distinction between them necessary.

In NWP models, detailed angular integrations of all the contributions to the diffuse radiation are not feasible. Typically in NWP models, and also the case for DMI’s HIRLAM models, the contributions to the diffuse radiation is limited to either upward or downward diffuse radiation.

3.1 Definitions

By use of Paltridge and Platt (1976, TABLE 3), the following components of solar radiation (or solar irradiance) in $\frac{W}{m^2}$ to be used throughout the thesis are defined below.

1 Radiation with wavelengths less than about 5 $\mu$m (Paltridge and Platt, 1976, TABLE 2.3), which constitutes most of the radiation from the Sun.
Figure 3.1: Illustration of the sensitivity of the direct radiation received at the solar collector to the slope of the collector relative to the direct radiation beam and the insensitivity of the diffuse radiation received at the solar collector to the slope of the collector; the two solar collectors receive (more or less) the same amount of diffuse radiation, but not the same amount of direct radiation.

- **Direct normal solar radiation,** $F_{\text{direct}}$:
  - solar radiation (or irradiance) that has not been subjected to scattering and is incident upon a surface perpendicular to the direction of the beam.

- **Direct horizontal radiation,** $F_{\text{direct}}^\downarrow$:
  - direct normal solar radiation incident on a horizontal surface, that is, $F_{\text{direct}}^\downarrow = F_{\text{direct}} \cdot \cos \theta$, when the angle of the direct solar beam and the normal to the surface is $\theta$ (see e.g. figure 3.2).

- **Diffuse horizontal solar radiation,** $F_{\text{diffuse}}^\downarrow$:
  - downward solar radiation (or irradiance) incident on a horizontal surface from a solid angle of $2 \pi$ (the sky) with the exception of the solid angle subtended by the Sun’s disc.

- **Global radiation,** $F^\downarrow$:
  - downward solar radiation (or irradiance) incident on a horizontal surface from a solid angle of $2 \pi$ (the sky). It is the sum of the direct horizontal, $F_{\text{direct}}^\downarrow$, and the diffuse horizontal radiation, $F_{\text{diffuse}}^\downarrow$.

Throughout this thesis, the direct horizontal radiation, $F_{\text{direct}}^\downarrow$ will be referred to as “direct radiation” and the diffuse horizontal radiation, $F_{\text{diffuse}}^\downarrow$ will be referred to as “diffuse radiation”.
3.2 Shortwave Radiation in DMI’s HIRLAM models

The attenuation of solar radiation in an absorbing and scattering medium — like the atmosphere — is described by a complex equation, “the transfer equation”, involving detailed wavelength and angular integrations of the radiation (Thomas and Stamnes, 1999). To obtain the computational speed required in NWP forecasting, some sort of approximation to the this highly detailed transfer equation is often needed — the degree of which depends on the problem at hand. The transfer of radiation within the atmosphere is in DMI’s operational NWP model calculated with a very fast radiation scheme (Savijärvi, 1990; Wyser et al., 1999), wherein the interaction between the solar shortwave radiation and the atmosphere is highly parametrised, and the diffuse radiation is treated as going either up or down. The radiation scheme is documented in Sass et al. (1994).

Since the distance between the Earth and the Sun has an annual change, so does the solar radiation at the Earth’s distance from the Sun. This solar radiation is parametrised in terms of the running day from January 1, \( d \), as

\[
S = S_0 \cdot (1 + 0.034221 \cdot \cos(\frac{2 \pi d}{365}) + 0.00128 \cdot \sin(\frac{2 \pi d}{365}) + 0.00719 \cdot \cos(2 \frac{2 \pi d}{365}))
\]

(Paltridge and Platt, 1976; Savijärvi, 1990). \( S \) is the extraterrestrial radiation normal to the solar beam (Badescu, 2008). The annual change in the Earth-Sun distance and thereby in \( S \) is about \( \pm 3.3 \% \) (Paltridge and Platt, 1976, Chap. 3), which amounts to about \( \pm 45 \% \). The quantity \( S_0 \) is referred to as the “solar constant” and is the total solar radiation (that is, integrated over all wavelengths) at the mean distance of the Earth from the Sun.
the Sun (Thomas and Stamnes, 1999, Chap. 9). It is, in fact, not a constant and a more appropriate term is “total solar irradiance” (Thomas and Stamnes, 1999). In the radiation scheme of DMI’s HIRLAM models, this total solar irradiance, $S_0$, is equal to 1365 $\text{W m}^{-2}$.

At a given solar zenith angle $\theta$ (defined in figure 3.2), the total downward solar radiation, that is, the global radiation (see e.g. chapter 3), received at the top of the atmosphere (toa), $F_{\text{down}}(\text{toa})$, before any attenuation by absorption or scattering of radiation, is

$$F_{\text{down}}(\text{toa}) = S \cdot \cos \theta$$  \hspace{1cm} (3.2)

by the cosine law. Before atmospheric absorption and scattering, this radiation is all direct and coming from the direction of the Sun. The zenith angle $\theta$ (or cosine of the zenith angle) on a given location is parametrised in terms of the latitude, $\Phi$, the solar declination, $\delta$, and the local hour angle, $T$ (described in Paltridge and Platt (1976, Chap. 3)), as

$$\cos(\theta) = \sin(\delta)\sin(\Phi) + \cos(\delta)\cos(\Phi)\cos(T)$$  \hspace{1cm} (3.3)

Paltridge and Platt, 1976; Savijärvi, 1990). The solar declination, $\delta$, is parametrised in terms of the running day from January 1, $d$, as

$$\delta = 0.006918 - 0.399912 \cdot \cos(\frac{2\pi}{365}d) + 0.070257 \cdot \sin(\frac{2\pi}{365}d) - 0.006758 \cdot \cos(2\frac{2\pi}{365}d) + 0.000907 \cdot \sin(2\frac{2\pi}{365}d) - 0.002697 \cdot \cos(3\frac{2\pi}{365}d) + 0.001480 \cdot \sin(3\frac{2\pi}{365}d)$$  \hspace{1cm} (3.4)

Paltridge and Platt, 1976, Chap. 3).

In a clear atmosphere, the diffuse radiation at the surface is found from the empirical formula

$$F_{\text{diffuse}} = 100 \cdot (1 - \exp(-2.865 \cdot h)),$$

where $h$ is the solar elevation, which is $90^\circ - \theta$. This equation an environmental adjustment of an equation in Paltridge and Platt (1976, Chap. 6). The direct radiation at the surface, $F_{\text{direct}}$, is the difference between the global radiation, $F_{\text{down}}(\text{surface})$, and the diffuse radiation, $F_{\text{diffuse}}(\text{surface})$ at the surface.

---

2 The mean distance of the Earth from the Sun is one astronomical unit, which is $1.5 \cdot 10^{11}$ m (Wallace and Hobbs, 2006).

3 The solar declination is the angle between the plane spanned by the Earth and the Sun and the celestial equator, which is a projection of the terrestrial equator onto a celestial sphere surrounding the Earth. It varies between $+23.5^\circ$ (on June 22) and $-23.5^\circ$ (on December 22) (Paltridge and Platt, 1976, Chap. 3).
Cloud-free atmosphere  In a cloud-free atmosphere, the total downward solar radiation received at the ground, $F_{\downarrow}^{\text{surface}}$, is obtained by reducing the radiation impinging upon the top of the atmosphere, $F_{\downarrow}^{\text{toa}} = S \cdot \cos \theta$ according to equation (3.2), by atmospheric absorption (mainly by ozone, water vapour, and absorbing aerosols) and scattering (mainly by atmospheric molecules and scattering aerosols) parametrised as

$$
F_{\downarrow}^{\text{surface}} = F_{\downarrow}^{\text{toa}} \left( 1 - 0.24 (\cos \theta)^{-0.5} - a_a \cdot 0.11 \left( \frac{u}{\cos \theta} \right)^{0.25} - a_s \cdot \frac{0.28}{1 + 6.33 \cos \theta} + a_s \cdot 0.07 \alpha \right)
$$

(Savijärvi, 1990). The parametrisation in equation (3.5) takes the following absorption and scattering into account: i) absorption by ozone of ultraviolet and visible solar radiation, the degree of which is determined by the solar zenith angle (and thereby by the path through the atmosphere) as $0.024 \cdot (\cos \theta)^{-0.5}$, ii) absorption of solar radiation by water vapour, CO$_2$, and O$_2$, which depends on the (scaled) water vapour amount $u$ and the solar zenith angle, $\theta$ as $0.11 \cdot \left( \frac{u}{\cos \theta} \right)^{0.25}$, iii) atmospheric scatter (in a cloud-free atmosphere this is scatter by atmospheric molecules — also referred to as Rayleigh scatter), which depends on the solar zenith angle as $\frac{0.28}{1 + 6.33 \cos \theta}$, and iv) downward scatter of the reflected radiation given by $0.07 \cdot \alpha$, when the surface albedo is $\alpha$. Reflection at the surface gives an upward radiation determined by the incident beam and the surface albedo, $F_{\downarrow}^{\text{toa}} \cdot \alpha$. Atmospheric scatter will re-direct a small portion ($F_{\downarrow}^{\text{toa}} \cdot \alpha \cdot 0.07$) of the reflected radiation back toward the surface.

Scattering and absorption by aerosols is included by enhancing the water vapour absorption by multiplying with $a_a$, and the atmospheric scattering by multiplying the atmospheric scatter with $a_s$. In DMI’s operational HIRLAM model, $a_a$ and $a_s$ are equal to 1.2 and 1.25, respectively, thereby increasing the absorption by 20 % and the scattering by 25 %. These values are found by fitting model calculations to observations and should be applicable to North European conditions (Savijärvi, 1990).

Cloudy atmosphere  The radiative properties of clouds depend strongly on the size of the cloud particles (Wyser et al., 1999). Compared to large cloud droplets, smaller cloud droplets increase the scatter and reflectivity of the cloud, so that more solar radiation is reflected back to space and less reaches the surface (Wyser et al., 1999). Transmission of radiation through a cloud is parametrised in terms of the vertically integrated cloud condensate (the total amount of water droplets and ice particles (in $\frac{g}{m^2}$)) within the
column of the model containing clouds, the cosine of the zenith angle, $\cos \theta$, and the effective radius, $r_e$, of the cloud particles. For radiation purposes, the effective radius is an appropriate measure of the average particle size in a distribution of spherical droplets in a water cloud (Wyser et al., 1999).

In calculating the effective radius, $r_e$, for a cloud, a distinction between cloud droplets and ice particles is made. At a given vertical level, only one effective radius is computed, which should then be representative of all cloud layers above the level in which $r_e$ is calculated. To obtain this, the effective radius is found as a weighted average, where the weights consist of the amount of cloud condensate at the level under consideration normalised by the total amount of cloud condensate from the top of the atmosphere to that level. The full parametrisation of shortwave radiation through clouds can be found in Wyser et al. (1999).
4 Data

4.1 Observations

24 of DMI’s pyranometers, situated on different locations in Denmark, measuring global radiation, form the observational basis for the evaluation and verification of global radiation forecasts treated in this work. Location and number of the pyranometers can be seen in figure 4.1. Data from the pyranometers are reported as an average of measurements during the preceding hour.

In the process of verifying model calculations of global radiation against observations, the first step has been to investigate the quality of the observations. For correct measurement of global radiation it is important that the pyranometer is in level (and not tilted toward or away from the Sun) and that the glass dome protecting the pyranometer is clear, see figure 4.2. To ensure the above, the pyranometers are inspected 2-12 times a year (Nordstrøm, 2005). There is in addition a continuous control of data from the pyranometers so that any evidently erroneous values are detected (John Cappelen, personal communication). This gives reasonable expectations to the measurements of DMI’s pyranometers, which are of the type Star-pyranometer made in Austria by P. H. Schenk. They are sensitive to radiation with wavelengths in the interval 0.3-3 µm (Nielsen, 2005). After two years of running operationally (which was the case in 2005), the uncertainty on hourly values was reported to be ±8.4 % at a confidence interval of 95 % (Nordstrøm et al., 2005). This value was found from propagation of uncertainty (Nordstrøm et al., 2005; Taylor, 1997) and a reported stability of 1 % per year of an operational pyranometer (Nielsen, 2005). Here, the uncertainty on the observations is estimated to be 10 %.

4.2 Matching Observations with Model Calculations

Forecasts from NWP models are typically calculated on a horizontal grid. The observations available to evaluate and verify the forecasts are located at points that are (most likely) not coincident with the forecast grid, and
Figure 4.1: Location and number of 24 of DMI's pyranometers
in such a case, it is necessary to select a method to match the observations to the forecast grid (Jolliffe and Stephenson, 2012, section 6.2). Two such methods exist: upscaling and downscaling. In upscaling, the observation is compared with the forecast value at the closest grid point thereby preserving the forecasted value and in downscaling, forecast values are calculated often by an interpolation method at the observation point. The result of matching forecast grid points with observation points depends on the method (upscaling or downscaling) being used (Jolliffe and Stephenson, 2012, Chap. 6). Likewise, the choice of downscaling (interpolation method) applied might influence the result of evaluation.

4.3 Model calculations

DMI is a member of the international program HIRLAM (High Resolution Limited Area Model), the aim of which is to develop, maintain, and improve NWP models for operational use by the participating institutions (information on the HIRLAM corporation can be found at http://www.hirlam.org). A HIRLAM model is a high-resolution limited area (that is, covering a small domain) NWP model. At DMI, the HIRLAM models exist in a number of versions differing in horizontal and vertical resolution as well as in geographical coverage.

In 2009, the operational HIRLAM model, S03 (with a horizontal resolution of 0.03°), was nested into the lower resolution HIRLAM model, T15 (with a horizontal resolution of 0.15°), which again was nested into the global low-resolution model, IFS, at ECMWF (European Centre for Medium-Range Weather Forecasts). In 2011, the resolution of IFS was comparable with the HIRLAM model T15, and the S05 model (with a horizontal resolution of 0.05°) used in DMI-EPS is now directly nested into the global
IFS model. At time of verification, the global IFS model had a horizontal resolution of 0.15° and 91 vertical layers, while the S03 and the S05 models had 40 vertical layers. For model S03, S05, and T15, a new forecast was initiated at 00, 06, 12, 18 UTC. See table 4.1 for details on model setup and figure 4.3 for the model domain of S03 and S05, which are (almost) identical.

In DMI’s HIRLAM models, calculations of the accumulated global, direct, and diffuse radiation are provided every hour. In taking differences of the accumulated radiation, values accumulated within the preceding hour are obtained.

### 4.4 Multidimensional Data

As often in meteorology, a data set contains some kind of multidimensionality — either by forecasts of several weather components at the same location or by the same weather component simultaneously forecasted at different locations, or a combination of this (Jolliffe and Stephenson, 2012). To match the global radiation measurements at the 24 pyranometers seen in figure 4.1, the data set — for either single or ensemble forecasts — consists of forecasts simultaneously predicted at 24 points, and this introduces multidimensionality into the data set.
<table>
<thead>
<tr>
<th>Model name</th>
<th>S05</th>
<th>S03</th>
<th>T15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of vertical layers</td>
<td>40</td>
<td>40</td>
<td>40</td>
</tr>
<tr>
<td>Horizontal resolution</td>
<td>0.05°(≈ 5 km)</td>
<td>0.03°(≈ 3 km)</td>
<td>0.15°(≈ 16 km)</td>
</tr>
<tr>
<td>Time step</td>
<td>150 s</td>
<td>150 s</td>
<td>400 s</td>
</tr>
<tr>
<td>Forecast length</td>
<td>54 hours</td>
<td>54 hours</td>
<td>60 hours</td>
</tr>
<tr>
<td>Number of forecasts per day</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Global model</td>
<td>IFS</td>
<td>IFS</td>
<td>IFS</td>
</tr>
</tbody>
</table>

**Table 4.1:** Model setup at time of verification
5 Performance of Global Radiation Forecasts

The delivery of accurate forecasts of global, direct, and diffuse radiation to the solar heating unit developed at the Technical University of Denmark initiated an evaluation of the solar radiation forecasts from DMI’s HIRLAM models. This evaluation had not previously been performed at DMI and was expected to unveil problems and successes in the model calculations. The performance of global radiation forecasts was evaluated against observations at the 24 DMI pyranometers.

5.1 Data

5.1.1 Observations

The period of verification was April 1 to June 1 2009. Inspection of the observations of this period revealed some obviously incorrect values: i) night values of a few $\frac{W}{m^2}$ over long periods of time and ii) individual values that are unrealistically high and which for some reason slipped through the data control. To compensate for these errors in the observations, night values of global radiation of a few $\frac{W}{m^2}$ have been set equal to zero and unrealistically high values of global radiation have been marked as invalid.

5.1.2 Model Calculations of Global Radiation

To investigate global radiation forecasts from the — at time of verification — operational HIRLAM model S03, a time series matching the observations have been constructed from a 24 hour long forecast for each 00 UTC forecast within the time period April 1 to June 1. Matching forecasts with observations, the method of upscaling (see e.g. section 4.2) has been applied.

It is highly challenging in model calculations to capture small scale clouds. Since they give rise to locally different atmospheric conditions and hence global radiation, some discrepancies between observations and model calculations are expected. To confine these discrepancies, a distinction between clear days and non-clear days is made. April 2009 was in Denmark
characterized by sunny weather and is therefore considered appropriate for collecting a statistical sample of clear days 1.

5.2 Comparison of Model Calculations with Observations

Figure 5.1 shows the measured and calculated global radiation at the verifying location 6188 (Sjælsmark in Zealand) (see e.g. figure 4.1) on four days in April. April 25 was a clear day and exhibits the characteristics — a smooth curve peaking at noon — typical of clear days. In general, on clear days, and also valid for April 25, model calculations fit the observations within the uncertainties.

As expected, on cloudy days, the fit between model calculations and observations is reduced. This is illustrated in figure 5.2 for some days in May. The model tends to overestimate the global radiation compared to the observations, which might be related to HIRLAM’s tendency to underestimate the cloud cover (Bent Hansen Sass, personal communication). On May 23 and 24, the shape of the curve representing model calculations (red) resembles the curve of the observations (blue), but the timing is a couple of hours off.

On April 26 in figure 5.1, the forecasted global radiation at noon exceeds the “noon-radiation” on clear days. This is the case for several verifying locations, but here only shown for station 6188. These values are unrealistically high and do not agree with observations, which show a depression in the

1 A “clear day” is here loosely defined as a prevailing clear day, since actually only few days during sunny April 2009 were completely cloud-free in the whole country from sunrise to sunset.
Figure 5.2: Global radiation at the verifying location 6188 (Sjælsmark); *blue*: observations; *cyan*: uncertainty on observations (±10.0%); *red*: model calculations

Figure 5.3: Calculations from DMI’s operational model of diffuse (*thick red*) and direct (*thin green*) radiation at the verifying location station 6188 (Sjælsmark); Yellow, vertical bars mark clear days.

global radiation in the late morning in consistency with satellite pictures\(^2\), which show a cloud cover over northern Zealand around that time.

### 5.3 Comparing Model Calculations with DTU Observations

Decomposing the global radiation into its direct and diffuse parts and analyzing them separately is expected to give a deeper understanding of shortcomings of the radiation scheme, some of which are displayed on April 26 in figure [5.3]. According to figure [5.4], the peak in the calculated global

\(^2\)www.sat.dundee.ac.uk
radiation at noon on April 26 consists of only diffuse radiation (the direct radiation is zero).

At DTU, global, direct and diffuse radiation are measured at the weather station at the Department of Civil Engineering at DTU. They have most kindly placed their radiation data from 1 April to 1 May 2009 at disposal for a comparison with DMI’s diffuse and direct radiation computations.

In figure 5.4 and 5.5, this peak in global radiation on April 26 with zero direct radiation doesn’t agree with DTU observations, which do show a depression around noon in the direct radiation (in accordance with satellite pictures), but which have a value of about 150 $\text{W/m}^2$.

Running test cases showed that the radiation scheme of HIRLAM gives 100% diffuse radiation in case of a complete cloud cover — regardless of the optical thickness of the cloud.

Disagreements between observations and model calculations is also seen
on April 27 in figure 5.4 and 5.5. However, on April 27 the model is to some extent able to capture the shape of the observations for both direct and diffuse radiation, although the model in general overestimates both quantities compared to the observed DTU values.

Another point to be made from figure 5.4 is that on certain days (e.g., on April 28), the model calculations result in a diffuse radiation typical of a clear day, while DTU observations show elevated diffuse radiation, characteristic of a day with some cloud cover. Satellite pictures confirm the presence of clouds over Denmark for most of that day. The “clear day” seen in the model calculations might be related to HIRLAM’s tendency to underestimate the cloud extent and hence overestimate the global radiation reaching the ground (Bent Hansen Sass, personal communication).

5.4 A Revision of the Radiation Scheme in DMI’s HIRLAM models

In general, two problems of the radiation scheme have been encountered from this evaluation of global radiation forecasts: i) events of high values of global radiation consisting of only diffuse radiation, which do not agree with observations and ii) events of “clear days” in terms of radiation pattern that disagree with observations.

Since the commence of this evaluation, test cases with a revised radiation scheme by Bent Hansen Sass and Kristian Pagh Nielsen have been performed to deal with i) above. Before the implementation of the changes to the shortwave parametrisations, they made all radiation entering (and leaving) a cloud layer diffuse, and in case of a complete cloud cover, all radiation became diffuse. The revision has enabled a correction of a few errors within the radiation scheme — which should rectify the cases of very high values of global radiation — and has led to a change in the shortwave parametrisations to allow a transmission of direct radiation through a cloud layer depending on the integrated amount of cloud condensate, $W$, within the cloud, the effective radius, $r_e$, of the cloud particles, and the solar zenith angle, $\theta$.

The major change in the shortwave parametrisations is to assume an exponential attenuation of the direct radiation traversing a cloud of optical thickness $\tau_{\lambda,z}$ as

$$F_{\text{direct (surface)}} = F_{\text{direct (at cloud top)}} \cdot \exp\left(-\frac{\tau_{\lambda,z}}{\cos \theta}\right)$$

3www.sat.dundee.ac.uk
4The optical thickness is a measure of the number and strength of optically active particles — the either absorb or scatter radiation along the radiation beam (Thomas and Stamnes, 1999, Chap. 2).
5The relationship between an optical path, $\tau_{\lambda}$, related to the direction of the solar beam, and an optical thickness, $\tau_{\lambda,z}$, related to the vertical axis of a typical NWP model, is given by $\tau_{\lambda} = \frac{\tau_{\lambda,z}}{\cos \theta}$.
(confer Beer’s Law on exponential attenuation of a monochromatic beam of radiation entering an absorbing and/or scattering medium (Paltridge and Platt, 1976; Thomas and Stamnes, 1999) — described in appendix A). The optical thickness, $\tau_{\lambda,z}$, of a cloud can in the visible part of the solar spectrum be approximated by $\tau_{\lambda,z} \approx \frac{3W}{2r_e}$ (Stephens, 1984), which gives

$$ F_{\text{direct}}(\text{surface}) = F_{\text{direct}}(\text{at cloud top}) \cdot \exp\left(-\frac{3W}{2r_e \cos \theta}\right). \quad (5.1) $$

Multiplying $F_{\text{direct}}(\text{surface})$ with $\cos \theta$ gives $F_{\text{direct}}^{\downarrow}(\text{surface})$.

A combination of a large value of integrated cloud condensate, $W$, and a small effective radius, $r_e$, (i.e. a large water content spread over many small cloud droplets) effectively impedes the direct radiation passing through the cloud layer. The cloud is said to be optically thick. The opposite — a small value of $W$ and a large effective radius, $r_e$, (i.e. a small water content spread over a few large cloud droplets) — allows, to a large extent, transfer of direct radiation through the cloud. By comparison, the cloud is in this case optically thin. This is in accordance with results obtained by Wyser et al. (1999, Figure 2) showing the influence of the integrated cloud condensate, $W$, and the effective radius, $r_e$, on the transmission of global radiation through a water cloud.

This new parametrization (equation (5.1)) ensures that optically thin clouds can transfer direct radiation to the ground — also in case of a complete cloud cover. This is illustrated in figure 5.6 where the direct (horizontal) radiation — normalised by the value for zero cloud condensate — received at the surface, $F_{\text{direct}}^{\downarrow}(\text{surface})$, is plotted against $W$ in a case of a $100 \%$ cloud cover. A value of integrated cloud condensate, $W$, below about 50 $g m^{-2}$ allows direct radiation at the ground. With the old radiation scheme, this was not possible, since radiation leaving a cloud was considered as diffuse and in case of a complete cloud cover, all radiation became diffuse.

A highly detailed radiative transfer model, the DISORT model (described in Stamnes et al. (1988)), has been run by Kristian Pagh Nielsen and is included in figure 5.6 for a comparison with the revised radiation scheme of DMI’s HIRLAM model. The resemblance between the two is highly satisfying and indicates of a considerable improvement and success of the HIRLAM shortwave parametrisations of direct radiation. Note that the direct radiation from the old radiation scheme would be zero for any value of $W$ in case of a complete cloud cover. This revised radiation scheme is now implemented in DMI’s HIRLAM models.

A justification for implementing a simple exponential attenuation of the polychromatic (here, visible) direct radiation (as Beer’s law describes for monochromatic radiation entering an absorbing and/or scattering medium — described in appendix A) can be found from the following considerations. When direct radiation enters a cloud, cloud droplets and ice particles attenuate the incident radiation primarily by scatter (Stephens, 1984). In
Figure 5.6: Normalised direct radiation as a function of the integrated amount of cloud condensate, \( W \), within a 100% cloud cover from 500 to 1000 m above the surface. The zenith angle is 30° and the surface albedo 0.35. The red full line shows results of the revised radiation scheme and the green dashed line the results of the DISORT model — a highly detailed radiation model in which the number of wavelengths and solid angles in which to discretise equation (A.6) can be chosen.

the visible part of the spectrum, this scatter is to a first approximation independent of wavelength. This condition that all visible wavelengths are scattered equally makes (most) clouds appear white. The transmission of direct radiation through a cloud can (at least in the visible part of the solar spectrum) be considered independent of wavelength, and this makes the application of a simple exponential attenuation of the direct radiation incident at the top of the cloud possible.

The remedy of point ii) above is an improvement in the prediction of clouds — their location, their horizontal and vertical extent, and their exact timing. All this may highly affect the global radiation received at a pyranometer or a solar collectors. The exact location of a cloud, its extent and exact timing might be some of the most difficult parameters to forecast in NWP modelling. Clouds may even appear on such a small scale, that they cannot (yet) be resolved in present NWP models despite their increasing resolution.
6 DMI’s Ensemble Prediction System

An NWP forecast is in general subject to forecast uncertainties (or forecast errors) occurring from a combination of i) uncertainties in determining the initial state of the atmosphere arising from inaccuracy and incompleteness in the observations of the atmosphere (Epstein, 1969), and ii) uncertainties arising from deficiencies in the NWP model formulation (Leith, 1974). The information contained in a deterministic forecast is therefore best appreciated if complemented with estimates of these uncertainties in the forecast value.

Deficiencies in NWP models relate both to the errors introduced by the difference in resolution between a numerical model and the real atmosphere — illustrated in figure 6.1 — and to physical processes in the atmosphere appearing on such small scales, that they cannot be resolved by the NWP model (Wilks, 2006, Chap. 6). These physical processes must be parametrized in terms of variables that can be resolved by the model, and since these small-scale processes are often not fully determined by the resolved variable, all parametrizations might introduce uncertainties in the NWP formulation (Wilks, 2006, Chap. 6).

Figure 6.1: Illustration of the difference between a real world and the world as represented by an NWP model. From Wilks (2006, FIGURE 6.19)
6.1 Ensemble Prediction

Forecast uncertainties arising from model deficiencies are addressed in developing NWP models and ensemble prediction has (until recently) mainly focused on forecast uncertainties related to uncertainties in determining the initial state in an NWP model.

Because of the uncertainty in determining the initial state of the atmosphere, a multitude of states will all be consistent with the observations. In the high-dimensional phase space of a typical NWP model — in which an estimated state of the atmosphere is a point and its time evolution is along a trajectory in the phase space to another point — this multitude of states can be associated with a probability distribution function (PDF) characterizing the probability of each of these initial atmospheric states. At any time (initial or future), a PDF can be ascribed to the multitude of possible atmospheric states — a multitude in which the true atmospheric state is one. From the initial time to a given forecast time, the initial PDF transforms into a forecast PDF quantifying the uncertainty in the forecast.

A feasible way of addressing forecast uncertainty is by running an ensemble of forecasts initiated from slightly perturbed conditions. At the initial time, the PDF is assumed to be represented by a finite sample of all possible initial states (Molteni et al., 1996). Each of the ensemble members is then integrated forward in time by an NWP model, and at any forecast time, the properties of the PDF are assumed to be described by the ensemble members (Molteni et al., 1996).

Figure 6.2 attempts to illustrate the evolution of both the PDF of a two-dimensional phase space and the ensemble members representing the PDF. At “T = 0”, the cyan cross represents the atmospheric state from which a traditional forecast would be initiated. This point in phase space represents only one in a number of atmospheric states all consistent with the uncertainty in the initial state. The PDF of these states is represented by the small ellipse (Wilks, 2006, Chap. 6). A time integration moves the three initial states through phase space along trajectories, which in this case diverge. From “T = 0” to “T = 24”, the PDF expands implying less agreement between the three ensemble members on the atmospheric state at “T = 24”. The shape, or spread, of the PDF is related to the uncertainty in the forecast. A small ensemble spread would generally imply little uncertainty in the forecast value and thereby a high confidence in it. Conversely, a large ensemble spread would imply large uncertainty and low confidence in the forecast value. In the high-dimensional phase space of a typical NWP model, some directions are associated with states that with time diverge and to sample the forecast PDF appropriately, a degree of dispersion of the ensemble members is necessary (Wilks, 2006, Chap. 7). In figure 6.2, if at “T = 0” the top black cross and the blue cross were sampled as an ensemble, their dispersion at “T = 24” would be too small to represent
Figure 6.2: A simple illustration of the time evolution in a two-dimensional phase space (from 0 to 24 hours) of the true state of the atmosphere (red), and an ensemble prediction system consisting of three members: the standard forecast initiated from the “best estimate” of the initial conditions \( T = 0 \) and two forecasts started from atmospheric states consistent with the uncertainty in the initial state (black). The ovals indicate the PDF and from left to right, its time evolution is illustrated.

The figure is inspired by a figure in: http://www.metoffice.gov.uk/research/areas/data-assimilation-and-ensembles/ensembles/ensemble-forecasting/explanation
the forecast uncertainty and to capture the true state of the atmosphere.

Generating ensemble members that at all time (initial and future) describe the features of the PDF, that is, behave as a random sample of the PDF, is the ideal goal in ensemble prediction (Gneiting et al., 2008; Tracton and Kalnay, 1993). If 1) the applied NWP model is accurate and 2) if the ensemble members at the initial time is a random sample of the PDF of the analysis (Gombos and Hansen, 2007), that is, if the ensemble represents the uncertainty in the analysis (Toth and Kalnay, 1993), the ensemble members will approximate the PDF (Molteni et al., 1996; Wilks, 2006) and be able to capture its features. Requirement 2 above ensures that at the initial time, the ensemble members approximates the initial PDF, while requirement 1 ensures that after integrating these ensemble members forward in time, they will approximate the PDF at the forecast time. If the above two requirements are fulfilled, the true state of the atmosphere will behave as one more member of the ensemble, that is, the true state of the atmosphere will be statistically indistinguishable from the ensemble members (Wilks, 2006, Chap. 7). This condition, that the true atmospheric state behaves statistically like any of the ensemble members, is called consistency of the ensemble (Wilks, 2006, Chap. 7). A degree of ensemble dispersion is a prerequisite of ensemble consistency (Wilks, 2006, Chap. 7).

With no a priori information on the uncertainties in the initial state, generating ensemble members that reflect this uncertainty is challenging.

### 6.1.1 Initial Condition Perturbations

A number of methods differing in complexity of generating the initial ensemble members exist, but they all submit to the following form illustrated for ensemble member $y_j$

$$y_j(t = 0) \equiv y_0(t = 0) + \delta y_j(t = 0),$$  \hspace{1cm} (6.1)

(Buizza et al., 1999), where $y_0$ is the unperturbed control forecast and $\delta y_j(t = 0)$ is the initial condition (IC) perturbation of the control forecast $y_0$. The control forecast at time $t = 0$, $y_0(t = 0)$, is the most recent analysis. A challenging in ensemble prediction is to generate ensemble perturbations that reflect the uncertainty in the analysis.

A model integration of equations of the type

$$\frac{\partial y_j}{\partial t} = A(y_j, t) + P(y_j, t),$$  \hspace{1cm} (6.2)

where $A$ and $P$ are the dynamic (non-parametrised) and physical (parametrised) processes, respectively, of the ensemble member $y_j$ from the initial time and

1 The analysis is the estimated state of the atmosphere that serves as the initial conditions of an NWP forecast.
to a time $t$ gives

$$y_j(t) = \int_{t=0}^{t} [A(y_j, t) + P(y_j, t)]dt.$$ 

Here, $\frac{\partial y_j}{\partial t}$ in equation (6.2) could be the time derivative of either the velocity components as governed by the equation of motion (or momentum) originating from Newton’s second law, of the temperature governed by the thermodynamic energy equation, or the density governed by the continuity equation. Combined, these three governing equations are often referred to as the primitive equations. “$A$” represents the causes of a change as described by the primitive equations and “$P$” represents additional forces from atmospheric processes that are less well described in a typical NWP model. These could be heating by solar radiation, changes of both heat and velocity components due to convection, and changes in velocity components due to turbulence and surface friction.

### 6.1.2 Model Perturbations

Until recently, the main focus in developing ensemble prediction systems (EPSs) has been on medium-range forecasts (Molteni et al., 1996; Tracton and Kalnay, 1993) that typically cannot resolve local features of the weather. Increased computer resources has created an interest in addressing forecast uncertainties in the local weather by developing short-range, high resolution, limited-area EPSs. These are now evolving rapidly in many centres (Bowler et al., 2008). The use of short-range, high-resolution EPSs has necessitated an additional simulation of deficiencies in the NWP model formulations (Bowler et al., 2008; Buizza et al., 1999). This concerns requirement 1 above. Generating an EPS that takes into account both the uncertainty in the analysis and the imperfections of the NWP model poses an additional challenge to ensemble prediction. Within the last decade, there has been enhanced focus on the effect of model deficiencies on forecast uncertainty, and on how to use ensemble prediction to address these model imperfections. Dealing with model errors in EPS is, however, thought to be an even greater challenge than simulating initial condition perturbations (Bowler et al., 2008).

One approach to simulate forecast uncertainty associated with model deficiencies is to include, besides the initial condition perturbations, multiple NWP models, which could, for example, consists of two or more competing model schemes describing some physical parametrisations (Bowler et al., 2008; Buizza et al., 1999; Feddersen, 2009; Wilks, 2006). Another approach to simulate this uncertainty is to introduce random perturbations to these physical parametrisations (Bowler et al., 2008; Feddersen, 2009; Wilks, 2006). These can be obtained by perturbing the total tendency $\frac{\partial y_j}{\partial t}$.
in equation (6.2) of each ensemble member $y_j$ in the following way
\[ \frac{\partial y_j}{\partial t} = A(y_j, t) + P(y_j, t) + r_j(\lambda, \phi, z, t) \cdot P(y_j, t). \] (6.3)

The term $r_j(\lambda, \phi, z, t)$ is a stochastic perturbation coefficient with $\lambda$ and $\phi$ as horizontal coordinates and $z$ as the vertical coordinate. To limit its growth, the time evolution in the perturbation coefficient $r_j(\lambda, \phi, z, t)$ is generated by an auto-regression model\(^2\) as
\[ r_j(\lambda, \phi, z, t + T) = a \cdot r_j(\lambda, \phi, z, t) + s_j, \] (6.4)

where $a$ is a coefficient and $s_j$ is a uniformly distributed random number.

A time integration then gives for ensemble member $y_j$
\[ y_j(t) = \int_{t=0}^{t} [A(y_j, t) + P(y_j, t) + r_j(\lambda, \phi, z, t) \cdot P(y_j, t)] dt \] (6.5)

starting, (with $t = 0$), from the perturbed initial condition of equation (6.1).

### 6.2 Construction of DMI-EPS

At DMI, a short-range, high resolution ensemble prediction system (DMI-EPS) has during 2009 been developed and verified for the prognostic parameters temperature, wind speed, and precipitation [Feddersen, 2009].

#### 6.2.1 Initial Condition Perturbations

Because of the limited forecast period of short-range, high-resolution ensemble forecasts, and more complex methods in which the model perturbations are developed in the early part of a forecast (as the singular-vector approach used at ECMWF [Molteni et al., 1996]) may not be suitable and other means of producing the ensemble perturbations must be applied (Bowler et al., 2008). In DMI-EPS, a scaled lagged average forecasting (SLAF) method (described in Ebisuzaki and Kalnay [1992]), to generate the initial perturbations, is used. In the SLAF method, the difference between a very-short-range forecast predicting the present analysis and this analysis (here denoted the forecast error) is used as perturbations ($\delta y_j(t = 0)$) of the control analysis ($y_0(t = 0)$) in equation (6.1) to estimate the uncertainty in the analysis. The perturbations have the following form
\[ \delta y_j(t = 0) = \pm \alpha_n(y_n(n) - y_0(t = 0)) \] (6.6)

\(^2\)An auto-regression model is, loosely speaking, a model to predict an outcome of a system at a time $T + t$ from the outcome of the system at a previous time $t$. 

where \( y_0, (n) \) is an \( n \) hour old and unperturbed (control) forecast propagated \( n \) hours forward in time to predict the present analysis.\(^4\) Equation (6.1) can together with equation (6.6) be translated to

\[
\text{initial condition} = \text{analysis} \pm \alpha_n (\text{forecast}_n \text{ hour old}(n) - \text{analysis}) \quad (6.7)
\]

(Feddersen, 2009).

An old forecast generally has larger forecast errors (it has become less skillful compared to a young forecast. Not to let these larger forecast errors influence the estimates of the uncertainty in the analysis, a scaling factor \( \alpha \) is introduced to control the magnitude of the forecast error, so that larger forecast errors of older forecasts are damped more than smaller forecast errors of younger forecasts (Feddersen, 2009; Toth and Kalnay, 1993). Presently, a 6 hour old and a 12 hour old forecast generate through equation (6.7) four perturbed initial conditions in addition to the unperturbed control forecast, which have the form

\[
\begin{align*}
\text{IC}_0 &= \text{analysis} \\
\text{IC}_{6^+} &= \text{analysis} + \alpha_6 \cdot (\text{forecast}_{6 \text{ hour old}}(6) - \text{analysis}) \\
\text{IC}_{6^-} &= \text{analysis} - \alpha_6 \cdot (\text{forecast}_{6 \text{ hour old}}(6) - \text{analysis}) \\
\text{IC}_{12^+} &= \text{analysis} + \alpha_{12} \cdot (\text{forecast}_{12 \text{ hour old}}(12) - \text{analysis}) \\
\text{IC}_{12^-} &= \text{analysis} - \alpha_{12} \cdot (\text{forecast}_{12 \text{ hour old}}(12) - \text{analysis}).
\end{align*}
\]

(6.8)

The scaling factors were at time of verification \( \alpha_6 = 0.80 \) and \( \alpha_{12} = 0.56 \) (Feddersen, 2009).

### 6.2.2 Model perturbations

In DMI-EPS, the perturbation coefficient, \( r_j(\lambda, \phi, z, t) \), in equation (6.3) is allowed to vary horizontally, but not vertically, so \( r_j(\lambda, \phi, z, t) = r_j(\lambda, \phi, t) \). To limit calculations, each horizontal domain \( D \) has its own perturbation coefficient, \( r_j \), and random number, \( s_j \), (associated with the ensemble member \( y_j \)) – denoted by \( \langle \ldots \rangle_D \) in equation (6.9), and \( T \) is the time interval between the updates of \( r_j \). This gives the perturbation coefficient

\[
r_j(\lambda, \phi, t + T) = a(r_j(\lambda, \phi, t))_D + \langle s_j \rangle_D.
\]

(6.9)

Values presently employed in DMI-EPS can be found in table 6.1. In order to prevent the perturbation coefficients from escalating in the course of a forecast, \( r_j \) is limited to the interval \([-0.5, 0.5]\) as indicated in table 6.1.

---

\(^4\)Here \( n \), which denotes a previous forecast that predicts the present analysis, is separated from \( t \), which denotes a time integration from the present analysis and into the future, i.e. a forecast.
Parameter | Value
--- | ---
$T$ | 45 min
$D$ | $53 \times 53$ grid points
$s_j$ | $\in \{-0.15, 0.15\}$
$a$ | 0.9
$r_j$ | $\in \{-0.5, 0.5\}$

Table 6.1: Values used to construct the perturbation coefficients (Feddersen, 2009)

<table>
<thead>
<tr>
<th>STRACO</th>
<th>KF/RK</th>
<th>STRACO</th>
</tr>
</thead>
<tbody>
<tr>
<td>no stoc. phy.</td>
<td>stoc. physics</td>
<td>no stoc. phy.</td>
</tr>
<tr>
<td>IC1 = IC₀</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>IC2 = IC₆⁺</td>
<td>2</td>
<td>7</td>
</tr>
<tr>
<td>IC3 = IC₆⁻</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>IC4 = IC₁₂⁺</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>IC5 = IC₁₂⁻</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td>group</td>
<td>I</td>
<td>II</td>
</tr>
</tbody>
</table>

Table 6.2: Configuration of the 25 ensemble members in DMI-EPS; IC: initial conditions; use and no use of stochastic physics ("stoc. phys." and "no stoc. phy.") and application of a perturbed roughness length ("pert. rough"); the computation of the initial conditions can be seen in equation (6.8).

In DMI-EPS, the total tendencies $\frac{\partial y}{\partial t}$ in equation (6.3) of the four three-dimensional model variables of temperature, wind speed, humidity and cloud water are randomly perturbed according to equation (6.3) and (6.9).

To simulate model deficiencies, each of the five initial conditions emerging from the application of equation (6.6) and shown in equation (6.8) is combined with two different cloud schemes — the STRACO cloud scheme (documented in Sass (2002)) and the Kain-Fritsch/Rasch-Kristjánsson (KF/RK) cloud scheme (documented in Kain (2004); Rasch and Kristjánsson (1998)) — yielding a total of 10 ensemble members. Model deficiencies are further simulated by subjecting each ensemble member to stochastic perturbations, which increases the number of ensemble members to 20. According to Feddersen (2009), five ensemble members (members 21 to 25) have been dedicated to studying the impact of perturbing the roughness lengths for urban areas in addition to the application of stochastic physics. At the start of a forecast, a roughness length in meters in the interval $\{0.05, 1.1\}$ is chosen at random for each of the ensemble members 21 to 25. The roughness length for the other 20 ensemble members (1 to 20) is equal to 1. An overview of the configuration of all 25 ensemble members can be seen in table 6.2.
7 A Framework for Forecast Verification

The purpose of forecast verification is to assess the quality of forecasts. Forecast quality describes the association between the forecasts and the corresponding observations for a number of forecast-observation pairs (Wilks, 2006). Forecast quality can be assessed in a number of ways each describing a different aspect of the quality of forecasts. Different approaches of assessing these aspects exist involving both measures and graphical displays.

Approaches of assessing forecast quality have been applied to both individual ensemble members of DMI-EPS as well as their derivatives as the ensemble mean and the ensemble median.

Aspects of forecast quality are described in section 7.1. In section 7.2, a description of some methods to address quality of ensemble forecasts is given.

7.1 Aspects of Forecast Quality

7.1.1 Traditional Measures of Forecast Quality

The more traditional measures of forecast quality applicable to continuous forecasts of a scalar variable (Wilks, 2006, Chap. 7) includes forecast bias, forecast accuracy, and forecast skill.

Bias

Forecast bias describes any systematic deviation between a forecast $y$ and a matching observation $x$ and is often measured by the mean error (ME), which is given by

$$ ME = \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i). \quad (7.1) $$

In the forecast example illustrated in table 7.1, the forecasts display no bias since the difference between the mean of the observations and the mean of the forecasts is zero.
Accuracy

Forecast accuracy is often measured by the mean squared error (MSE) or its related root mean squared error (RMSE), which are defined as

\[
\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i)^2 \\
\text{RMSE} = \sqrt{\text{MSE}}
\]

(7.2)

respectively. Both ME, MSE, and RMSE will be equal to or above zero. They are all negatively oriented, so that small values are preferred to large values, and for perfect forecasts MSE, and RMSE, are all equal to zero. In the forecast example in table 7.1, the forecasts are not completely accurate, since neither MSE (nor RMSE) is zero. In taking the square in the MSE, and hence in the RMSE, large deviations between forecasts and observations will be highly penalised.

Skill

Forecast skill is often measured by a forecast skill score, SS, which is a relative accuracy measure, that is, the accuracy of a set of forecasts relative to a set of unskillful reference forecasts (Jolliffe and Stephenson 2012; Wilks 2006). For a given set of forecast-observation pairs, the skill score for a measure of accuracy, \( A \) (which could be MSE or RMSE) with respect to the accuracy measure for a set of unskillful reference forecasts, \( A_{\text{ref}} \), is

\[
\text{SS} = \frac{A - A_{\text{ref}}}{A_{\text{perf}} - A_{\text{ref}}} \cdot 100 \%
\]

(7.3)

where \( A_{\text{perf}} \) is the value of the accuracy measure that would be obtained for a set of perfect forecasts (Wilks 2006, Chap. 7). The unskillful reference forecasts might be climatology or persistence (Wilks 2006, Chap. 7). Skill scores, as defined here, can be interpreted as a percentage improvements over the unskillful reference forecasts.

In this work, a skill score based on MSE is generated. According to equation (7.3), the skill score can be written as

\[
\text{SS} = \frac{\text{MSE} - \text{MSE}_{\text{ref}}}{\text{MSE}_{\text{perf}} - \text{MSE}_{\text{ref}}} \cdot 100 \% = \left( 1 - \frac{\text{MSE}}{\text{MSE}_{\text{ref}}} \right) \cdot 100 \%.
\]

(7.4)

The last equation follows from MSE being zero for perfect forecasts, that is \( \text{MSE}_{\text{perf}} = 0 \). If the MSE for the forecasts being evaluated equals the MSE for perfect forecasts, that is, \( \text{MSE} = \text{MSE}_{\text{perf}} \), the skill score attains its maximum value of 100 %. If the forecast MSE equals the MSE for the unskillful reference forecasts, that is, \( \text{MSE} = \text{MSE}_{\text{ref}} \), SS = 0 %, implying
Table 7.1: A simple example illustrating forecasts with corresponding observations; the values could be mm of rain.

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2.0</td>
<td>3.0</td>
</tr>
<tr>
<td>2.0</td>
<td>0.0</td>
</tr>
<tr>
<td>2.0</td>
<td>4.0</td>
</tr>
</tbody>
</table>

that the forecasts contain no improvement compared to the unskillful reference forecasts \(\text{Wilks, 2006, Chap. 7}\). If the forecasts being evaluated are inferior to the unskillful reference forecasts, the skill score becomes negative, so \(SS < 0\%\) \(\text{Wilks, 2006, Chap. 7}\).

7.1.2 General approach to Forecast Verification

The framework of verification introduced by \(\text{Murphy and Winkler, 1987}\), offers a general approach to forecast verification with access to other aspects of forecast quality than those described above. This approach of verification is based on the joint distribution of forecasts and observations. Denoting observations by \(x\) — and a specific value by \(X\) — and forecasts by \(y\) — and a specific value by \(Y\), \(p(x, y)\) denotes the joint distribution of all relevant values of \(X\) and \(Y\). This joint distribution \(p(x, y)\) contains information about the forecasts, the observations, and the relationships between the forecasts and the observations \(\text{Murphy et al., 1989}\). For a given data set, the joint distribution \(p(x, y)\) is represented by a distribution of the relative occurrences (or relative frequencies) of the simultaneous events of forecasting a specific value \(Y\), while observing a specific value \(X\) for all relevant combinations of \(X\) and \(Y\).

Any joint distribution can be factored into a conditional and a marginal distribution in two ways \(\text{Murphy and Winkler, 1987; Wilks, 2006}\). The information contained in the joint distribution, \(p(y, x)\) becomes more accessible, when it is factored in either of these ways \(\text{Murphy and Winkler, 1987; Murphy et al., 1989; Wilks, 2006}\).

One factorization is the calibration-refinement factorization, which has the form

\[
p(x, y) = p(x|y) \cdot p(y)
\]

\(\text{Murphy et al., 1989; Wilks, 2006}\). This factorization of \(p(x, y)\) involves the conditional distribution \(p(x|y)\), which is the distribution of the observations given a forecast value \(y\), and the marginal distribution of the forecasts \(p(y)\). The other factorization is called the likelihood-base rate factorization and has the form

\[
p(x, y) = p(y|x) \cdot p(x)
\]

\(\text{Murphy et al., 1989; Wilks, 2006}\).
It involves the conditional distribution \( p(y|x) \) of the forecasts given an observed value \( x \) and the marginal distribution of the observations \( p(x) \).

**Reliability**

Forecast reliability (or forecast calibration) is related to the conditional distribution of the observations associated with a specific forecast value, \( Y \), \( p(x|y = Y) \) in equation (Murphy 1993).

Forecasts possess good reliability by a high degree of similarity between the mean of the observations associated with a specific forecast \( Y \), \( E(x|y = Y) \), and that forecast value \( Y \) (Murphy et al. 1989). According to Murphy (1993), evaluation of reliability can assess a question like this: “Does the mean observed temperature on those occasions on which the predicted temperature is 10°C correspond to that value?”. If

\[
E(x|y = Y) = Y \quad \text{for all} \quad Y, \tag{7.7}
\]

where \( E(x|y = Y) \) is the expected (or mean) value of the observations given a specific forecast value \( Y \), the forecasts are said to be completely reliable (or perfectly calibrated) (Murphy and Winkler 1987; Murphy et al. 1989). In table 7.1, the forecasts are completely reliable, since the mean of the observations on those occasions on which the predicted amount of rain is 0.2 equals 0.2.

In forecasting, two types of bias might be present: i) unconditional (or systematic or overall) bias and ii) conditional bias, that is, a forecast bias that depends on (or is conditional on) the forecast value. A presence of either of these biases imply a lack of reliability of the forecasts, which can be appreciated from the following considerations by use of the reliability criterion in equation (7.7). In a case of 10 forecasts, \( Y_1, \ldots, Y_{10} \), if

\[
E(x|y = Y) > Y \quad \text{for} \quad Y_1, Y_2, Y_3, Y_4, \text{ and } Y_5
\]

and

\[
E(x|y = Y) < Y \quad \text{for} \quad Y_6, Y_7, Y_8, Y_9, \text{ and } Y_{10} \tag{7.8}
\]

the forecasts values \( Y_1, \ldots, Y_{10} \) are said to be conditionally biased (Murphy et al. 1989) and are not completely reliable according to equation (7.7). If

\[
E(x|y = Y) > Y \quad \text{for} \quad Y_1, \ldots, Y_8, \tag{7.9}
\]

the forecasts \( Y_1, \ldots, Y_8 \) suffer from an unconditional (or a systematic or an overall) bias and the forecasts, \( Y_1, \ldots, Y_{10} \), are not completely reliable — again with reference to equation (7.7). That is, either unconditional biases or conditional biases lead to forecasts that are not completely reliable.
Forecast Observation  \[ p(y|x = X) \]
\[ \begin{array}{c|c|c}
 2.0 & 1.0 & p(y = 2.0|x = 1.0) = 0.25 \\
 2.0 & 3.0 & p(y = 2.0|x = 3.0) = 0.25 \\
 2.0 & 0.0 & p(y = 2.0|x = 0.0) = 0.25 \\
 2.0 & 4.0 & p(y = 2.0|x = 4.0) = 0.25 \\
\end{array} \]

Table 7.2: A simple example illustrating completely reliable forecasts with no discriminative power and no sharpness; the values could be mm of rain.

**Sharpness**

Forecast sharpness, or refinement, is an attribute of the forecasts and is a characteristic of the marginal distribution of the forecast values \[ p(y) \] in equation (7.5) (Wilks, 2006, Chap 7). The distribution of forecasts indicates how often different forecast values are issued. A forecasting system that produces the same forecast on each forecast occasion is not sharp (Murphy et al., 1989). In table 7.2, the forecasts completely lack sharpness since only one value, the average of the observations, is forecast at all forecast occasions. Within the field of meteorology, sharp forecasts can easily be produced. The challenge is to ensure that these forecast values correspond to the subsequent observed values (Wilks, 2006, Chap. 7). Except for very simple forecasts as shown in table 7.1 that contain no variation and completely lack sharpness, evaluation of sharpness is usually applied to probability forecasts of a dichotomous (or binary) event and complete sharpness might be difficult to define for forecasts of continuous observations (Murphy et al., 1989). However, for perfectly accurate and reliable forecasts the distribution of forecast values, \[ p(y) \], should be identical to the distribution of the observed values, \[ p(x) \] (Murphy, 1993; Murphy et al., 1989).

**Discrimination**

Forecast discrimination measures the ability of the forecasts to discriminate between observations that differ (Murphy, 1993; Weigel and Manson, 2011) regardless of the actual forecast value. Discrimination is a fundamental quality attribute since it indicates the usefulness of a set of forecasts after being appropriately post-processed (Murphy, 1993; Weigel and Manson, 2011). Different measures of discrimination exists. Perfectly reliable forecasts are effectively useless if they lacks discriminative power (Weigel and Manson, 2011), which is illustrated by the simple example in table 7.2, which is an extension of table 7.1. The forecasts are completely reliable since \[ E(x|y = 2.0) = 2.0 \] for all values of \( Y \) (in this case, there is just one value), but completely lack discriminative power, since the forecasts cannot discriminate between different observations and cases of rain and no rain.
Table 7.3: A comparison of two forecast-observations pairs; in square brackets, the corresponding rank of the observation and of the forecast is seen.

<table>
<thead>
<tr>
<th>Observation</th>
<th>Forecast</th>
<th>t = 1</th>
</tr>
</thead>
</table>

Within the framework of [Murphy and Winkler (1987)] discrimination is a characteristic of the conditional distribution of the forecasts for a specific value of the observations $X$, that is, $p(y|x = X)$. In table 7.2, the value of $p(y|x = X)$ is the same (0.25) for all four values of $X$. That is, a forecast value of 2.0 mm is with equal probability followed by an observation of either 1.0 mm, 3.0 mm, 0.0 mm, or 4.0 mm. When $p(y|x = X)$ is the same for all values of $X$, the forecasts are not able to discriminate between the different observations ([Murphy et al. 1989]). To clarify, an example of forecasts that are perfectly discriminatory, but not completely reliable is given in appendix B. The forecast value is actually never observed. The forecasts would probably not be considered very informative despite their complete reliability, since they are all equal to the mean of the observations, and contain no variation. In real life, deductions about forecasts from only four forecast-observation pairs is, of course, questionable.

Another measure of discrimination is the generalised discrimination score proposed by [Manson and Weigel (2009)] with application to binary, categorical, and continuous observations and based on the joint distribution $p(x, y)$. In constructing the generalised discrimination score, $D$, all possible sets of two forecast-observation pairs are constructed from the verification data. For each of these sets, the question is asked whether the forecasts can be used to successfully distinguish (or rank) the observations ([Weigel and Manson, 2011]). In table 7.3, two forecast-observations pairs are compared, and the forecasts are able to correctly discriminate (or rank) the corresponding observations while in table 7.4, this is not the case. The proportion of times when two forecasts can be used to correctly rank the corresponding observations yields the generalised discrimination score, $D$, given by

$$D = \frac{1}{2}(\tau + 1),$$  \hspace{1cm} (7.10)

([Weigel and Manson, 2011]), where $\tau$ is Kendall’s rank correlation coefficient ([Bhattacharyya and Johnson, 1977, Chap. 15]), which measures the degree of association between two variables. Since discrimination is not concerned with the forecasts taken at face value, a rank correlation coefficient will suffice in defining the generalised discrimination score.

By its definition, the generalised discrimination score, $D$, indicates how often the forecasts correctly discriminate the observations ([Weigel and Manson, 2011]). If the forecasts do not contain any useful information, the proba-
Table 7.4: A comparison of two forecast-observations pairs; in square brackets, the corresponding rank of observations and forecasts is seen.

<table>
<thead>
<tr>
<th>Day</th>
<th>Continuous case Observations</th>
<th>Continuous case Forecasts</th>
<th>Binary case Observations</th>
<th>Binary case Forecasts</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t = \text{July 10 14.00 UTC}$</td>
<td>22°C</td>
<td>23°C</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$t = \text{July 11 14.00 UTC}$</td>
<td>23°C</td>
<td>21°C</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 7.5: An example of forecasts and corresponding observations for continuous and binary observations.

The probability that the forecasts correctly discriminate two observations is equivalent to random guessing (which would be equal to 50 %) and one would obtain $D = 0.5$ (in this case there is no association between forecasts and observations, and so $\tau = 0$). The more successfully the forecasts are able to discriminate the observations, the closer the score is to 1, since a strong association (or correlation) between forecasts and observations will give $\tau$ close to 1. Forecasts that consistently rank the observations in the wrong way, will give $D = 0$ (in this case, a strong negative association between forecasts and observations will yield $\tau = -1$).

When the number of distinct observations increases — as from binary to categorical and to continuous observations, the discriminative power usually decreases (Weigel and Manson, 2011). This can be explained by the additional precision that is required to discriminate between continuous observations compared to binary observations (Weigel and Manson, 2011), which may be defined as observations being above or below some threshold. In the continuous case, the forecasts will have to successfully discriminate between values that might differ by only small amounts (Weigel and Manson, 2011). As an example, table 7.3 should illustrate the difference in precision required for continuous observations compared to binary observations. The table shows (fictitious) national average of maximum temperatures and corresponding observed events for two days in July. In creating the binary events, temperatures above the assumed average for July is 15°C make an event (1) and temperatures below 15°C no event (0). In the continuous case, the forecasts will have to correctly discriminate and therefore correctly forecast, which of the two days is the warmest. In the binary case, the forecasts will (only) have to discern between days when the national average temperature is above or below 15°C, which would (probably) be easier to forecast correctly.
7.2 Quality of Ensemble Forecasts

Methods to assess quality of ensemble forecasts exist. Multidimensional ensemble forecasts of continuous observations can be difficult to verify in a comprehensive manner (Jolliffe and Stephenson, 2012, Chap. 1). In this work, reliability and discrimination have been applied to ensemble forecasts of continuous observations.

7.2.1 Consistency

If the ensemble members can be considered as random samples of the PDF both at the initial time and at any later forecast time, the true state of the atmosphere will behave like one more member of the ensemble and the true atmospheric state will be statistically indistinguishable from the ensemble members (Gombos and Hansen, 2007; Wilks, 2006, Chap. 7). This condition is called consistency of the ensemble (Wilks, 2006, Chap. 7).

A common approach to assess the degree of consistency of ensemble forecasts, that is, to assess whether the ensemble members and the truth (materialised by the verifying observation) are statistically indistinguishable is by the shape of a scalar rank histogram (also known as a verification rank histogram or a Talagrand histogram) (Gombos and Hansen, 2007; Wilks, 2006). From consistent ensemble forecasts, questions like: “What is the probability that the daily accumulated global radiation tomorrow exceeds a given threshold?” or a more distributional question like: “Within what interval of global radiation values does 90 % of the forecasts lie in 12 hours?” can be answered. These questions, that is, the forecasts, will be reliable. The derivation of reliable probability forecasts is central to ensemble forecasting (Hamil and Colucci, 1997), and evaluation of ensemble consistency has over the years received special attention (see e.g. Gneiting et al. (2008); Hamill (2001); Jolliffe and Stephenson (2012); Wilks (2006)).

In the literature, consistency and reliability are used interchangeably to describe the statistical indistinguishability of the true state of the atmosphere and the ensemble members. Here, consistency describes this condition, while reliability characterises the reliability of the probability forecasts derived from the ensemble.

In constructing the scalar rank histogram, consider $N$ ensemble forecasts, each of which consists of $M$ ensemble members, and $N$ corresponding observations (Jolliffe and Stephenson, 2012, Chap. 8). Now, for each of these $N$ ensemble forecasts with corresponding observation, the value of the observation is ranked according to the values of the ensemble members. If the observation is smaller than all $M$ ensemble members, then its rank is 1, and if the observation is larger than all ensemble members, then its rank is $M + 1$ (Jolliffe and Stephenson, 2012, Chap. 8). If the $M$ ensemble members and the verifying observation belong to the same distribution (the PDF),
that is, if ensemble members and the verifying observation are statistically indistinguishable, the rank of the observation within the \( M + 1 \) values would be equally likely to take on any of the rank values 1, 2, \ldots, \( M + 1 \). (Jolliffe and Stephenson, 2012; Wilks, 2006). Determining the rank of the observation with respect to the ensemble members for \( N \) observations and \( N \) ensemble forecasts (of each \( M \) forecasts) gives when displayed in a histogram the rank histogram. If the ensemble forecasts are consistent, the ensemble members and the observations are statistically indistinguishable (Wilks, 2006, Chap. 7), and the histogram of the \( N \) observation ranks will be uniform with an expected number of \( \frac{N}{M+1} \) counts per rank value apart from deviations due to sampling variability (Jolliffe and Stephenson, 2012; Wilks, 2006).

A uniform scalar rank histogram implies that the ensemble forecast and the verifying observation are random samples from the PDF — i.e. they are statistically indistinguishable. However, sampling the ensemble members from a distribution different from the PDF, can under special circumstances result in a uniform rank histogram (Hamill, 2001) and the interpretation of rank histograms should be treated carefully. Flatness of rank histograms is a necessary, but not sufficient criterion for ensemble consistency (Feddersen, 2009; Jolliffe and Stephenson, 2012; Wilks, 2006).

Deviations from uniformity of the rank histogram can be used to diagnose deficiencies of the ensemble consistency (Hamill, 2001; Jolliffe and Stephenson, 2012; Wilks, 2006). Bias in the ensemble (or systematic (or overall) deviations between the ensemble members and the observation) will result in a sloped rank histogram, since too, often the observation will be on one side of the ensemble, that is, either too often smaller than or too often larger than the ensemble forecasts, and this will too often occupy the smallest or the largest rank values (Hamill, 2001; Wilks, 2006 — see figure 7.1, where the ensemble forecasts have a tendency to over-forecast leaving the observations too often at one of the smallest rank values. An under-dispersive ensemble, will result in a U-shaped rank histogram, since too often the ensemble is not able to capture the observation, which will too often be either smaller than or larger than the ensemble forecasts (Hamill, 2001; Wilks, 2006) as may be caused by the presence of a conditional forecast bias among the ensemble members. This under-dispersion will overpopulate both the smallest and the largest rank values resulting in a rank histogram with a U-shaped appearance.

### 7.2.2 Multidimensional Consistency

The scalar rank histogram is in essence a verification tool to assess the consistency of ensemble forecasts of a scalar variable.

For the data set treated here, a multidimensionality is introduced by the 24 simultaneous forecasts corresponding to the 24 locations of the verifying observations (see e.g. section 4.3). To access the consistency of en-
ensemble forecasts of multidimensional data, a number of methods have been proposed. Two such methods are the multivariate rank (MVR) histogram (Gneiting et al., 2008; Jolliffe and Stephenson, 2012) and the minimum spanning tree (MST) histogram (Gneiting et al., 2008; Jolliffe and Stephenson, 2012) described below.

**Multivariate Rank Histogram**

The multivariate rank (MVR) histogram is a generalisation of the scalar rank histogram and assess — like their scalar counterpart — how the observations rank with respect to the individual ensemble members. Being a generalisation of the scalar rank histogram, the shape of the MVR histogram can be interpreted in the same way as the shape of the scalar rank histogram. In constructing MVR histograms, the challenge is to find a definition of how to rank multidimensional vectors (Jolliffe and Stephenson, 2012). Below, a procedure of determining the rank of a $K$-dimensional observation with respect to the $M$ $K$-dimensional ensemble forecasts, is suggested (Gneiting et al., 2008; Jolliffe and Stephenson, 2012). In the $K$-dimensional space spanned by the $K = 24$ verifying locations simultaneously predicted, the $K$-dimensional vector, $x_{i,1}, \ldots, x_{i,k}, \ldots, x_{i,K}$, is denoted by $X_i$ (with $i = 1, \ldots, N$) and the ensemble forecast $Y_{i,j}$ denote the $K$-dimensional vector, $y_{i,j,1}, \ldots, y_{i,j,k}, \ldots, y_{i,j,K}$ is denoted by $Y_{i,j}$ (with $i = 1, \ldots, N$). Table 7.6 attempts to illustrate how this work. Here, with $K = 4$ and $M = 3$, the $K$-dimensional observation vector, $X_i$ would be equal to $(25, 23, 22, 24)$, and the $K$-dimensional vector for ensemble member 1, $Y_{i,1}$, would be $(28, 27, 25, 25)$, and for ensemble member 2, $Y_{i,2}$, the $K$-dimensional vector would be $(25, 22, 20, 25)$, and for ensemble member 3, $Y_{i,3}$,
At time point $i$: forecast date 2011-08-06, initial time 00, and lead time 6 h.

<table>
<thead>
<tr>
<th>Station</th>
<th>Member 1</th>
<th>Member 2</th>
<th>Member 3</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>6019</td>
<td>28</td>
<td>25</td>
<td>24</td>
<td>25</td>
</tr>
<tr>
<td>6031</td>
<td>27</td>
<td>22</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>6041</td>
<td>25</td>
<td>20</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>6049</td>
<td>25</td>
<td>25</td>
<td>24</td>
<td>25</td>
</tr>
</tbody>
</table>

$V_1$, $V_2$, $V_3$, $V_0$

**Table 7.6:** Illustration of the $K$-dimensional vectors for an ensemble of three members at a given time point $i$ corresponding to a certain forecast date, say 20110806, and a certain initial time, say 00, and a certain lead time, say 6 hours, which corresponds to the time 6 August 2011 06 UTC.

the $K$-dimensional vector would be $(24, 23, 21, 24)$ for the given combination of forecast date, initial time, and forecast length, $i$.

The procedure is then as follows (Jolliffe and Stephenson, 2012, Chap. 8):

1. Define $V_0 = X_i$ and $V_j = Y_{i,j}$ with $j \in 1, \ldots, M$. Now, $V_{0,k}$ denotes the $k$th dimension of the observations and $V_{j,k}$ the $k$th dimension of the $j$th ensemble member.

2. Then, determine the “pre-rank”, $\rho_j$, of each vector $V_j$ with $j \in 0, 1, \ldots, M$ by

$$\rho_j = 1 + \sum_{l=0}^{M} q_{j,l} \text{ with } q_{j,l} = \begin{cases} 1 & \text{if } V_{j,k} \geq V_{l,k} \text{ for all } k \in 1, \ldots, K \\ 0 & \text{otherwise} \end{cases}$$

(7.11)

that is, if the $K$-dimensional ensemble vector $V_j$ is larger than the $K$-dimensional ensemble vector $V_l$ for all $K$-dimensions (in table 7.6 $K = 4$), the pre-rank, $\rho_j$, of ensemble member “$j$” is increased by 1. In the example illustrated in table 7.6 the pre-rank of ensemble member 1 is 4, since $V_1$ is clearly larger than any of the other vectors, the pre-rank of the observations is 3, while the pre-ranks for ensemble member 2 and 3 are both 1.

3. Finally, the multivariate rank $r$ is determined. If the pre-rank of the observation, $\rho_0$, is not tied, then $r$ is given by $\rho_0$. If there are ties, these are solved at random. In the example in table 7.6 the pre-rank of the observations is in this case not tied, so $r = 3$.

Repeating these steps for $N$ forecast-observation pairs and displaying the $N$ multivariate ranks of the observations, $r$, in a histogram, yields the MVR histogram.
Minimum Spanning Tree Histogram

Another method to evaluate the reliability of multidimensional ensemble forecasts is the minimum spanning tree (MST) histogram (Gneiting et al., 2008; Jolliffe and Stephenson, 2012). It is likewise applicable if a number of locations are simultaneously predicted by ensemble of forecasts.

When plotting a $K$-dimensional ensemble forecast in a diagram, a spanning tree can be constructed by connecting all $K$ forecast points with line segments without generating any closed loops (Jolliffe and Stephenson, 2012, Chap. 8). There are several ways of generating a spanning tree (of connecting the points in figure 7.2 without any closed loops), but the one that gives the smallest sum, when adding the length of all the line segments, defines the minimum spanning tree (MST). This is best illustrated when the dimension of the data is 2 as in figure 7.2. Here, $K = 2$, and $M = 25$.

For a consistent EPS, the ensemble members should be statistically indistinguishable from the verifying observation, which according to Jolliffe and Stephenson (2012, Chap. 8) implies that the $K$ dimensional distance of the observation from any of the ensemble members, should on average be similar to the average mutual distance of the ensemble members from each other. This relation is tested by the MST histogram. The statistical indistinguishability implies that the MST lengths should — on average — not be significantly affected if a random ensemble member is replaced by the verifying observation (Jolliffe and Stephenson, 2012). In figure 7.2, the distance of the observation from any of the ensemble members seems, on average, to exceed the average mutual distance of the ensemble members.
The procedure for generating MST is according to Jolliffe and Stephenson (2012, Chap. 8): the total length, $l_0$, of this minimum spanning tree spanned by the $M = 25$ ensemble members in the $K = 24$-dimensional space, is generated. Now, replacing the $j$th ensemble member by the verifying observation gives another total MST length, $l_j$. For consistent ensemble forecasts, $l_0$ should not be systematically larger or smaller than any $l_j \in \{l_1, \ldots, l_M\}$, that is, replacing any one of the ensemble members by the verifying observation should not affect the associated total MST length. Ranking $l_0$ within the set of MST lengths $\{l_0, l_1, \ldots, l_M\}$, it should be equally likely for the rank of $l_0$, which is denoted by $r$, to take on any of the values $\{1, \ldots, M + 1\}$. Whether this is the case can be assessed by determining $r$ for all forecast-observation pairs and then displaying the values obtained in a histogram — the MST-histogram. If the forecasts are reliable, the MST histogram should be uniform (Jolliffe and Stephenson, 2012, Chap. 8).

MST histograms can generally not be considered as a multidimensional generalization of rank histograms, as is the case of the multivariable rank histogram, and they need to be interpreted differently (Jolliffe and Stephenson, 2012). Under-dispersive ensemble forecasts, which are often associated with U-shaped rank histograms, often yield negatively sloped MST histograms (Jolliffe and Stephenson, 2012).

### 7.2.3 Discrimination

Discrimination of ensemble forecast has often been accessed probabilistically, that is, after the ensemble has been transformed into probability forecasts (Weigel and Manson, 2011). A generalised discrimination score, $D$, applicable to ensemble forecasts of continuous observations has been proposed by Weigel and Manson (2011) as an expansion of the generalised discrimination score proposed by Manson and Weigel (2009) and described in section 7.1.2.

The construction of a generalised discrimination score, $D$, applicable to ensemble forecasts of a continuous observation, requires a definition of how to discriminate, or rank, two ensemble forecasts. For single deterministic forecasts, it is trivial to decide which one of two forecasts is larger and should (if the forecasts possess discriminative power) be associated with the larger one of the two corresponding observations (see e.g. table 7.3 and 7.4). This decision is less obvious for ensemble forecasts. Weigel and Manson (2011) consider 3 hypothetical 5-member ensemble forecasts of temperature ($^\circ$C) with $y_1 = (22, 23, 26, 27, 32)$, $y_2 = (28, 31, 33, 24, 36)$, and $y_3 = (24, 25, 26, 27, 28)$ (from Weigel and Manson (2011)). Most people would intuitive label $y_2$ larger than both $y_1$ and $y_3$. The situation is less obvious when comparing $y_1$ and $y_3$. A method of ranking ensemble forecasts proposed by Weigel and Manson (2011) is described in appendix C and makes the generalised discrimination score applicable to ensemble forecasts.
7.2.4 Summary of Aspects of Forecast Quality

In table 7.7, a summary of the aspects of forecasts quality described in this chapter can be found. A possibility of the quality aspect to handle multidimensional forecasts is indicated in the fourth and the fifth column for single forecast and ensemble forecasts, respectively.
<table>
<thead>
<tr>
<th>Aspect</th>
<th>Measure</th>
<th>Definition</th>
<th>Multidimensionality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>ME</td>
<td>Describes the difference between mean forecast and mean observation</td>
<td>yes*</td>
</tr>
<tr>
<td>Accuracy</td>
<td>MSE,</td>
<td>Describes the average correspondence between individual pairs of forecasts and observations</td>
<td>yes*</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Skill</td>
<td>SS</td>
<td>Describes the accuracy of forecasts relative to the accuracy of unskillful reference forecasts</td>
<td>yes*</td>
</tr>
<tr>
<td>Reliability**</td>
<td></td>
<td>Describes the correspondence between the mean observation associated with a particular forecast and that forecast</td>
<td>no</td>
</tr>
<tr>
<td>Discrimination</td>
<td>D</td>
<td>Describes the ability of the forecasts to discriminate between different observations</td>
<td>no</td>
</tr>
<tr>
<td>Sharpness</td>
<td></td>
<td>Describes the variability of the forecasts</td>
<td>yes*</td>
</tr>
</tbody>
</table>

Table 7.7: Aspects of forecast quality with short definitions from Murphy [1993]; the aspect of forecast quality is shown in column one, a typical measure of this quality aspect is shown in column two, a short definition of the aspect is given in column three, and in the fourth and fifth column, the application of the forecast aspect to multidimensional data sets is indicated for single forecasts in column four (single forecasts) and to ensemble forecasts in column five (ensemble forecasts). Here, an asterisk (*) indicates that the forecast aspect can handle multidimensional data, but only implicitly, that is, in the way the given measure or aspect of forecast quality is defined here. This is the case for the ME, MSE, and RMSE. A double asterisk (**) indicates that concerning ensemble forecasts, consistency (rather than reliability) of the ensemble, from which reliable probability forecasts can be derived (Wilks [2006, Chap. 7]), is evaluated.
8 Verification of Global Radiation from DMI-EPS

To comply with accuracy requirements to solar radiation forecasts arisen from the development of a smart solar heating unit, DMI-EPS has since August 2011 made hourly calculations of global, direct and diffuse radiation. With the aim of assessing the possibility of complementing forecasts of global radiation with uncertainty estimates as derived from ensemble forecasts of DMI-EPS, this chapter is devoted to verifying these ensemble forecasts.

8.1 The Data

The period of verification is from August 5 2011 to November 14 2011, which amounts to 102 days and the ensemble forecasts are verified against global radiation measurements at the 24 DMI pyranometers.

8.1.1 Ensemble Calculations of Global Radiation

The data set consists of 100 days of forecasts (from August 5 2011 to November 12 2011) with 4 forecasts per day each consisting of 54 hourly values. This gives $N = 21600$ forecast occasions with corresponding observations within the period of verification from August 5 to November 14 2011. In verifying the S05 ensemble forecasts of global radiation against DMI’s 24 observation stations, the forecasts have been downscaled by bilinear interpolation (see e.g. section 4.3) to the locations of the 24 verifying observations (that is, the 24 pyranometers shown in figure 4.1), which gives a data set of 24 simultaneous ensemble forecasts. The number of ensemble forecasts, $M$, is 25 (see e.g. table 6.2).

In generating ASCII files from the GRIB format of the DMI-EPS calculations, discontinuities in the data series have been introduced due to occasional storage problems. For all included time points, action has been taken to ensure that data for all 25 ensemble members exist, which makes $N = 19002$. 

50
8.1.2 The Observations

Inspection of the observations covering the period of verification have revealed a few obviously incorrect values: i) night values of a few $\frac{W}{m^2}$ over long periods of time and ii) individual values that are unrealistically high, and which for some reason slipped through the data control — described in section 4.1. To compensate for the type i) errors, night values above zero have simply been set equal to zero and concerning errors of type ii), the “top of the atmosphere radiation” global radiation, $F_{\downarrow}^{\text{toa}}$, has been used to evaluate the ground measurements, and all values above $F_{\downarrow}^{\text{toa}}$ have been marked as invalid.

8.1.3 The Clearness Index

Global radiation reaching the surface of the Earth consists of a part governed by astronomical parameters giving rise to diurnal and seasonal variations and another part governed by short-term weather parameters such as frequency and height of clouds and their optical properties, atmospheric aerosols, atmospheric water vapour, and ground albedo (Badescu, 2008). The global radiation reaching the ground can be considered as a sum of two components — a deterministic (or astronomical) component and a stochastic (or meteorological) component. This meteorological or stochastic component can be isolated by introducing the clearness index defined by

$$K_c = \frac{F_{\downarrow}^{\text{surface}}}{F_{\downarrow}^{\text{toa}}},$$

where $F_{\downarrow}^{\text{toa}}$ is the extraterrestrial horizontal solar radiation (see e.g. equation 3.2). $F_{\downarrow}^{\text{surface}}$ is the global radiation received at the surface of the Earth (Badescu, 2008). The clearness index $K_c$ then accounts for all meteorological influences on the global radiation (Badescu, 2008) and has been calculated in addition to the global radiation.

With this definition of the clearness index, it cannot be defined at night times, when $F_{\downarrow}^{\text{toa}}$ is zero. To eliminate any possible problems with model calculations for very high zenith angles (or for very low solar elevations), these invalid night values are expanded to zenith angles greater than 88°.

To avoid having night values — when the global radiation is zero for both observations and model calculations (i.e. a trivial forecast difference of zero) — influence the verification of global radiation from DMI-EPS, the undefined night values of the clearness index have been used to determine undefined night values of the global radiation, and, except if otherwise stated, these undefined night values are excluded from the verification.

In validating global radiation calculations from DMI-EPS, the clearness index data have been applied, when the elimination of the astronomical signal in the data was important.
8.2 Assessing the Performance of the Ensemble Mean

By averaging the ensemble members, some of the errors present in the individual forecasts should cancel, and the ensemble mean should, in general, provide an improved forecast compared to a traditional forecast (Toth and Kalnay, 1993; Wilks, 2006). Hence, the performance of the ensemble mean has been investigated.

In phase space, the ensemble mean will be the center of the distribution of all the ensemble members — at least until at clustering of the ensemble members appear (Wilks, 2006, Chap. 6), which is the case at the final forecast time in figure 8.1. In case of clustering, the ensemble mean will in phase space not be near any of the ensemble members (Wilks, 2006, Chap. 6), that is, the atmospheric state represented by the ensemble mean might be far from any predicted states. Being a more robust measure of centre, the ensemble median might in case of clustering do better compared to the ensemble mean. Depending on the distribution of the ensemble forecasts, the ensemble median is therefore speculated to be able to surpass the ensemble mean. Hence, both are calculated and together with the control forecast subjected to verification methods with the aim of investigating and comparing the performance of the three forecasts. Of course, the ensemble mean is not a real forecast, but a quantity derived from the ensemble. When sorting the 25 ensemble members, the ensemble median will be equal to the 13th member with 12 members above and 12 members below. In investigating the three forecasts, both the more traditional measures of forecast performance such as ME, RMSE, and skill score as well as aspects of forecast performance...
such as sharpness, reliability, and discrimination are addressed.

8.2.1 Sharpness

Recall from section 7.1.2 that for continuous observations, complete sharpness of the forecasts is difficult to define, but that for completely accurate and reliable forecasts, the distribution of the forecasts $p(y)$ should be identical to the distribution of the observations $p(x)$ (Murphy et al., 1989).

In figure 8.2, the relative frequency distribution of the observations, $p(x)$, is seen in (a) and the relative frequency distribution, $p(y)$, of the control forecast is seen in (b), the ensemble mean in (c), and the ensemble median in (d). Here, the clearness index has been used to emphasize features in the distributions by a change of both the distribution and the scale of data. Although not identical, the distributions of clearness indices, bear an overall resemblance. However, the distribution of the ensemble mean, the control forecast, and the ensemble median are all characterized by a higher degree of symmetry as opposed to the corresponding distribution of the observations. High values of clearness index (greater than $\approx 0.8$) found in the observations, are not seen in either of the three forecasts: the control forecast, the ensemble mean, and the ensemble median — best illustrated in the boxplot in figure 8.3. A more uniform distribution is exhibited by the ensemble mean and to some degree also by the ensemble median compared to the other two distributions. Smoothing by taking means and medians (not so pronounced for the ensemble median) seems to reduce the variability found in the observations and better captured by the control forecast. All three forecasts seem relatively broad and covers most values of the observations, but not values above $\approx 0.8$. The distributions of the control forecast resembles the distribution of the observations more compared to the other two distributions, and therefore seems more sharp than both the ensemble mean and the ensemble median.

8.2.2 Bias, MSE, RMSE, and Skill Score

The more common measures of forecast quality as ME, MSE, and RMSE, are defined as

$$\text{ME} = \frac{1}{n} \sum_{n=1}^{n \in N} (y_n - x_n)$$ (8.2)

$$\text{MSE} = \frac{1}{n} \sum_{n=1}^{n \in N} (y_n - x_n)^2$$ (8.3)

$$\text{RMSE} = \sqrt{\text{MSE}}$$ (8.4)
Figure 8.2: Distribution of the clearness index for the observations (a), the control forecast (b), the ensemble mean (c), and the ensemble median (d) for the \( N \) forecast occasions and the \( K \) locations of the verifying observations; the distributions have been normalized.
The clearness index is a measure of how well a forecast matches the observed data. It is calculated according to equation (7.1) and (7.2), but here \( n \) is the number in the total number of \( N = N \cdot K \) time and space combinations in the data set over which to average — in figure 8.3 this is determined by the forecast lead time —, and \( y_n \) and \( x_n \) are the forecast and the observation, respectively, at time-space point \( n \). Both the measure of forecast bias (ME), the measure of forecast accuracy (MSE and RMSE), (and forecast skill score) are applicable to scalar forecasts (Wilks, 2006, Chap. 7) (see e.g. chapter 7). They are all here defined to incorporate the multidimensionality of the data set by averaging over a pooled data (sub)set consisting of both time and space points.

In generating the ME and the RMSE, illustrated in figure 8.4, the following procedure has been applied — here illustrated for the ensemble mean:

1. the ensemble mean is calculated for all \( N \) time-space combinations, and

2. for the ME the difference, and for the MSE the squared difference between the ensemble mean and the matching observation is calculated at each of the \( N \) time-space combinations, and

3. the average of this quantity is calculated for each lead time, that is, over \( n \) time-space combinations.

Taking the square root of the MSE gives the root mean squared error, RMSE. In calculating the bias and the MSE for the ensemble median, the median is...
calculated in point 1 above instead of the ensemble mean and in calculating the bias and the MSE for the control forecast, point 1 is skipped. Note that in squaring the deviation between a forecast and the corresponding observation, the MSE and the RMSE become highly sensitive to large deviations in the data set. In forecasting accurately the global radiation received at solar collector, penalising large deviation between a forecast and a corresponding observation is desirable.

In general, the biases in figure 8.4(a) are modest. A strong declining trend is seen in the first $\approx 12$ hours of the forecast lead time. This is related to spin-up effects, which might appear in the first hours of an NWP forecast as, for instance, clouds build up. From a turning point occurring at a forecast lead time of about 16 hours for the control forecast and 24 hours for the ensemble mean and the ensemble median an increasing trend is seen. The control forecast with its much faster rise towards zero performs better compared to both the ensemble mean and the control forecast. Over forecast lead times of 6 – 48 hours, the biases are negative with the ensemble mean displaying the largest absolute bias. This suggest a degree of, although modest, overall or systematic under-forecasting for all three forecasts, but most pronounced for the ensemble mean and the ensemble median. This corresponds with the figures in the “ME” column of table 8.1.

Values of RMSE found in this study and displayed in figure 8.4(b) are comparable with values reported in [Diaz et al. (2012), in Lorenz et al. (2011)] for single deterministic forecasts (compare for example with the high-resolution bias-corrected HIRLAM-CI data, the pixel-averaged WRF-MT data, or the direct model output WRF-UJAEN data), and in [Lorenz et al. (2012)]. This lends support to the three forecasts: the control, the ensemble mean, and the ensemble median from DMI-EPS. From the figure, accuracy is seen to decrease with lead time. The control forecast shows a larger RMSE for all lead times compared to both the ensemble median and the ensemble mean, which has the lowest RMSE. The overall values are shown in the “RMSE” column in table 8.1.

The strong diurnal variation in the squared forecast deviation symmetric around 12 UTC in figure 8.5 combined with the 6 hour interval between two consecutive forecasts (issued at 00, 06, 12, 18 UTC every day), causes the 6 hour variation in RMSE (thin dotted lines) seen in figure 8.4(b). A given

---

1 The equation governing the state of the atmosphere are differential equations describing changes within the atmosphere. To use these equations in NWP forecasts, an initial state of the atmosphere must be specified. Typically, this initial state is based on observations, which compared with the NWP grid are sparse in coverage and must be interpolated to the forecast grid. This introduces errors in the initial state of the NWP forecast and initially the NWP model might predict large changes of the atmospheric state. This is referred to as spin-up effects and these are illustrated by the strongly declining bias in figure 8.4(a). As the effects of this initial state disappear, the changes of the atmospheric state reach a certain level. Information from [http://www.drjack.info/IBPU/model-basics.html](http://www.drjack.info/IBPU/model-basics.html)
Figure 8.4: Bias (a) and RMSE (b) of global radiation calculated as an average per lead times; a Loess smoothing has been applied to emphasize trends.
Table 8.1: An average over all $N$ time space combinations of ME and RMSE in

<table>
<thead>
<tr>
<th></th>
<th>ME (bias)</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ensemble mean</td>
<td>−5.7</td>
<td>92</td>
</tr>
<tr>
<td>Control forecast</td>
<td>−2.9</td>
<td>110</td>
</tr>
<tr>
<td>Ensemble median</td>
<td>−4.2</td>
<td>97</td>
</tr>
</tbody>
</table>

Figure 8.5: Mean of the squared forecast deviation as a function of time of day (UTC) for the ensemble mean; the squared forecast deviation enters into the calculation of both the MSE and the RMSE.

lead time (of say 3 hours) will with four forecasts per day (00, 06, 12, and 18 UTC) result in four different times of the day (03, 09, 15, and 21 UTC). The complete data set for a given lead time consists exclusively of those four times of the day. How these four distinct times are distributed in figure 8.5 affects the value of the MSE (and of the RMSE) according to the generation of them (equation (8.3) and (8.4)). With the 6 hour cycle in the forecasts, two lead times with a 6 hour interval result in the same four hours of the day (that is, a lead time of 9 hours result in 09, 15, 21, and 03 UTC), which are equal to the times of the 3 hour lead time. On average and according to figure 8.5 the same squared deviations enters the calculation of the RMSE, and this induces the 6 hour variation in the RMSE seen in figure 8.4(b).

In this work, the accuracy measure used in calculating the skill score is the MSE defined in equation (8.3). The sample climatology\footnote{The climatology is estimated by the mean of the observations (or a set of observations), which is called the sample climatology (Wilks, 2006, Chap. 7).}, or sample
mean, defined as
\[
\bar{x} = \frac{1}{n} \sum_{n=1}^{n \in \mathbb{N}} x_n
\]
is used as the references forecast and \( n \) is the number in the \( \mathbb{N} \) time-space combinations matching a given lead time. This gives the following accuracy measure for the unskillful reference forecasts
\[
\text{MSE}_{\text{ref}} = \text{MSE}_{\text{clima}} = \frac{1}{n} \sum_{n=1}^{n \in \mathbb{N}} (\bar{x} - x_n)^2
\]
\(^{\text{(Wilks, 2006, Chap. 7)}}\), and the following skill score
\[
\text{SS} = \left( 1 - \frac{\text{MSE}}{\text{MSE}_{\text{clima}}} \right) \cdot 100 \%
\]
according to equation \(^{\text{(8.6)}}\).

For \( \text{MSE}_{\text{clima}} \), it is implied that the climatology, estimated by \( \bar{x} \), does not change in the course of the \( n \) forecast occasions (i.e. during either space or time of the \( n \) forecasts) \(^{\text{(Wilks, 2006, Chap. 7)}}\). Failing to account for a time varying climatology can according to equation\(^{\text{(8.5)}}\) produce an unrealistically large \( \text{MSE}_{\text{clima}} \) \(^{\text{(Wilks, 2006, Chap. 7)}}\). During the verification period of several months, the true climatology (or level) of global radiation is expected to change as illustrated in figure \(8.6(a)\). To reduce the effects of a changing climatology in the course of the verification period (August 5 to November 14), the clearness index is used in calculating skill scores. Despite a changing clearness index throughout this period, a clear and unambiguous trend is not seen in figure \(8.6(b)\). The clearness index seems to more fluctuate around a mean value compared to the global radiation in figure \(8.6(a)\).

Figure \(8.7\) displays the skill score as a function of lead time. With increasing lead time, the skill score is seen to decrease. Table \(8.2\) shows a selection of values found in figure \(8.7\). It is seen that the ensemble mean performs better in skill score (based on \( \text{MSE} \) with climatological values as the reference forecasts) compared to the control forecast. A value of 0.461 in skill score is interpreted as a 46.1 % increase in performance of the ensemble mean compared to the unskillful climatology, \( \bar{x} \). For the control forecast, this value is 27 %. The decreasing trend in skill score with lead time is highly expected. It is more surprising that the 48 hour horizon skill score for the ensemble mean (0.380) corresponds to the 6 hour horizon for the control forecast (0.382). The modest skill scores — especially pronounced

\(^{3}\text{The annual period in the global radiation caused by astronomical constellations gives, in the course of the verification period, a gradual change in the climatology of the global radiation.}\)
Figure 8.6: Time evolution in the observed global radiation [a] and from this the computed clearness index [b] for verification station 6135; a moving average has been applied (red curve) to underline the variability of the climatology.

Figure 8.7: Skill score as a function of lead time; the MSEs, on which the skill score is based, is averaged over a data set defined by the lead time (see e.g. section 8.2.2). A Loess smoothing has been applied to emphasize trends.
for the control forecast — might be related to the use of the clearness index for which the highly predictable astronomical signal present in the global radiation is eliminated leaving a “pure” weather signal.

### 8.2.3 Reliability

Recall from section 7.1.2 that a set of forecasts is said to be completely reliable if $E(x|y = Y) = Y$ for all forecast values $Y$, where $E(x|y = Y)$ is the mean value of the observations associated with a particular forecast value $Y$ (Murphy et al., 1989). The conditional distribution $p(x|y = Y)$ of observations for a given forecast value $Y$ together with the marginal distribution of forecasts $p(y = Y)$ for selected or binned values of $Y$ provide information about unconditional as well as conditional biases in the forecasts — information that becomes available in the graphical representations in figure 8.8.

Within the distribution of observations associated with (or conditional on) a forecast value $Y$, $p(x|y = Y)$, displayed in figure 8.8, the thick line shows the median of this distribution, the two dashed lines the 25th and 75th quantiles (or percentiles), and the two dotted lines the 10th and the 90th quantiles. The blue histograms displayed in figure 8.8(b) illustrates the distribution of forecasts $p(y = Y)$ also found in figure 8.2(b). The diagonal line in both figure 8.8(a) and 8.8(b) is the line of perfect correspondence between forecasts and observations (Murphy et al., 1989), that is, of completely reliable forecasts, is included to compare with the median values of the observations. Deviations of the conditional medians from this diagonal line are assumed to indicate that the forecasts contain bias (Murphy et al., 1989) either conditionally or unconditionally, and, hence, that they are not completely reliable.

Inspecting the conditional quantile plots in figure 8.8 reveals both a presence of a condition bias, that is, a forecast bias that depends on the
Figure 8.8: The conditional distribution $p(x|y)$ of the control forecast for global radiation (a) and the clearness index (b) calculated for all $N$ forecast occasions at each of the $K$ verifying locations. The “conditional quantile plots” of quantiles of the conditional distribution $p(x|y)$ are inspired by Murphy et al. (1989).
forecast value and an absence of unconditional biases, that is a systematic bias irrespective of the forecast value that is apparent in figure 8.4(a). An unconditional bias in the forecasts would in figure 8.8 appear as a line of the observation medians entirely to one side of the diagonal line — above the line in case of systematic under-forecasting and below the line in case of systematic over-forecasting. For the control forecast in figure 8.8(a), the median lies along the diagonal line of perfect correspondence a least below a forecast value of about 500 $\text{W m}^{-2}$. A value of global radiation below $\approx 150 \text{W m}^{-2}$ is associated with some degree of under-forecasting and a value above $\approx 150 \text{W m}^{-2}$ (most pronounced for forecast values from about 500 to about 800 $\text{W m}^{-2}$) with over-forecasting, which agrees with HIRLAM’s tendency to underestimate the cloud cover (Bent Hansen Sass, personal communication) and thereby over-estimating the amount of global radiation reaching the surface. From this plot, an area of focus could be to reduce high values (in the interval from 500 to 800 $\text{W m}^{-2}$) of global radiation caused by the underestimate of clouds — a problem that is not easily alleviated in NWP models because of the difficulties in forecasting the exact timing and location of small-scale clouds, some of which appear on a scale too small to be resolved by the grid of the NWP model.

A change of variable to the clearness index reveals a strong conditional bias with under-forecasting for small values of the clearness index and a similar degree of over-forecasting for high values of the clearness index. A value of $\approx 0.4$ — the central value in the distribution of forecasts of clearness index — separates under-forecasting (for values below $\approx 0.4$) from over-forecasting (for values above $\approx 0.4$). In a situation of light cloud cover corresponding to a high clearness index, the forecasts over-estimate the clearness index in correspondence with HIRLAM’s tendency to under-estimate clouds. Conversely, in a situation of heavy cloud cover corresponding to a low clearness index, the forecasts seem to over-estimate the cloud cover and hence under-estimate the clearness index (observed values are higher than forecasted values). The focus area from this figure would probably be more evenly distributed over both low values and high values of the clearness index and, hence, more evenly distributed between situations of a low to modest cloud cover associated with over-forecasting of the clearness index and thereby under-forecasting of the cloud cover and situations of a modest to an extended cloud cover associated with under-forecasting of the clearness index and thereby over-forecasting of the cloud cover.

While the conditional biases in figure 8.8(a) are modest at least until a forecast value of about 500 $\text{W m}^{-2}$ — indicating relatively unbiased forecasts in the range $0 - 500 \text{W m}^{-2}$ —, the picture is different for the clearness index in figure 8.8(b). Here, the slope of the median values is different from the line of perfect correspondence between forecasts and observations. This deteriorate appearance is believed to be caused by a combination of the rearrangement
of data using the clearness index instead of the measured values of global radiation and by the resulting change of scale.

Similar “conditional quantile” plots of the ensemble mean and the ensemble median as well as individual ensemble members correspond qualitatively to the plots for the control forecasts and are therefore not shown. However, the ensemble mean displays somewhat less conditional bias compared to the control forecast.

8.2.4 Discrimination

To investigate the ability of a set of forecasts to discriminate between observations, both the approach of Murphy et al. (1989) based on the conditional distribution $p(y|x)$ and the approach of Manson and Weigel (2009) based on the general discrimination score, $D$, (described in section 7.1.2) have been applied.

Assessed by the conditional distribution $p(y|x)$

In figure 8.9(a) and 8.9(b), the conditional distribution, $p(y|x)$, is plotted for three different values of the observation, $X$: the lower quartile (25 %), the median (50 %), and the upper quantile (75 %) of the distribution of the observations. For the data subset corresponding to the 12 UTC global radiation, these are $73 \frac{W}{m^2}$, $320 \frac{W}{m^2}$, and $589 \frac{W}{m^2}$, respectively. To reduce sampling variability, only forecasts associated with an observation within a 5 % interval of the observed value are considered. The number of forecasts used to estimate the conditional distribution $p(y|x)$ is between 350 and 1000 depending on the value of the observation and is a combination of time and space points. In figure 8.9, data for the ensemble mean (a) and for the control forecast (b) is seen. For each of the three observation values, the distributions of the corresponding forecasts have distinct peaks (modes) at values matching the observations for both the ensemble mean in figure 8.9(a) and the control forecast in figure 8.9(b). The distributions of forecasts associated with observations of $73 \frac{W}{m^2}$ and $589 \frac{W}{m^2}$, respectively, are well separated for both the ensemble mean in figure 8.9(a) and the control forecast in figure 8.9(b).

According to Murphy et al. (1989), a forecast $Y$ is perfectly discriminatory if $p(y = Y|x) = 0$ for all values of $X$ except one, that is, for perfect discrimination, all observations associated with a given forecast $Y$, should be equal to a given observation $X$. In figure 8.9(a), this is equivalent to requiring that for a given forecast, say $Y = 100$, the relative frequencies $p(y = 100|x = 320)$ and $p(y = 100|x = 589)$ both be equal to zero and $p(y = 100|x = 73)$ different from zero. This is nearly the case here for fore-
Figure 8.9: The conditional probability $p(y|x)$ for the ensemble mean [(a)] and the control forecast [(b)] for a data subset within the $N$ time-space points consistent with the 12 UTC global radiation; the figure is inspired by Murphy et al. [1989, Fig. 8].
cast in the range 0-150 $\frac{W}{m^2}$. The relative frequencies of $p(y = 100|x = 320)$ and $p(y = 100|x = 589)$ are both close to zero, while $p(y = 100|x = 73)$ is well above. However, this is not the case from values of about 300 to 500 $\frac{W}{m^2}$, where both figures reveal some overlap between the forecasts distributions corresponding to observations of 320 $\frac{W}{m^2}$ and 589 $\frac{W}{m^2}$. For those values, both forecasts distributions, $p(y|x = 320)$ and $p(y|x = 589)$, contain relative frequencies above zero and display a modest ability of both the ensemble mean and the control forecast to discriminate between those two well separated observation values.

From the two plots in figure 8.9, the assessment of which, the ensemble mean or the control forecast, is best at discriminating is not obvious. The modes of the three forecast distributions correspond better with the observations for the control forecast compared to the ensemble mean, for which the forecast distributions tend to be more peaked — at least for the two forecast distributions corresponding to $x = 73$ and $x = 320$ $\frac{W}{m^2}$.

**Assessed by the generalised discrimination score**

To avoid having trivial discriminations — for example a forecasting system being able to discriminate between global radiation in the morning and at midday or to discriminate between the global radiation in different parts of the country at a given time — influencing the evaluation of the discriminative power of global radiation, the calculation of the generalised discrimination score, $D$, is based on the clearness index, which allows the inclusion of the $K = 24$ verifying locations. The reason for this is that by using the clearness index, time, date and location are eliminated leaving only a weather component.

The generalised discrimination score, $D$, is in essence a scalar verification score, but by using the clearness index, the $N$ forecast-observation pairs at each of the $K$ verifying locations are included in the calculation of $D$.

Eliminating night values, leaves for each lead time between $\approx 3500$ and $\approx 4500$ data points (including the $K$ verifying locations) for which any forecast-observation pair must be compared with any other forecast-observation pair.

The generalised discrimination score is seen to decrease with lead time. For all lead times, both the ensemble mean and the ensemble median score higher in the generalised discrimination score compared to the control forecast. Both the ensemble mean and the control forecast show a strong resemblance in the value of $D$, which decreases from about 78 % to 72 % for the ensemble mean and the ensemble median and from about 77 % to 70 % for the control forecast. Compared to the control forecast, the ensemble mean (and the ensemble median) scores about 3 % higher in discrimination score, $D$.

Each lead time is associated with between 3500 and 4500 forecast-observation pairs. Of these observations, nearly just as many are distinct
corresponding to weather situations that might be very similar, but seldom identical. This results in observations that might differ by tiny amounts. This might explain the somewhat modest discrimination score seen in figure 8.10. A higher precision of the forecasts is required to discriminate between continuous observations that differ by only small amounts. Another reason for the modest discrimination score might be related to the data set consisting of the clearness index for which the highly predictable astronomical signal present in the global radiation is eliminated leaving the more stochastic weather component. This might imply a conservative estimate of the discriminative power of the forecasts.

The two different ways of assessing the discriminative power of the ensemble mean and the control forecast have resulted in different outcomes. Using the generalised discrimination score, the ensemble mean performs better compared to the control forecast. This is, however, not evident using the conditional distribution $p(y|x)$ to assess the discriminative power.
8.3 Assessing the Performance of DMI-EPS

A central aspect of ensemble forecasting is its capacity to give information about the uncertainty in a forecast (Wilks, 2006, Chap. 6). The second part of this chapter is devoted to assessing the quality of DMI-EPS’ ensemble forecasts of global radiation.

8.3.1 Application of quality measures applicable to single forecasts

Treating each ensemble member as an individual deterministic forecast, the deviation \( y_i - x_i \) for each forecast-observation pair \( i = 1, \ldots, N \), with \( N = 19,002 \), is computed at each of the \( K = 24 \) verifying locations and this is generated for each of the \( M = 25 \) ensemble forecasts. The resulting distribution of this pooled data set is divided after the hour of the day (UTC), and the result can be seen in figure 8.11. For each hour, the forecast deviations are generally distributed around zero with no forecast deviation at night — when both observations and model calculations are zero. Despite large deviations between forecasts and observations, no deviations seem physically impossible. A strong symmetry is present in the distributions in figure 8.11. Around 12 UTC, the distributions broaden and at 12 UTC, the distribution contains the largest absolute forecast deviations. A prevailing small negative forecast deviation is seen indicating a modest degree of under-forecasting. However, as suggested by figure 8.8, this picture might be more complicated with a presence of conditional bias.

Two quality measures of deterministic forecasts are applied to all 25 ensemble members. In figure 8.12, the forecast accuracy measure, the RMSE (see e.g. section 8.2.2), is seen, and in figure 8.13, the measure of forecast skill, the skill score is seen. The ensemble members are grouped according to the applied cloud scheme (STRACO and KF/RK), to “stochastic physics” and “no stochastic physics” of applied stochastic physics and no applied stochastic physics, respectively, and to their initial conditions (IC1 - IC5) (see e.g. table 6.2).

In general, the RMSE in figure 8.12 increases with lead time (this was also the case for the RMSEs of the control forecast, of the ensemble mean, and of the ensemble median in figure 8.4(b)). In using no stochastic physics (lower panels in figure 8.12), the two cloud schemes have different impacts on the RMSE with a high degree of clustering and little mixing and with the STRACO scheme in general superior to KF/RK. Without the application of stochastic physics, the RMSE for the two cloud scheme are slightly more adjacent for IC2 and IC4 compared to IC1, IC3, and IC5. In applying stochastic physics (upper panels), a considerable degree of overlap between the two cloud schemes is seen. In this case, however, the KF/RK scheme performs somewhat better with lower RMSEs. Differences in RMSE between
Figure 8.11: The distribution of the forecast deviations — difference between a forecast and the corresponding observation, \( y_i - x_i \), for each \( i \) forecast occasion with \( i = 1, \ldots, N \). The deviations are calculated for each \( j = 1, \ldots, M \) ensemble forecast and for each location of the verifying observations \( k = 1, \ldots, K \); the forecast deviations have been split by hour of day (UTC) indicated in the strips (0-23); night values are included; in each panel, the mean of the corresponding distribution is indicated; in each panel, the distribution contains \( 19002 \times 24 \) (simultaneously predicted locations) \( \times 25 \) (ensemble forecasts) \( \times 24 \) (hours) = 475,050 values.
Figure 8.12: RMSE as a function of lead time calculated for each of the $M = 25$ ensemble members; the data set consists of all $N$ forecast-observation pairs at each of the $K = 24$ verifying locations matching a given lead time. The RMSE for the ensemble members are grouped according to applied cloud scheme (the “STRACO” cloud scheme and the “KF/RK” cloud scheme), no application of stochastic physics (“no stochastic physics”) and application of stochastic physics (“stochastic physics”), and initial conditions (see e.g. table 6.2); note that in the upper panel, two ensemble members fulfill the requirements of application of stochastic physics and STRACO cloud scheme for each initial condition, and this results in twice as many points for the STRACO scheme (purple “+”) compared to the KF/RK scheme (cyan “×”) — for IC1 ensemble members 6 and 21 contribute and for IC2 ensemble members 7 and 22 contribute, etc. (consult e.g. table 6.2).
the initial conditions are not pronounced for either cloud scheme. From the above analysis of RMSE, the application of stochastic physics seems to be able to convert groups of ensemble members into a more unified distribution, in which one cannot tell from the group of ensembles (1 to V — see e.g. table 6.2), the value of the RMSE (for a given value of the lead time). Without stochastic physics, the different cloud schemes result in some clustering of the RMSEs with distinct relationships between RMSE and lead time.

Figure 8.13 shows a decreasing trend in the skill score, SS, with lead time (which was also the case for the skill scores for the control forecast and the ensemble mean in figure 8.7). Here, only the part of the data set corresponding to 12 UTC is included, which gives the following lead times: 0, 6, 18, 24, 30, 36, 42, 48, 54. In general, the conclusions of the figure resembles the conclusions of figure 8.12, but the limited data set emphasizes features of the ensemble forecasts somewhat more concealed in figure 8.12. A clear distinction in the skill scores of the two cloud schemes between applying stochastic physics (upper panel) and not applying stochastic (lower panel) exists. The application of stochastic physics brings the two cloud schemes closer regarding the skill score as defined here. With no application of stochastic physics, there is little mixing in skill scores between the two cloud schemes. The STRACO cloud scheme performs better (with higher skill score) than the KF/RK cloud scheme, but the closeness in skill score between the two cloud schemes depends, to some degree, on the initial condition perturbations. At the 54 hour lead time, the difference in skill score between the two cloud schemes is, for the initial conditions IC1, IC3, and IC5, substantial with no application of stochastic physics.

Considering the ensemble as reliable and the ensemble members as a random sample from the PDF, both the RMSE and its associated skill score (here based on MSE with climatology as the unskillful reference forecast) will deviate from one ensemble member to the next, but supposedly more in a statistical sense — in line with the upper panels in figure 8.12 and 8.13 — and probably less in line with the lower panels of these figures.

Among ensemble members with no application of stochastic physics in figure 8.12 and 8.13 a clear pattern in the different initial conditions is seen in both figure 8.12 and 8.13. The quality measure (RMSE or SS) of the two cloud schemes resembles each other more for IC2 and IC4 compared with IC3 and IC5. With respect to both RMSE and SS, using a 6 hour old forecast to generate the ensemble members gives the same pattern as using a 12 hour old forecast. Rather, the similarity in performance of the two cloud schemes seems to be related more to the sign of the perturbation in generating the initial conditions (see e.g. table 6.2 and equation (6.8)).

Values of RMSE of the individual ensemble members are comparable with values reported in Díaz et al. (2012), Lorenz et al. (2011) and Lorenz et al. (2012), which lends support to the individual forecasts of DMI-EPS.
Figure 8.13: Skill scores (based on $\text{MSE}_{\text{clima}}$) as a function of lead time calculated for each of the $M = 25$ ensemble members; the data set consists of 12 UTC data at all $K$ verifying locations matching a given lead time. The skill scores are grouped according to applied cloud schemes (STRACO and KF/RK), no application of stochastic physics (“no stochastic physics”) and application of stochastic physics (“stochastic physics”), and initial conditions (see e.g. table 6.2); note that in the upper panel, two ensemble members fulfill the requirements of application of stochastic physics and STRACO cloud scheme for each initial condition, and this results in twice as many points for the STRACO scheme (purple “+”) compared to the KF/RK scheme (cyan “x”) — for IC1, ensemble members 6 and 21 contribute and for IC2, ensemble members 7 and 22 contribute, etc. (consult e.g. table 6.2).
8.3.2 Application of quality measures applicable to ensemble forecasts

For EPSs, the future state of a meteorological parameter should ideally fall within the ensemble spread, or the ensemble range or the ensemble dispersion, and the amount of spread should ideally be related to the uncertainty of the forecast (Toth and Kalnay, 1993). A small ensemble spread, that is, a high degree of agreement between the individual ensemble members, would intuitively imply a high confidence in the forecasts.

In figure 8.14(a), the capture rate is shown as a function of forecast lead time both excluding and including uncertainties on the observations. Due to sampling variability, the expected capture rate for a reliable ensemble of size \( M \) is \( \frac{M - 1}{M + 1} \) (Feddersen, 2009; Jolliffe and Stephenson, 2012) — less than 100 %. Including 10 % uncertainty on the observations, increases the capture rate, but it is about 10 % off compared to the expected capture rate of \( \frac{M - 1}{M + 1} \), which is about 92 %. Both curves (of including and not including observational uncertainties), increase with lead time until a lead time of about 18 hours. From then on, the curves increases only slightly. In figure 8.14(b), the spread of, or the dispersion of, the ensemble, here expressed as one standard deviation, is plotted against lead time. The ensemble spread increases until a lead time of about 42 hours, thereafter the ensemble spread seems to stabilise. Beyond a lead time of about 18 hours, the further divergence of the ensemble members, that is, the increase in ensemble spread (figure 8.14(b)) seems to be able to ensure that the capture rate is kept at a certain level throughout the forecast period despite the typical deterioration of forecasts with increasing lead time as illustrated in figure 8.4(b) and 8.12 for the RMSE and in figure 8.7 and 8.13 for the skill score.

Scalar Reliability

According to Jolliffe and Stephenson (2012, Chap. 8) and Bowler (2008), the RMSE of the ensemble mean should in case of consistency of the ensemble match the mean of the standard deviation of the ensemble — apart from a factor \( \sqrt{\frac{M - 1}{M}} \) that depends on the size of the ensemble. In figure 8.15, the RMSE as defined by equation (7.2) is plotted as a function of the mean ensemble standard deviation. The multidimensionality is here eliminated by averaging over the \( N \) time points in calculating the RMSE and the mean ensemble standard deviation. This gives a pair of values for each of the 24 verifying locations.

The ensemble standard deviation is about 20 % lower than the RMSE. Complete consistency of the ensemble forecasts of global radiation is not evident from the “RMSE versus ensemble spread”-relation.
Figure 8.14: The capture rate (a) and ensemble spread (b) — expressed by one standard deviation from mean — as an average per lead time in a data set consisting of global radiation for \(N\) forecast-observation pairs; a Loess smoothing has been applied to emphasize trends.
Mean RMSE = 92
Mean ensemble standard deviation = 72

Figure 8.15: RMSE versus mean of the ensemble standard deviation multiplied by $\sqrt{\frac{M+1}{M}}$ (indicated by an asterisk (*)); an average of the 24 values is indicated.

Multidimensional Reliability

**Multivariate Rank Histogram** To evaluate the multidimensional reliability of the $K = 24$ simultaneously ensemble predictions each consisting of $N$ ensemble forecasts with matching observations (see e.g. section 8.1), a multivariate rank (MVR) histogram has been constructed according to Gneiting et al. (2008); Jolliffe and Stephenson (2012) and is displayed in figure 8.16.

From the MVR histogram, both a slight tendency of the ensemble forecasts to over-forecast and a number of events, where the observation is exceeded by all ensemble members is seen. The highly overpopulated rank, $r = 1$, together with the slightly negative slope in the distribution suggests a degree of over-forecasting of the ensemble (see e.g. Gneiting et al. (2008, Figure 3)), which corresponds with HIRLAM’s tendency to underestimate the cloud extent (Bent Hansen Sass, personal communication).

In the MVR histogram (figure 8.16), all rank values — except rank 1 — are close to being equally populated. One could argue that the requirement that for any two ensemble “vectors” $V_j$ and $V_l$, $V_j$ must exceed $V_l$ for all $K$ dimensions for the coefficient $q_{j,l}$ to be different from zero (and for the pre-rank, $\rho_j$, to be different from 1), is in this case difficult to fulfil for $K = 24$ dimensions. This would imply that most “pre-ranks”, $\rho_j$ with $j \in \{0, 1, \ldots, M\}$ will be equal to 1 according to equation (7.11). The “pre-
The “rank” of the observation, $\rho_0$, will then often be tied and its multivariate rank, $r$, solved at random. This will tend to populate all rank values, $1, \ldots, 26$, equally, which will result in a uniform multivariate rank histogram.

**Minimum Spanning Tree Histogram**  The other method to evaluate the reliability of multidimensional ensemble forecasts is the minimum spanning tree (MST) histogram (Jolliffe and Stephenson, 2012, Chap. 8). It is applicable if a number of locations are simultaneously predicted (Jolliffe and Stephenson, 2012), which is the situation in this work with $K = 24$ simultaneously predicted ensemble forecasts (see e.g. section 8.1).

The distances of the spanning tree can be calculated in several ways applicable to different characteristics of the data set. In this case, the minimum spanning trees have been calculated using a Euclidean distance in a $K = 24$ dimensional space. The Euclidean distance might not be appropriate if the variances in the data set are different for some of the $K$ dimensions, and if further a covariance between (some of) the $K$ dimensions exists (Gombos and Hansen, 2007). The $K$ dimensions of the data consist of the 24 locations measuring the same quantity, and it is therefore assumed that the variances of the data in different dimensions on average are similar. The use of the clearness index should reduce some of the covariance between dimensions.

---

4The situation is different if the $K$ dimensions consists of $K = 2$ meteorological variables — like temperature and precipitation at the same location, since temperature and precipitation might not have the same variance.

5Concerning global radiation, a covariance between the $K$ verifying locations exists due to the latitude and longitude dependence of the solar radiation. By use of the clearness index, this is eliminated, but a covariance due to a similar weather situation at two or
With the large number of forecast realizations, \( N \), available here, it is believed that the similarity of the two distributions — the distribution at a certain forecast time of the ensemble forecasts and of the observation (considered as a random sample) — will not be sensitive to the choice of norm \cite{gombos2007}. From those considerations, the Euclidean distance has been applied in the minimum spanning tree.  

The strong negative slope (see e.g. section 7.2) suggests an under-dispersive ensemble. Of the rank values \( \{1, \ldots, M+1\} \), the smallest rank value, that is, \( r = 1 \), is highly overpopulated as seen in figure 8.17 indicating that the inclusion of the verifying observation increases the total length of the MST and hence that the notion of the ensemble members and the verifying observation being statistically indistinguishable is not evident.

**Discrimination**

To avoid having trivial discriminations — for example a forecasting system being able to discriminate between the 12 UTC global radiation on August 23 and on November 12 — influencing the evaluation of the discriminative power of global radiation from DMI-EPS, the calculation of the generalised discrimination score, \( D \), is based on the clearness index, which allows the inclusion of the \( K = 24 \) verifying locations, since the clearness is a normalised quantity sensitive to only the weather situation — for example the amount of and distribution of clouds.

The generalised discrimination score, \( D \), is a scalar verification score, but by using the clearness index, the \( N \) forecast-observation pairs at each

\[ \text{more neighbouring locations will still be present.} \]
of the $K$ verifying locations are included in the calculation of $D$.

Eliminating night values, leaves for each lead time between $\approx 3500$ and $\approx 4500$ data points (including $K$ verifying locations) from which any forecast-observation pair must be compared with any other forecast-observation pair. With $\approx 4000$ values for each lead time, the process of ranking ensemble forecasts (described in section [C]) is quite time-consuming. To reduce computational time, a sample of 1000 randomly selected data points is generated for each lead time, and from those values, the generalised discrimination score is calculated.

As the number of distinct observations increases, the discriminative power of forecasts usually decreases because of the higher precision required in the continuous case to discriminate between a large number of observations that may differ by only small amounts (Weigel and Manson, 2011). To compare with ensemble forecasts of continuous observations, binary observations are generated by discerning between observations above or below the observation mean.

In figure 8.18, the generalised discrimination score is seen to decrease with lead time for both the continuous and the binary observations as expected. For the continuous observations the generalised discrimination score decreases from about 78% and ends at about 72% at a 54 hour forecast lead time. In the case of binary observations, the generalised discrimination score...
tion score is about 8% high throughout the forecast period. It decreases from about 86% to about 85% in the first 24 hours of the forecast. From then on, a faster decrease is seen, and at a 54 hour forecast lead time, the generalised discrimination score has dropped to about 80%, which is highly comparable with $D$ calculated for the ensemble mean and the ensemble median in figure 8.10. In this case, the sample of 1000 values corresponds to 1000 (or just below) discrete observations, so here, the forecasts have to discriminate between observations that may differ by only tiny amounts, which might explain the generally modest discrimination score, $D$. From figure 8.18 the generalised discrimination score for binary observations is for all lead times above the score for the continuous observations as a result of the higher precision required for the forecasts to discriminate between observations that differ by only small amounts in the continuous case compared to the binary case with only two observation values. Another reason for the modest discrimination score for both the continuous and the binary observations might be related to the use of the clearness index for which the highly predictable astronomical signal present in the global radiation is eliminated leaving only a weather component. This might give a modest estimate of the discriminative power of the forecasts.
9 Discussion

9.1 Verification of Global Radiation from DMI-EPS

An important goal in ensemble prediction is to approximate by the ensemble members the forecast PDF, which contains information on the forecast uncertainty. Through verification techniques, the consistency, or reliability, of the ensemble forecasts of global radiation has been investigated.

9.1.1 Assessing the Performance of DMI-EPS

Generating an EPS that takes into account both the uncertainty in the analysis, which is unknown, and the imperfections of the NWP model, which might be difficult to assess, poses a challenge to ensemble prediction.

Consistency of DMI-EPS forecasts of global radiation

The immediate conclusion on the ensemble consistency, or the ensemble reliability, of DMI-EPS global radiation forecasts is somewhat ambiguous. However, the strong negative slope in the MST histogram in figure S.17 and the values of the “RMSE versus ensemble spread”-relation shown in figure S.15 both suggest an under-dispersive ensemble as caused by a conditional forecast bias. Under-dispersion is not evident from the MVR histogram in figure S.16. One could in case of the the MVR histogram argue that the requirement that for any two ensemble “vectors” $\mathbf{V}_j$ and $\mathbf{V}_l$, $\mathbf{V}_j$ must exceed $\mathbf{V}_l$ for all $K$ dimensions for the coefficient $q_{j,l}$ to be different from zero (and for the pre-rank, $\rho_j$, to be different from 1 — see e.g. equation (7.11)), is in this case difficult to fulfil for $K = 24$ dimensions. Most “pre-ranks”, $\rho_j$ would then be equal to 1 according to equation (7.11). The “pre-rank” of the observation, $\rho_0$, will then often be tied and its multivariate rank, $r$, solved at random. This will tend to populate all rank values, 1, $\ldots$, 26, equally, which will result in a uniform multivariate rank histogram.

The under-dispersiveness derived from the MST histogram and the “RMSE versus ensemble spread”-relation corresponds with the too low capture rate seen in figure S.14(a). In DMI-EPS, the capture rate is about 20 % off the
expected rate excluding uncertainties on the observations. This number is reduced to about 10% including observational uncertainty. This illustrates the importance of including uncertainties on observations in all kinds of verification. These capture rates correspond to figures reported in Bowler (2008) for the MOGREPS (Bowler et al., 2008) limited area ensemble prediction system of 24 ensemble members. Under-dispersion is ensemble forecasts is a common problem in ensemble prediction (Bowler, 2008; Buizza, 1997; Wilks, 2006), and how to alleviate it is not straightforward as it necessitates a better simulation of the uncertainty in the analysis and/or of the uncertainties relating to deficiencies in the NWP model formulation.

On those occasions when the observation lies just outside the ensemble range, the effect of excluding or including uncertainties on observations in evaluating ensemble dispersion either through rank histograms or through capture rates may be considerable (Bowler, 2008). Except in a few cases, observational uncertainties have not been addressed, and this might permeate most results. Whether the inclusion of uncertainties on observations alone will increase the dispersion is to be investigated, but the results obtained here point toward an under-dispersive ensemble and ensemble forecasts that will give too little uncertainty on the forecast values.

**Over-forecasting or under-forecasting**

The MVR histogram in figure 8.16 suggests a strong degree of over-forecasting (from the highly over-populated lowest rank value, \( r = 1 \), indicating that the observation is too often exceeded by all ensemble members), which by comparing with Gneiting et al. (2008, Figure 4), seems to be the case in the MST histogram in figure 8.17. A tendency of the ensemble to over-forecast corresponds with HIRLAM’s tendency to under-forecast clouds (Bent Hansen Sass, personal communication) and, hence, over-forecast global radiation at the surface. However, the slight under-forecasting evident from figure 8.11 and the over-forecasting established from figure 8.16 and 8.17 might seem contradictory. This difference in over-all bias is related to the difference between real values — as used in figure 8.11 and rank values — as used in figure 8.16 and 8.17 — for which information on the distance of the ensemble members to the observation is lost. To illustrate this, consider table 9.1. At time point “a”, the contribution to the deviation between the ensemble members and the observation amounts to \( 2 - 18 + 1 = -15 \), while the observation only exceeds one ensemble member resulting in rank 2 of the observation. At time point “b”, the contribution to the deviation between ensemble members and observation is \( 1 + 1 + 1 = 3 \) with all ensemble members exceeding the observation, which then has rank 1. The difference between rank 1 and rank 2 of the observation, might therefore contribute significantly to the difference between the ensemble forecasts and the observation.
Table 9.1: An example illustrating how a low rank value of the observation can be associated with a large negative difference between a forecast and the matching observation; in parentheses, the rank value is shown.

On those occasions, when all ensemble members exceed the observation, the contribution to the forecast deviation in figure 8.11 will be positive, but on those occasions when the observation exceed one or more of the ensemble members, the contribution to the forecast deviation in figure 8.11 might be considerable negative as illustrated in table 9.1. It is therefore possible to establish over-forecasting from the rank histograms in figure 8.16 and 8.17 and under-forecasting looking at the deviation between a given ensemble member and the observation.

As opposed to the rank histograms, the distributions shown in figure 8.11 are more sensitive to the actual values of the ensemble members. However, dispersion of the ensemble members is a necessary condition for ensemble consistency, and if the ensemble members were to be sampled from the PDF, all values would be valid, but, of course, not equally likely. From these considerations, I have most faith in the results from the MVR histogram with the reservations that all rank histograms should be interpreted with care and that despite searching the literature, a similar shape of a MVR histogram has not been found, which further complicates interpretation.

Initial Condition Perturbations

Several ways of generating initial condition perturbations simulating the uncertainty in the analysis exist. These methods differ in approach and complexity. In DMI-EPS, a SLAF method to generate ensemble perturbations is applied. Despite Boeing a simple way of generating the initial condition perturbations, results endorsing the SLAF method have been reported (Feddersen, 2009). Feddersen (2009) speculates, however, that a way of increasing the ensemble spread and thereby improve DMI-EPS forecasts could be to use a different method to perturb the initial conditions — for example, a breeding method (described in Toth and Kalnay (1993)).

Among ensemble members with no application of stochastic physics, two distinct patterns in the different initial conditions are seen in both figure 8.12 and 8.13 with a common pattern between IC3 and IC5, and a common
pattern for IC2 and IC4. An explanation for this distinct pattern in terms of the generation of the initial conditions using a 6 hour and a 12 hour old forecast is not evident. These distinct patterns might suggest a degree of clustering of the ensemble members. However, by application of stochastic physics this apparent clustering seems to disappear.

Accessing model deficiency by use of multiple models

The use of multiple models beside the initial condition perturbations in generating ensemble members is reported to both cluster the ensemble members by model with a PDF dictated more by parametrisation scheme and less by the forecast uncertainty [Bowler et al., 2008] and improve the ensemble forecasts by increasing the ensemble dispersion [Wilks, 2006, Chap. 6]. From figure 8.12 and 8.13 the use of multiple models in terms of cloud schemes seems to cluster the ensemble members by model. However, application of stochastic physics seems to alleviate this clustering. Whether the use of multiple cloud schemes improves the ensemble forecasts by increasing the ensemble dispersion is not easily assessed by the results obtained here, but the RMSE in figure 8.12 and the skill score in figure 8.13 might be distributed over a larger span of values compared to corresponding values with an application of stochastic physics in both figures. This is most pronounced for IC1, IC3, and IC5.

9.1.2 Assessing the Performance of the Ensemble Mean

An early goal of ensemble prediction was to obtain a forecast superior to a traditional forecast by averaging the ensemble members. In averaging, some of the errors present in a traditional forecast should cancel and, in general, provide an improved forecast [Tracton and Kalnay, 1993]. However, this will only be the case if initially the ensemble members represent the PDF of the analysis and if the NWP model is perfect [Toth and Kalnay, 1993; Tracton and Kalnay, 1993]. Imperfect NWP models and imperfect representations of the analysis PDF have shown to limit the effect of cancelling the errors [Tracton and Kalnay, 1993], and, averaging the ensemble members is now considered to be a simple application of ensemble forecasts [Wilks, 2006, Chap. 6].

Within the task of assessing the full potential of DMI-EPS, the ensemble mean has, nevertheless, been computed to be compared with the control forecast. Being a more robust measure of centre, the ensemble median has, in addition, been computed. However, the ensemble median does not, as speculated, perform better than — but highly comparable to — the ensemble mean. Concerning RMSE (figure 8.4(b)), skill score (table 8.2), reliability (assessed by the conditional quantile plots (illustrated in figure 8.8 for the control forecast)), and discrimination as measured by the general discrimi-
nation score, $D$, (figure 8.10), the ensemble mean performs better than the control forecast. The superiority of the ensemble mean compared to the control forecast might be an effect of reducing by averaging some of the errors present in the control forecast. From the conditional quantile plots in figure 8.8, the character of the forecast bias has been revealed.

A source of the modest skill score, SS, of the control forecast might be related to the use of the clearness index in calculating the score. By use of the clearness index, the highly predictable signals relating to astronomical effects present in the global radiation is eliminated leaving the more stochastic weather component, which might give a conservative estimate of the forecast skills score. Another source of the modest skill score might be related to the limited time (5 August 2011 - 14 November 2011) over which the sample climatology (or observation mean) used as the unskillful reference forecast in computing the skill score is estimated (Jolliffe and Stephenson, 2012, Chap. 1).

**Discrimination**

The somewhat modest generalised discrimination score evident from figure 8.10 might be related to i) the vast amount of distinct observations for the forecasts to discriminate between and to ii) the use of the clearness index in estimating the discriminative power by the generalised discrimination score. By use of the clearness index, the highly predictable astronomical signal is eliminated leaving only a weather component. This might imply a highly cautious estimate of the discriminative power of the forecasts.

In general, the discrimination score, $D$, is about 3 % higher for the ensemble mean compared to the control forecast. This number might not be exceptional, but I find it nevertheless surprising. I would have expected that the smoothing by averaging introduced by the ensemble mean would imply a loss of discriminative power compared to the control forecasts. However, a clear distinction in the discriminative power between the ensemble mean and the control forecast using the conditional distribution $p(y|x)$ in figure 8.9 is not seen. The difference between the two methods in evaluating the discriminative power of the ensemble mean and the control forecast might be explained by the different approaches based on each their distribution: the conditional distribution $p(y|x)$ in figure 8.9 and the joint distribution $p(x, y)$ in figure 8.10.

9.2 **Performance of Global Radiation Forecasts**

Evaluation of single forecasts of global radiation has led to a revision of the HIRLAM shortwave parametrisations, which consists of simple, yet successful, parametrisation of the transmission of solar radiation through the
atmosphere. Concerning direct radiation, the revised parametrisations performs most satisfying compared with a highly detailed model describing transfer of radiation within the atmosphere.

9.3 High-Resolution, Limited-Area Models

With increased resolution in limited area NWP models, more details on the atmospheric state can be resolved including small-scale clouds, which might not appear in a model of lower resolution. If, however, one of these small-scale clouds were to be forecasted at one grid point, but show up at a neighbouring grid point, both grid points are forecasted wrong in terms of cloud cover. In this way, high-resolution models might because of their degree of detail suffer a double penalty compared to lower resolution models. The high-resolution limited area ensemble prediction system at DMI may, as a high-resolution model, in some cases suffer from this double penalty, which might reduce the quality of forecasts of global radiation.

9.4 The Solar Heating Unit

The full implementation of the advanced control system making use of all state-of-the-art forecast models developed within the project “Solar/electric heating systems in the future energy system” has not been feasible. It is planned, in a highly desirable prolongation of the project, to extend the advanced control system to make use of all developed forecast methods and available information.
10 Conclusion & Outlook

DMI-EPS

For most considered aspects of forecast quality, DMI-EPS is documented to perform better than or comparable to the global ensemble prediction system at ECMWF (Feddersen, 2009) and is a highly valuable forecast tool at the weather service at DMI.

From the investigation of the consistency of the ensemble forecasts of global radiation, there seems to be evidence of some under-dispersiveness. Whether the inclusion of uncertainties on the observations will reduce this under-dispersiveness is still an open question, but the ensemble members will tend to express too little uncertainty on forecast values (Wilks, 2006, Chap. 6). Often in EPSs, the ensemble dispersion will be too small to capture the observation at the expectable fraction (Bowler, 2008; Buizza, 1997; Wilks, 2006). How to alleviate this under-dispersiveness is not straightforward as it requires better simulation of the uncertainty in the analysis and, maybe, better simulation of the uncertainties relating to deficiencies in the NWP model formulation.

Concerning global radiation forecasts, the effect of simulating model deficiencies by use of multiple cloud schemes seems modest, and in a future revision of DMI-EPS, simulating model deficiencies by the application of two cloud schemes might be reconsidered and compared to the effect on ensemble dispersion of using a more complex way of generating the initial condition perturbations.

In DMI-EPS, the tendencies of temperature, wind speed, humidity, and cloud condensate are subjected to stochastic perturbations, which is a combination of meteorological variables of the primitive equations and variables of high impact and interest to common users of meteorological forecasts. A vast number of meteorological variables and parameters exist, each of which can be perturb stochastically. Bowler et al. (2008) apply stochastic perturbation to adjustable parameters in the physical parametrisations necessary in NWP models and introducing uncertainties in the NWP model. However, the stochastic perturbations applied in DMI-EPS seem highly successful as they are able to make ensemble members clustered by either model or initial condition un-clustered. Alternative, or additional, meteorological variables...
to subject to stochastic perturbations could still be considered in a revision of DMI-EPS. Inspired by (Bowler et al., 2008), these could be poorly know parameters in some of the parametrisations. Which parametrisations to select for stochastic perturbations should be deliberated.

Application of statistical post-processing might be a way to improve the ensemble forecasts of global radiation. However, more research into the character of the bias and its variation with weather patterns is necessary.

DMI-EPS — The Ensemble Mean

From the superiority of the ensemble mean compared to the control forecast, the value of the operational forecast might, maybe also for other parameters, benefit from complementing it with the value of the ensemble mean. A cluster analysis may, in addition to providing valuable information on the distribution of the ensemble members, also provide information on the appropriateness of the ensemble mean as characterising the ensemble.

Revision of the Radiation Scheme

In consequence of the evaluation of global radiation from DMI’s operational HIRLAM model, revisions of the radiation scheme have been implemented. The implementation has resulted in a radiation scheme that concerning direct radiation performs most satisfying compared to a detailed model describing the transmission of radiation through the atmosphere.
Appendices
A Derivation of Beer’s Law

When radiation enters a medium containing optically active matter, interactions between the radiation and the matter will weaken or attenuate the radiation traversing the medium by either absorbing or scatter the radiation. This attenuation or extinction depends on the optical characteristic of the medium (Thomas and Stamnes, 1999, Chap. 2) and is described by Beer’s law the derivation of which is found below.

In figure A.1(a) a small slap of volume $dV$, thickness $ds$ and surface area $dA$. The slap contains matter of $dN$ particles that are all optically active, that is, they interact with radiation passing through $dV$. A beam of radiation of a given wavelength $\lambda$ is incident normally to the slap — as illustrated to the left in figure A.1(a). As the beam of radiation passes through the slap, it interacts with the particles through either absorption or scattering and less radiation escapes at the opposite site of the slap (Thomas and Stamnes, 1999, Chap. 2) — the right side in figure A.1(a). The beam of radiation has been subjected to attenuation (or to extinction).

It has been found experimentally that the weakening or attenuation of a radiation beam depends linearly upon both the incident radiation and the amount of optically active matter along the beam direction (Thomas and Stamnes, 1999, Chap. 2). This amount must be proportional to the length $ds$. That is, the amount of optically active matter along the beam direction is proportional to the distance within the medium traversed by the beam, which in figure A.1(a) is $ds$. This gives

$$dI_\lambda \propto -I_\lambda ds$$

with the change in radiation to the left of the proportionality sign and the dependent quantities to the right. The constant of proportionality is denoted by $k_\lambda$ and called the extinction coefficient. It is a characteristic of the matter and might vary with wavelength. The above equation then becomes

$$dI_\lambda = -k_\lambda I_\lambda ds, \quad (A.1)$$

which is known as Beer’s law. By integrating equation (A.1) over a finite path from $0$ to $S$ along the beam direction (figure A.1(b)), the more common
Figure A.1: (a) Radiation passing through a thin slab; (b) Radiation passing through a medium. The figure is inspired by [Thomas and Stamnes (1999, Figure 2.4)]

The form of Beer’s law is obtained

\[ I_\lambda(S, \hat{\Omega}) = I_\lambda(0, \hat{\Omega}) \cdot \exp\left(-\int_0^S k_\lambda \, ds\right). \]  

(A.2)

This form gives the radiation leaving the medium, \( I_\lambda(S, \hat{\Omega}) \), as a function of the radiation entering the medium \( I_\lambda(0, \hat{\Omega}) \), where \( \hat{\Omega} \) is direction of propagation of the beam — which can be determined by a zenith and an azimuth angle. In an inhomogeneous medium, the extinction coefficient, \( k_\lambda \), might change along the beam direction.

Defining an optical path \( \tau_\lambda \) results in the following expression of Beer’s law

\[ I_\lambda(S, \hat{\Omega}) = I_\lambda(0, \hat{\Omega}) \cdot \exp(-\tau_\lambda). \]  

(Paltridge and Platt, 1976; Thomas and Stamnes, 1999), where \( \tau_\lambda \) is a measure of the strength and number of optically active particles (that scatter or absorb radiation) along the beam of radiation, and the radiation is seen to decay exponentially with \( \tau_\lambda \) along the beam direction (Thomas and Stamnes, 1999, Chap. 2). Note that \( \tau_\lambda \) may change along the path of the beam.

\( I_\lambda \) is the spectral radiance, which is the energy with wavelengths in the interval \( \{\lambda, \lambda + d\lambda\} \) that in the time interval \( dt \) passes the area element \( dA \)
and flows into a solid angle $d\omega$ centered on the direction $\hat{\Omega}$

$$I_\lambda = \frac{d^4E}{dA \cdot \cos \theta \cdot dt \cdot d\lambda \cdot d\omega} \quad (A.3)$$

where $\theta$ is the angle between the normal to the area $dA$ and $\hat{\Omega}$. The quantity $dA \cdot \cos \theta$ is the projection of $dA$ onto the direction $\hat{\Omega}$ (Thomas and Stamnes, 1999, Chap. 2) (see e.g. figure A.2).

The spectral irradiance, $F_\lambda$, expresses the energy with wavelengths in the interval $\{\lambda, \lambda + d\lambda\}$ that crosses an area $dA$ in the time interval $\{t, t + dt\}$

$$F_\lambda = \frac{d^3E}{dA \cdot dt \cdot d\lambda} \quad (A.4)$$

(Thomas and Stamnes, 1999, Chap. 2). Combining equation (A.4) and (A.3), the following relation between spectral radiance and spectral irradiance is obtained

$$F_\lambda = \int_{4\pi} I_\lambda \cos \theta d\omega. \quad (A.5)$$

From integrating the spectral radiance, $I_\lambda$, times $\cos \theta$ over all $4\pi$ solid angles, the total (or net) spectral irradiance $F_\lambda$ is obtained.

The total downward irradiance (that is, the global radiation), $F^\downarrow$, is obtained by integrating the spectral irradiance, $F_\lambda$, over a hemisphere.

$$F^\downarrow = \int_0^\infty \int_{2\pi} I_\lambda \cos \theta d\omega d\lambda \quad (A.6)$$

The inner integral ($\int_{2\pi} I_\lambda \cos \theta d\omega$) is the downward spectral irradiance according to equation (A.5), which is then integrated over all wavelengths.

Integrating both the spectral radiance, $I_\lambda$, and the spectral irradiance, $F_\lambda$, over wavelengths gives the radiance $I$ and the irradiance $F$, respectively, with units of $\frac{W}{m^2 \cdot sr}$ and $\frac{W}{m^2}$, respectively (Thomas and Stamnes, 1999, Wallace and Hobbs, 2006). Equation (A.5) then becomes

$$F = \int_0^\infty F_\lambda d\lambda \quad (A.7)$$

$$= \int_0^\infty \int_{4\pi} I_\lambda \cos \theta d\omega d\lambda \quad (A.8)$$

$$= \int_{4\pi} I \cos \theta \quad (A.9)$$

This equation illustrates the relation between radiance and irradiance with unit of $\frac{W}{m^2 \cdot sr}$ and $\frac{W}{m^2}$, respectively. The irradiance, $F$, is the radiance integrated over some solid angle — for example a hemisphere as in equation (A.0), or, to obtain the net irradiance, over all solid angles as in equation (A.7).
Figure A.2: In (a) the irradiance (in $\frac{W}{m^2}$), $F$, would be all radiant energy passing through the area $dA$ per second. The radiance $I(\Omega)$ (in $\frac{W}{m^2sr}$) in (a) is the radiant energy passing through an area $dAcos\theta$ in the direction $\Omega$ (or into an infinitesimal cone centered at $\Omega$) per second. Because of the lack of perspective in (a) the projection in one dimension of $dA$ onto the direction of $\Omega$, which is at an angle $\theta$ with $n$, is illustrated in (b). The projection of $dA$ onto the direction $\Omega$ becomes $dAcos\theta$. 
B Discrimination

Consider the set of forecast-observation pairs in table B.1.

<table>
<thead>
<tr>
<th>Forecast</th>
<th>Observation</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2.0</td>
<td>1.0</td>
</tr>
<tr>
<td>5.0</td>
<td>3.0</td>
</tr>
<tr>
<td>5.0</td>
<td>3.0</td>
</tr>
<tr>
<td>8.0</td>
<td>9.0</td>
</tr>
</tbody>
</table>

Table B.1: A simple example illustrating forecasts that are not completely reliable, but possess discrimination. The values could be mm of rain.

Because that neither

\[
E(x|y = 2.0) = 1.0 \neq 2.0
\]

nor

\[
E(x|y = 5.0) = 3.0 \neq 5.0
\]

nor

\[
E(x|y = 8.0) = 9.0 \neq 8.0,
\]

the requirements to complete reliability are not fulfilled (see e.g. equation ..
The conditional distributions \( p(y|x) \) become
\[
\begin{align*}
p(y = 2.0|x = 1.0) &= 1.0 \\
p(y = 2.0|x = 3.0) &= 0.0 \\
p(y = 2.0|x = 9.0) &= 0.0 \\
p(y = 5.0|x = 1.0) &= 0.0 \\
p(y = 5.0|x = 3.0) &= 1.0 \\
p(y = 5.0|x = 9.0) &= 0.0 \\
p(y = 8.0|x = 1.0) &= 0.0 \\
p(y = 8.0|x = 3.0) &= 0.0 \\
p(y = 8.0|x = 9.0) &= 1.0
\end{align*}
\]

For each forecast value \( Y \), \( p(y = Y|x) \) equals zero for all values of \( X \) except for one, and \( Y \) is therefore perfect discriminatory (Murphy et al., 1989), which should become clear from the following consideration: a forecast value of 2.0 \( mm \) is always followed by an observation of 1.0 \( mm \), and a forecast of 5.0 \( mm \) is always followed by an observation of 3.0 \( mm \), and a forecast of 8.0 \( mm \) is always followed by an observation of 9.0 \( mm \). Although the forecasts are not reliable, each forecast \( Y \) is able to discriminate between the three different events of 1.0 \( mm \), 3.0 \( mm \), and 9.0 \( mm \).
Generalized Discrimination Score

For two ensemble forecasts $y_s$ and $y_t$ of size $M$ (i.e. $y_s = (y_{s,1}, \ldots, y_{s,M})$ and $y_t = (y_{t,1}, \ldots, y_{t,M})$), the following value is constructed:

$$F_{s,t} = \frac{\sum_{j=1}^{M} r_{s,t,j} - \frac{M(M+1)}{2}}{M^2},$$

(C.1)

with $r_{s,t,j}$ begin the rank of $y_{s,j}$ with respect to the set of pooled ensemble members $y_{s,1}, y_{s,2}, \ldots, y_{s,M}, y_{t,1}, y_{t,2}, \ldots, y_{t,M}$, if sorted in ascending order (Weigel and Manson [2011]).

If in the pooled sample of $2M$ ensemble members, the ensemble members of $y_s$ occupy the $M$ smallest ranks $(1, \ldots, M)$ — which will sum to $\frac{M}{2}(M+1)$ — all ensemble members of $y_s$ are exceeded by those of $y_t$ and the numerator will be 0, which corresponds to the intuitive. If the converse is true (if all ensemble members of $y_s$ exceeds those of $y_t$), the first term of the numerator will be equal to $\frac{M}{2}(2M + (M + 1))$ (M ranks occupying ranks from $M + 1$ to $2M$) and so $F_{s,t}$ will be 1. If every second of the ranks from 1 to $2M$ is occupied by $y_s$ (the others occupied by $y_t$), $F_{s,t} = 0.5$. In such a situation, it cannot be determine which ensemble forecast $y_s$ or $y_t$ is greater. If more than half of the ensemble members of $y_s$ ($y_t$) exceeds those of $y_t$ ($y_s$), $F_{s,t} > (\leq)0.5$.

Determining the rank of $y_s$, $R_s$, within the set of $N$ ensemble forecasts $y_1, y_2, \ldots, y_N$

$$R_s = 1 + \sum_{t=1, t\neq s}^{N} u_{s,t} \quad \text{with} \quad (C.2)$$

$$u_{s,t} = \begin{cases} 
1 & \text{if } F_{s,t} > 0.5 \\
0.5 & \text{if } F_{s,t} = 0.5 \\
0 & \text{if } F_{s,t} < 0.5.
\end{cases}$$

\footnote{A sum of integers is the number of integers in the sum $(n - k + 1)$ times the sum of the last term in the sum $(n)$ and the first term $(k)$, and then this number is divided by 2. The sum of integers from $k$ to $n$, $\sum_{i=k}^{n} i$ is $\frac{(n-k+1)(n+k)}{2}$.}
The rank of $y_s$ When all ensemble members of $y_t$ exceeds those of $y_s$ and $F_{s,t} = 0$, $u_{s,t} = 0$. Comparing the two ensemble forecasts $y_s$ and $y_t$ will in this case not increase the rank, $R_s$, of ensemble forecast $y_s$. In a case when it cannot be determined if $y_s > y_t$ or vice verse, $F_{s,t} = 0.5$ and the contribution to $R_s$ is 0.5.

With $N$ observations ($x = x_1, \ldots, x_N$) and $N$ ensemble ranks ($R = R_1, \ldots, R_N$), the following generalized discrimination score is obtained

$$D = \frac{1}{2}(\tau_{R,x} + 1) \quad \text{(C.3)}$$

as defined in equation (C.2).
Summary

To comply with an increasing demand for sustainable energy sources, a solar heating unit is being developed at the Technical University of Denmark. To make optimal use — environmentally and economically —, this heating unit is equipped with an intelligent control system using forecasts of the heat consumption of the house and the amount of available solar energy. In order to best make use of this solar heat unit, accurate forecasts of the available solar radiation are essential. However, because of its sensitivity to local meteorological conditions, the solar radiation received at the surface of the Earth can be highly fluctuating and challenging to forecast accurately.

Within the project on developing a solar heating unit, it is the role of the Danish Meteorological Institute (DMI) — and thus of my Ph.D. project — to deliver accurate forecasts of the global, direct, and diffuse radiation. An evaluation of these forecasts has revealed shortcomings in the radiation scheme of DMI's HIRLAM models, which has subsequently been revised adjusting the ratio of direct to diffuse radiation penetrating a cloud layer and reaching the surface of the Earth.

As a mean of complying with the accuracy requirements to forecasts of global, direct, and diffuse radiation, the uncertainty of these forecasts is of interest. Forecasts uncertainties become accessible by running an ensemble of forecasts. To this end, global, direct, and diffuse radiation have since August 2011 been output parameters from DMI’s high-resolution ensemble prediction system — aimed at capturing small-scale weather features of which a solar heating unit can be expected to be sensitive.

From the investigation of the degree to which the ensemble members and the truth — here materialised by the verifying observation — are statistically indistinguishable, the appropriateness of complementing forecast values with uncertainty estimates derived from the ensemble forecasts has been assessed. A degree of under-dispersion of the ensemble members is evident, and the ensemble forecasts will tend to express too little uncertainty in the forecast values of global radiation. Under-dispersiveness of ensemble forecasts is a familiar problem in ensemble prediction. Some of the under-dispersiveness might be attributed to uncertainties on the observations.

Concerning RMSE, skill score, and discrimination, the ensemble mean is
seen to perform well compared to the control forecasts. It might be valuable, also for other parameters, to somehow complement the operational forecast with the value of the ensemble mean.
Bibliography


Klaus Wyser, Laura Rontu, and Hannu Savijärvi. Introducing the Effective Radius into Fast Radiation Scheme of a Mesoscale Model. Contributions to Atmospheric Physics, 72(3), 1999.