Weather Forecast Requirements and Evaluation Methods

Technical Report · July 2016
DOI: 10.13140/RG.2.2.10959.94883

CITATIONS
0

READS
22

5 authors, including:

Marion Schroedter-Homscheidt
German Aerospace Center (DLR)
234 PUBLICATIONS 1,760 CITATIONS

Martin Gastón
Centro Nacional de Energías Renovables
64 PUBLICATIONS 521 CITATIONS

Carlos M Fernández-Peruchena
Centro Nacional de Energías Renovables
98 PUBLICATIONS 380 CITATIONS

Isabel Martínez Marco
Agencia Estatal de Meteorología
19 PUBLICATIONS 102 CITATIONS

Some of the authors of this publication are also working on these related projects:

Final thesis View project
PreFlexMS View project
## Deliverable D4.1

**Weather Forecast Requirements and Evaluation Methods**

<table>
<thead>
<tr>
<th>Dissemination level¹</th>
<th>PU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nature²</td>
<td>R</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Document Identifier</th>
<th>PREFLEXMS_DEL_D4.2_20160531_v2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Status</td>
<td>Version 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lead beneficiaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>DLR: M. Schroedter-Homscheidt</td>
</tr>
<tr>
<td>AEMET: Jose Antonio Garcia-Moya Zapata, Isabel Martinez Marco</td>
</tr>
<tr>
<td>CENER: Carlos Fernandez Peruchena, Martin Gaston Romeo</td>
</tr>
</tbody>
</table>

¹ Dissemination level: PU = Public, PP = Restricted to other programme participants (including the JU), RE = Restricted to a group specified by the consortium (including the JU), CO = Confidential, only for members of the consortium (including the JU)

² Nature of the deliverable: R = Report, P = Prototype, D = Demonstrator, O = Other
Executive summary

PreFlexMS develops a dispatch optimizer and provides an optimized production schedule. The outcomes will be used to assess the absolute profitability of a new PreFlexMS CSP as well as the incremental value of the various innovative components of the PreFlexMS CSP.

This deliverable has been prepared by DLR in collaboration with CENER and AEMET within WP4 ‘Weather forecasting and measurement for renewable energy predictability’. It aims at collecting all input needed for the WP4 from the market scenario analysis and the dispatch optimizer developers team in PreFlexMS, presents what is available in weather forecasts and how this needs to be converted. It also serves as a handbook for metrics in forecast evaluation.

We reviewed and extracted the relevant information from previous PreFlexMS results on market scenarios about cost functions for a day-ahead spot market with hourly prices for different countries (chapter 1). Detailed discussions with the power plant simulation team helped to clarify use cases being planned for the optimizer. This interaction was needed to clarify about available information on uncertainty from numerical weather prediction (NWP) providers and how this could be used in the optimizer. Chapter 2 summarizes this information and defines the parameters direct normal irradiance, wind speed and ambient temperature as provided by NWP.

Chapter 3 shortly introduces principles of deterministic numerical weather prediction. Afterwards a suite of evaluation metrics is defined, discriminating between indicators of dispersion, distribution similarity, variability, and temporal distortion.

Chapter 4 shortly introduces principles of probabilistic numerical weather prediction. Afterwards it defines evaluation metrics which can be of use for the power plant simulation team in evaluating the ensemble of schedules being created in the optimizer based on ensemble numerical weather prediction. And finally, statistical post-processing for calibrating ensemble forecasts is discussed.
Table of Contents

1. Information provided from the market scenario analysis 10
   1.1.1 The ‘business-as-usual’ scenario 10
   1.1.2 The ‘Advanced support scheme’ scenario 11
   1.1.3 The ‘Market price’ scenario 12
   1.1.4 Summary time dependencies per country 16

2. Information from the dispatch optimizer developers 19
   2.1 Resulting requirements for meteorological forecasts 19
      2.1.1 Use case: Optimizer use to provide a day ahead schedule 20
      2.1.2 Use case: Optimizer use to provide an update of a schedule 20
   2.2 Parameters needed 20
      2.2.1 Direct normal irradiance 20
      2.2.2 Wind speed 21
      2.2.3 Ambient Temperature 22
   2.3 Missing data definition 22
   2.4 Temporal resolution 22
   2.5 Temporal availability 22
   2.6 Uncertainty information as input to the dispatch optimizer 22
   2.7 MAE vs RMSE vs accumulated values as error metric 23

3. Deterministic forecasts evaluation metrics 24
   3.1 Principles of deterministic weather forecasts 24
   3.2 Metrics to evaluate deterministic forecasts 24
      3.2.1 Class A – Indicators of dispersion 25
      3.2.2 Class B – indicators of distribution similarity 26
      3.2.3 Class C – indicators of variability 27
      3.2.4 Class D – indicators of temporal distortion 27

4. Probabilistic forecast evaluation metrics 29
   4.1 Principles of probabilistic weather forecasts 29
   4.2 Parameters needed and their temporal characteristics 30
   4.3 Uncertainty representation 30
   4.4 EPS verification 34
      4.4.1 Rank histogram 36
      4.4.2 Spread-error diagram 36
      4.4.3 Brier Score 37
      4.4.4 Ranked Probability Score 38
      4.4.5 Continuous Ranked Probability Score 39
      4.4.6 Quantile score 39
   4.5 Statistical post-processing for calibrating the EPS 41
      4.5.1 Bayesian Model Averaging 41
      4.5.2 Logistic Regression and Extended Logistic Regression 41

5. References 43
Acronyms

3D-Var  Three-Dimensional Variational data assimilation
4D-Var  Four-Dimensional Variational data assimilation
AEMET  Agencia Estatal de Meteorología
AEMET-SREPS  AEMET Short Range Ensemble Prediction System
AERONET  Aerosol Robotic Network
AIREP  Air Report
ALADIN  Aire Limitée Adaptation Dynamique Initialisation
AMSU  Advanced microwave sounding unit
AOD  Aerosol optical depth
APOLLO  AVHRR Processing scheme Over clouds, Land and Ocean
AROME  Applications of Research to Operations at Mesoscale
AROME/HARMONIE  Hirlam Aladin Regional/Meso-scale Operational NWP in Europe
ARPEGE  Model from Météo-France
ATOVs  Advanced TIROS-N Operational Vertical Sounder
BC  boundary condition
BMA  Bayesian Model Averaging
BSRN  Baseline Surface Radiation Network
BS  Brier Score
BSS  Brier Skill Score
CAMA  Pampa Camarones (Station in Chile)
CAMS  Copernicus Atmosphere Monitoring Service
CDF  cumulative distribution function
CENER  Centro Nacional de Energías Renovables
CMC-GDPS  Global Deterministic Prediction System from the Canadian Meteorological Centre
CNRM/GMME  Centre National de Recherches Méteorologiques
CPI  Combined Performance Index
CRPS  Continous Ranked Probability Score
CSP  Concentration Solar Power
DCH  direction changes in DNI
DHI  diffuse hemispherical irradiation
DLR  German Aerospace Center
DMO  Direct Model Output
DNI  direct normal irradiation
DNICast  Project acronym www.dnicast-project.net
DRIBU  buoys observations
DRIFTER  observations from drifting buoys
ECMWF  European Centre for Medium-Range Weather Forecast
ECMWF/EPS  ECMWF/Ensemble Prediction System
ECMWF/IFS  ECMWF/Integrated Forecast System
EDA  Ensemble Data Assimilation
ELR  extended logistic regression
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>em</td>
<td>ensemble mean</td>
</tr>
<tr>
<td>EMPS</td>
<td>Evora Molten Salt Platform</td>
</tr>
<tr>
<td>EnerMENA</td>
<td>project acronym</td>
</tr>
<tr>
<td>EPS</td>
<td>ensemble prediction system</td>
</tr>
<tr>
<td>ETKF</td>
<td>Ensemble Transform Kalman Filter</td>
</tr>
<tr>
<td>FONDEF</td>
<td>El Fondo de Fomento al Desarrollo Científico y Tecnológico, Chile</td>
</tr>
<tr>
<td>GHI</td>
<td>global horizontal irradiation</td>
</tr>
<tr>
<td>GIZ</td>
<td>Deutsche Gesellschaft für Internationale Zusammenarbeit</td>
</tr>
<tr>
<td>GRT</td>
<td>Graaff-Reinet (SAURAN station)</td>
</tr>
<tr>
<td>gSREPS</td>
<td>AEMET New Mesoscale Short Range Ensemble Prediction System</td>
</tr>
<tr>
<td>Heliosat-4</td>
<td>surface downwelling solar irradiance model by MINES/DLR</td>
</tr>
<tr>
<td>HIRLAM</td>
<td>High Resolution Limited Area Model</td>
</tr>
<tr>
<td>HITRAN</td>
<td>High Resolution Transmission Molecular Absorption Database</td>
</tr>
<tr>
<td>HRES</td>
<td>high-resolution forecast</td>
</tr>
<tr>
<td>IC</td>
<td>initial condition</td>
</tr>
<tr>
<td>IDEO</td>
<td>Inca de Oro (Station in Chile)</td>
</tr>
<tr>
<td>IR</td>
<td>infrared</td>
</tr>
<tr>
<td>JMA-GSM</td>
<td>Global Spectral Model from the Japan Meteorological Agency</td>
</tr>
<tr>
<td>JNNSM</td>
<td>Jawaharlal Nehru National Solar Mission</td>
</tr>
<tr>
<td>kcB</td>
<td>clear sky beam index</td>
</tr>
<tr>
<td>KSI</td>
<td>Kolgorov-Smirnov test integral</td>
</tr>
<tr>
<td>KZH</td>
<td>University of KwaZulu-Natal Howard College (SAURAN station)</td>
</tr>
<tr>
<td>KZW</td>
<td>University of KwaZulu-Natal Westville (SAURAN station)</td>
</tr>
<tr>
<td>LAF</td>
<td>lagged average forecast</td>
</tr>
<tr>
<td>LAM</td>
<td>limited-area models</td>
</tr>
<tr>
<td>LBC</td>
<td>lateral boundary conditions</td>
</tr>
<tr>
<td>LR</td>
<td>logistic regression</td>
</tr>
<tr>
<td>LTEKF</td>
<td>Local Transform Ensemble Kalman Filter</td>
</tr>
<tr>
<td>LW</td>
<td>long wave</td>
</tr>
<tr>
<td>MAD</td>
<td>mean absolute difference</td>
</tr>
<tr>
<td>MAE</td>
<td>mean absolute error</td>
</tr>
<tr>
<td>McRad</td>
<td>radiative transfer model used in ECMWF/IFS</td>
</tr>
<tr>
<td>MENA</td>
<td>Mediterranean Europe and Northern Africa regions</td>
</tr>
<tr>
<td>Met Office</td>
<td>United Kingdom Meteorological Office</td>
</tr>
<tr>
<td>MSG</td>
<td>Meteosat Second Generation</td>
</tr>
<tr>
<td>NCEP</td>
<td>National Center for Environmental Prediction</td>
</tr>
<tr>
<td>NCEP-GFS</td>
<td>the Global Forecast Model from NCEP</td>
</tr>
<tr>
<td>NEMS</td>
<td>NOAA Environmental Modeling System</td>
</tr>
<tr>
<td>NMHM</td>
<td>Non-hydrostatic Multi-Scale Model</td>
</tr>
<tr>
<td>NMU</td>
<td>Nelson Mandela Metropolitan University (SAURAN station)</td>
</tr>
<tr>
<td>NOAA</td>
<td>National Oceanic and Atmospheric Administration</td>
</tr>
<tr>
<td>NWP</td>
<td>numerical weather prediction</td>
</tr>
<tr>
<td>PDF</td>
<td>Probability Density Function</td>
</tr>
<tr>
<td>Term</td>
<td>Definition</td>
</tr>
<tr>
<td>--------</td>
<td>---------------------------------------------------------------------------</td>
</tr>
<tr>
<td>PILOT</td>
<td>observations from pilot sounding</td>
</tr>
<tr>
<td>PPA</td>
<td>Power purchase agreements</td>
</tr>
<tr>
<td>PSA</td>
<td>Plataforma Solar de Almeria</td>
</tr>
<tr>
<td>QC</td>
<td>quality control</td>
</tr>
<tr>
<td>QS</td>
<td>quantile score</td>
</tr>
<tr>
<td>R&amp;D</td>
<td>research and development</td>
</tr>
<tr>
<td>REFIT</td>
<td>Renewable Energy Feed-in Tariff</td>
</tr>
<tr>
<td>REIPPP</td>
<td>Renewable Energy Independent Power Producer Procurement programme</td>
</tr>
<tr>
<td>relMAD</td>
<td>relative mean absolute difference</td>
</tr>
<tr>
<td>relRMSD</td>
<td>relative root mean square difference</td>
</tr>
<tr>
<td>relSD</td>
<td>relative standard deviation of the residuals</td>
</tr>
<tr>
<td>REST2</td>
<td>a radiative transfer model</td>
</tr>
<tr>
<td>RMSD</td>
<td>root mean square difference</td>
</tr>
<tr>
<td>RMSE</td>
<td>root mean square error</td>
</tr>
<tr>
<td>ROC</td>
<td>relative operating characteristic curve</td>
</tr>
<tr>
<td>RPS</td>
<td>Ranked Probability Score</td>
</tr>
<tr>
<td>RPSS</td>
<td>ranked probability skill score</td>
</tr>
<tr>
<td>RR</td>
<td>Ramp Rate</td>
</tr>
<tr>
<td>RRTM</td>
<td>Rapid Radiation Transfer Model</td>
</tr>
<tr>
<td>RSI</td>
<td>rotating shadowband irradiometers</td>
</tr>
<tr>
<td>RUC</td>
<td>relative user characteristic</td>
</tr>
<tr>
<td>RUC1</td>
<td>hourly rapid updated circle</td>
</tr>
<tr>
<td>RVD</td>
<td>GIZ Richtersveld (SAURAN station)</td>
</tr>
<tr>
<td>SAURAN</td>
<td>Southern African Universities Radiometric Network</td>
</tr>
<tr>
<td>SHIP</td>
<td>automatic ship observations</td>
</tr>
<tr>
<td>SLAF</td>
<td>scaled algged averaged forecast technique</td>
</tr>
<tr>
<td>SPPT</td>
<td>Stochastically Perturbed Parameterization Tendencies</td>
</tr>
<tr>
<td>STA</td>
<td>Mangosuthu Univ. of Technology STARlab (SAURAN station)</td>
</tr>
<tr>
<td>STDV</td>
<td>standard deviation</td>
</tr>
<tr>
<td>SUN</td>
<td>Stellenbosch University (SAURAN station)</td>
</tr>
<tr>
<td>SUT</td>
<td>Eskom Sutherland (SAURAN station)</td>
</tr>
<tr>
<td>SW</td>
<td>shortwave</td>
</tr>
<tr>
<td>SYNOP</td>
<td>synoptical observations</td>
</tr>
<tr>
<td>SZA</td>
<td>solar zenith angle</td>
</tr>
<tr>
<td>TEMP</td>
<td>Temp measurement observations</td>
</tr>
<tr>
<td>TKE</td>
<td>turbulent kinetic energy</td>
</tr>
<tr>
<td>U95%</td>
<td>uncertainty at 95%</td>
</tr>
<tr>
<td>UFS</td>
<td>GIZ University of Free State (SAURAN station)</td>
</tr>
<tr>
<td>UNV</td>
<td>USAid Venda (SAURAN station)</td>
</tr>
<tr>
<td>UNZ</td>
<td>University of Zululand (SAURAN station)</td>
</tr>
<tr>
<td>UPR</td>
<td>GIZ University of Pretoria (SAURAN station)</td>
</tr>
<tr>
<td>UV</td>
<td>ultraviolet</td>
</tr>
<tr>
<td>VAN</td>
<td>Vanrhynsdorp (Station in South Africa)</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>VAN</td>
<td>GIZ Vanrhynsdorp (SAURAN station)</td>
</tr>
<tr>
<td>VRY</td>
<td>GIZ Vryheid (SAURAN station)</td>
</tr>
<tr>
<td>WCRP</td>
<td>World Climate Research Program</td>
</tr>
<tr>
<td>WMO</td>
<td>World Meteorological Organisation</td>
</tr>
</tbody>
</table>
1. Information provided from the market scenario analysis

Deliverable D2.1 ‘Pricing scenarios’ provides cost functions for a day-ahead spot market with hourly prices for different countries South Africa, India, Spain, Morocco, Chile, and Saudi Arabia. In this section we copy input from this deliverable with a focus on the aspects being relevant for WP4.

Current hourly prices (from 2014) are given for a whole year. Additionally, an outlook for 2030 hourly prices in a whole year is provided. Both datasets serve as input for the power plant simulation team.

In the case of wholesale market dominated price structures, average daily price curves are given for the winter and summer half year. The winter half year is defined as October to March for the Northern hemisphere (Spain, India, Saudi Arabia and Morocco) April to September for the Southern hemisphere (South Africa, Chile).

Three scenarios for 2030 are distinguished per country: (a) the ‘Business-as-usual support scheme’ where a PreFlexMS CSP competes only against traditional CSP power plants in nowadays market conditions and feed-in or prevailing power purchase agreement support schemes, (b) the ‘Advanced support scheme’ scenario where a PreFlexMS CSP competes partially against other renewables and conventional power plants as it obtains some support as a CSP plant and for the rest it is treated as a wholesale market participant, and (c) a ‘Market price’ scenario where a PreFlexMS CSP competes fully with all other renewables and conventional power plants as there is no CSP specific additional support scheme anymore.

1.1.1 The ‘business-as-usual’ scenario

For 2030 the same temporal behaviour is expected, only the values on the vertical axis are adapted according to expected cost reductions for conventional CSP power plants.

Spain
‘Special Regime’: Feed-in tariff for CSP for 25 years. This scheme was replaced in 2014 by a scheme with a feed-in tariff ensuring a fixed ‘reasonable’ return over the power plant’s lifetime. Penalties apply if regulatory power need is caused by the overall renewable energy schedule vs actual production.
Chile
Power purchase agreements (PPA) are used to contract CSP power plants. There is no known time-of-the-day dependency at the moment.

India
Jawaharlal Nehru National Solar Mission (JNNSM) programme: Power purchase agreements (PPA) in an auction where the lowest bids win. Inside the PPA a pre-determined price over 25 years is guaranteed. There is no time-of-the-day dependency involved.
Shortfalls below -20% of the declared schedule result in compensation payments to the grid operator.

Morocco
Power purchase agreements (PPA) are used to contract CSP power plants. Peak prices are paid between 17:00 to 22:00 local time. In off-peak hours only 85% of the peak tariff is paid.

Saudi Arabia
There are no contracted CSP plants at the moment, so there is no information about any time-of-the-day dependencies.

South Africa
Renewable Energy Feed-in Tariff (REFIT) programme 2008; Renewable Energy Independent Power Producer Procurement (REIPPP) programme 2011: Predetermined price for electricity for the next 20 years in a PPA with ESKOM (state-owned utility as single buyer). Feed-in tariffs vary in peak and off peak (baseline) hours. Peak hours are 16:30 to 21:30, baseline feed-in tariff hours are 5:00 – 16:30 and 21:30 – 22:00; and between 22:00 and 5:00 the feed-in tariff is zero. In peak hours the plants receive 270 % of the baseline tariff. There are no imbalance charges – not fulfilling the schedule as forecasted the day ago is not penalized.

1.1.2 The ‘Advanced support scheme’ scenario
In an advanced support scheme a fixed part is paid unless CSP technologies reach grid parity. This is paid additionally to a time-variant part reflecting the market price variations. Generally, it is expected that market prices will be decreasing during noon hours due to larger shares of photovoltaics (PV). This is due to the fact that PV generation needs to be traded as it occurs as PV-based generation has only restricted storage and self-consumption capacities.

Spain
The Spanish day-ahead market is run in an hourly temporal resolution.

Chile
The day-ahead market in Chile is run in an hourly temporal resolution.

India
The Indian day-ahead market is run in a quarter hourly temporal resolutions.

Morocco
There is no wholesale market with quoted prices available nowadays.
Saudi Arabia
There is no wholesale market with quoted prices available nowadays.

South Africa
The South African day-ahead market is run in an hourly temporal resolution.

1.1.3 The ‘Market price’ scenario
Spain
The Spanish day-ahead market is run in an hourly temporal resolution.

Chile
The day-ahead market in Chile is run in an hourly temporal resolution.

India
The Indian day-ahead market is run in a quarter hourly temporal resolution.

Morocco
There is no wholesale market with quoted prices available nowadays.

Saudi Arabia
There is no wholesale market with quoted prices available nowadays.

South Africa
The South African day-ahead market is run in an hourly temporal resolution.
Figure 3. Wholesale market prices in 2014 (hourly resolved; split for summer and winter half year). Note that such data is not available for Morocco and Saudi Arabia.
Figure 4. Scenario 3 (market price outlook 2030)
Figure 5. Seasonal wholesale profiles 2014 and 2030 for South Africa

Figure 6. Seasonal wholesale profiles 2014 and 2030 for India
1.1.4 Summary time dependencies per country

Figure 7. Overview of the pricing scenarios for South Africa

Figure 8. Overview of the pricing scenarios for India
Figure 9. Overview of the pricing scenarios for Spain

Figure 10. Overview of the pricing scenarios for Morocco
Figure 1. Overview of the pricing scenarios for Chile

Figure 12. Overview of the pricing scenarios for Saudi Arabia
2. Information from the dispatch optimizer developers

PreFlexMS develops a dispatch optimizer and provides an optimized production schedule. The outcomes will be used to assess the absolute profitability of a new PreFlexMS CSP as well as the incremental value of the various innovative components of the PreFlexMS CSP. Here we extract information from PreFlexMS D3.1 ‘Report on reference plant configuration for dispatch optimization and machine learning’ which is relevant for our work.

Priority in dispatch optimization within PreFlexMS is the maximization of revenues for sold electricity taking the once-through steam generator’s flexibility features into account. Other aspects as optimized operations of the solar field or the storage are not treated in PreFlexMS. One may refer to Hirsch et al. (2014) for a more detailed discussion on issues like short-term decision on start-up and shut-down procedures or the control of mass flow within the solar field. Within PreFlexMS it is assumed that the power plant is always in the thermodynamic equilibrium and that the transient effects in start-up and shut-down situations can be described by a parameterization.

Potential advantages as higher reliability of CSP power generation in negotiations for grid access negotiations are also not in the focus of PreFlexMS.

Also, the perspective of the grid operator being responsible for grid stability as a whole or a market agent being responsible for the trading of a whole renewable power plant portfolio are not under investigation here. We concentrate on the view of the power plant operator in PreFlexMS – this may include external forcings being imposed by the grid operator requests or the market agent.

Machine learning treats the dynamics of operating the power block and therefore is updated regularly in operations. It relies on observations of ambient temperature and wind speed (describing the cooling process) as obtained in the power block. The optimizer uses hourly forecasts of ambient temperature and wind speed and applies the machine learning results as a parameterization.

The optimizer provides an income in EUR for each hour and day. This is based on hourly production * remuneration.

A successful optimization is measured as follows:

- The reference case is a simple optimization approach (tbd by the dispatch optimizer team).
- It is assumed that the reference case when using the ideal (=correct) forecast is the truth.
- Real observations (DNI, T, WS) are taken as ideal (=correct) forecast – quality of result is only a function of the optimizer.
- Real forecasts (DNI, T, WS) are taken as realistic (= may be wrong) forecasts – quality of result is a function of optimizer and forecast quality. They may interact non-linearly. This may include the use and comparison of various forecast options.
- Any other optimization strategy performs better or worse compared to the reference case in EUR.
- It may happen that external constraints have to be followed (e.g. grid control commands). This results in a no/yes decision (or may be imposed by very large monetary incentives and therefore influence the EUR result).

A close collaboration with the dispatch optimizer team will define if there are any typical critical situations for the optimizer. This may impose priorities for evaluation metrics. This will be a later outcome in the project and is not part of this deliverable.

2.1 Resulting requirements for meteorological forecasts

In this study we deal with the assessment of 15 minute and hourly resolved forecasts in a 48 hour forecast horizon. We focus on two use cases providing different requirements for forecasting.
Generally in both use cases the aim is to
• suggest optimum plant operations/storage management for rest of today
• to cope with yesterday's day ahead forecast errors which become clear during today
• avoid short position at the end of the day and at any hour of today left at time optimizer run; KPI: energy being short as sum of rest of the hours during the day; This KPI can be transferred into EUR value by assuming e.g. a 50% penalty of the price curve as provided in the day ahead trading use case.

The economic point of view has several aspects:
- Avoid a short position in any hour/at the end of the day ( = being not able to deliver as much as promised).
- Avoid a long position in any hour when the storage is already filled ( = being not able to get potential higher prices).
- Make sure to avoid short position, but may also sure to fulfil any intra-day options to sell more. Drawback: this requires quite some assumption on the intra-day market volume/prices. This is not known for the moment.
- Ensure exact production according to the schedule in each hour to show reliability and predictability of CSP in each hour (this might be required for grid access license in future).
- Allow optimum technical operations of the power plant (not focus of the PreFlexMS project).

2.1.1 Use case: Optimizer use to provide a day ahead schedule
Run once per day – has a single 48 hour long forecast as input.
The forecast length is 48 hours. It starts at 00 UTC and provides values for today and tomorrow in hourly or 15 minutes temporal resolution.

Time between the release of the numerical weather prediction (NWP) run and closure date available for processing should be 2-3 hours. Therefore, take the latest available forecast (e.g. 6 UTC run available 11 UTC for market closure of 14 UTC allows 2-3 hours processing time).

2.1.2 Use case: Optimizer use to provide an update of a schedule
The first run is taken from the day-ahead schedule case, but hourly we evaluate and update the schedule several times.

Trigger moments for an update run of the optimizer can be
• new NWP input is available
• grid control command obtained
• deviation to observations found
• price incentive signal received.

Therefore, a short processing time is required.

2.2 Parameters needed

2.2.1 Direct normal irradiance
Due to interactions between the radiation and the atmosphere (scattering and absorption) the terrestrial solar radiation is divided into two components. Direct normal irradiance (DNI) refers to solar photons that reach the surface without being scattered or absorbed. Diffuse irradiance refers to such photons that reach the observer after one or more interactions with the atmosphere.

In the strict sense of the definition, DNI is the irradiance on a surface perpendicular (called also normal) to the vector from the observer to the center of the sun consisting of radiation that did not interact with the atmosphere. It does not include circumsolar radiation which is scattered into the forward direction and from the observer’s point of view seem to come from the sun disk.

This strict definition is useful for atmospheric physics and used in some radiative transfer models, but brings along a complication for ground observations: It is not possible to measure whether or not a photon was scattered if it
reaches the observer from the direction in which we see the solar disk. Therefore, DNI is interpreted differently in the world of measurements and also in solar energy applications.

Direct solar radiation is understood as the “radiation received from a small solid angle centered on the sun’s disk”. The size of this small solid angle for DNI measurements is recommended to be $5 \cdot 10^{-3}$ srad$^3$ (corresponding to 2.5° half angle) (WMO, 2010). This recommendation is approximately 10 times larger than the radius of the solar disk itself (yearly average 0.2665°). This is due to the fact that instruments for DNI measurements (pyrheliometers) have to be tracked to follow the path of the sun and small tracking errors must be allowed. The large field of view (FOV) of pyrheliometers reduces the effect of such tracking errors.

The circumsolar radiation is treated in several NWP models by the concept of using scaled optical depths. Among them the ECMWF/IFS (European Centre for Medium-Range Weather Forecasts; Integrated Forecast System) and AROME/HARMONIE (Hirlam Aladin Regional/Meso-scale Operational NWP In Europe) as being used in PreFlexMS. This method is described e.g. in Joseph and Wiscombe (1976) as used in the ECMWF IFS and Toll et al., 2016 describing the HIRLAM implementation which is used in AROME/HARMONIE as well. The strong forward peak of cirrus cloud and aerosol particles is numerically solved by assuming a scattering phase function without forward scattering, but by correcting this assumption by a reduced optical depth. It may create DNI values which are systematically larger than DNI as measured by a pyrheliometer in thin cirrus or aerosol conditions as the effective field of view of this method is larger than the 2.5° half-angle used by the pyrheliometer.

Requirement: Provide DNI as measured with pyrheliometers. Provide average power (W/m²) over the time interval.

Note: NWP provides aggregated irradiation for the last output interval (e.g. 15 minutes, 1 hour, 3 hours). There may be the need to re-map to 15 minutes average power (to be done by CENER).

### 2.2.2 Wind speed

Wind speed in 10 m height, wind gust and wind direction are important parameters for the CSP plant yield. The wind gust is also called peak gust and defined as the maximum 3-second wise averaged wind speed in a given time interval. Intervals of about 3 s are useful for most applications (WMO, 2010).

Requirement: Provide wind speed, direction, and wind gust in 10 m height. Additionally, wind speed is needed at approx. 200 m height (evaporator facility and tower height) for the simulation of thermal efficiencies and cooling processed.

Note: Meteorological observations of wind speed and direction are generally valid for the last 10 minutes before the official reporting time (WMO, 2010). The NWP centers therefore provide the same as NWP output. So, any wind speed information being forecasted or measured only represents 10 minutes (as discussed e.g. in the ECMWF IFS documentation (ECMWF, 2016). It may be available in a temporal resolution every hour, e.g. 10:00, 11:00, 12:00 etc. But it only means that the values given are representative for 9:51-10:00, 10:51 – 11:00, and 11:51-12:00, respectively.

For wind speed as a parameter, the average of all instantaneous wind speeds within this 10 minute interval is forecasted.

The instantaneous wind speeds are the wind speed modeled at each time step of the NWP model - the internal time step may be 30 seconds or few minutes – depending on the NWP model.

Note: For the wind gust, the maximum value of all 3-second averaged wind speeds in the whole interval (e.g. 10:01 – 11:00) is taken and reported as wind gust.

Recommendation: Wind speed should be interpolated linearly in time (will be done by CENER).

Note: Interpolating wind gust out of 3-hourly ECMWF output is not meaningful. Therefore, wind gust is not provided for ECMWF based forecasts.
2.2.3 **Ambient Temperature**

The ambient temperature is defined as the 2m height temperature as measured and forecasted in the meteorological sector.

Also, ambient temperatures are needed at 10 and approx. 200 m height (evaporator facility and tower height) for the simulation of thermal efficiencies and cooling processed.

**Requirement:** Provide average 2m air temperature over the time interval. Also provide air temperature in 10 and 200 m.

**Note:** NWP provides an instantaneous value at the nominal time of the forecast.

**Recommendation:** Air temperature should be interpolated linearly in time and provided as time interval averages (will be done by CENER).

2.3 **Missing data definition**

Missing/invalid data is marked by -999.0 or -999 for float and integer variables, respectively.

2.4 **Temporal resolution**

The dispatch optimizer is driven by 15 min or hourly resolved price forecasts. Therefore, also the meteorological forecasts need to have an hourly resolution or 15 min. AROME/HARMONIE directly offers this model output resolution.

DNI is the only variable to be provided in this project that is truly intermittent, since it can vary from values of ~1000 W/m² to zero in seconds, as it is strongly influenced by passing clouds. Consequently, a high-performance method will be applied to increase the temporal resolution from 3-h forecasted and post-processed DNI series to 15-min that dynamically assemble site information to provide high frequency DNI series (up to the temporal resolution of available measurements at the site). This will be performed by CENER by applying method like discussed in Fernandez-Peruchena et al. (2015 and 2016).

**Requirement:** NWP provides hourly irradiation; hourly instantaneous values of temperature, wind speed and wind direction; and maximum wind gust during the last 10 minutes before nominal time. They have to be converted into 15 min average values (CENER).

**Note:** This is not provided by standard NWP output. Any transformation/assumption has to be made.

2.5 **Temporal availability**

**Requirement:**

Use case ‘day-ahead schedule’: Be available 2-3 hours before closure date of the market.

Use case ‘update of schedule’: Forecasts should be made available several times during a day. There is no specific requirement on the delivery time of a forecast. Every forecast should be provided as soon as possible and will be used as soon as available.

2.6 **Uncertainty information as input to the dispatch optimizer**

For deterministic forecasts only statistically derived uncertainty information based on assessment of historical forecasts can be provided. They are not derived from the individual forecast itself. It may be a constant value, perhaps in percent of the forecasted value. It may be a constant database e.g. as function of solar zenith angle and/or cloud conditions. It may be a slowly variation over time included as an outcome based on machine learning. But it will certainly not rely on an assessment of each physical situation. This can only be provided by the probabilistic approach.

**Requirement:** Each deterministically forecasted value should be accompanied by uncertainty information. Uncertainty is described by 10%, 25%, 75% and 90% percentile (p10,p25,p75,p90).

For deterministic forecasts the uncertainty is derived by doing:

- Derive a time series of forecast minus observation (FC-OBS) deviations
- Split the FC-OBS into kcB (clear sky beam index being defined as the ratio of actual DNI to its corresponding value under clear sky conditions.) and solar zenith angle (sza) classes.
- Generate databases for p10, p25, p75, p90 values of all kcB/sza classes for FC-OBS from historic forecasts and observations.
- For each forecast value search for the kcB and sza class and obtain p10, p25, p75, p90 values out of the database.
- The data sent by CENER will include the forecast value, the p10, p25, p75, and p90 as 5 separate columns.

Requirement: Each probabilistic forecasted value should be provided with all ensemble members. No percentiles and probabilities are given.

2.7 MAE vs RMSE vs accumulated values as error metric

In case of a grid operator the root mean squared error (RMSE) is the error metric to optimize as it emphasizes the large errors affecting the electricity grid’s security while small errors are of less importance.

The mean absolute error (MAE) is more appropriate when the cost function is linear. In case of a market agent being responsible for a distributed renewable energy power plant portfolio the spatially averaged mean absolute error is of main interest. In case of a single power plant operator, the MAE of the local production is of interest.

On the other hand, the RMSE can be decomposed in several factors (including the bias), so it can give a better understanding of the systematic and non-systematic errors in our forecasts, and how much the post-processing method can reduce any biases.

Note: In case of a power plant with storage, a low MAE is helpful, but over a number of hours the storage allows absolute errors to cancel out over a certain period. Therefore, aggregated DNI may be a good proxy to be evaluated.

Requirement: For all investigations within PreFlexMS it is required to give bias, MAE, and RMSE.
3. Deterministic forecasts evaluation metrics

3.1 Principles of deterministic weather forecasts

Numerical weather prediction is generally performed by numerical integration of the hydrodynamic equations governing atmospheric motions. These non-linear equations describing the evolution of the atmosphere do not have analytical solutions even if the problem is “well posed”. If an analytical solution does not exist, we have to use the numerical techniques to find a certain approximation to the true solution of the system of equations and therefore we have to use computers. Given a description of the current state of the atmosphere, numerical models can be used to propagate this information forwards to produce a forecast for future weather. The state of the atmosphere is described by the spatial distribution of wind, temperature, and other weather variables. By extrapolating the computed tendencies ahead in time, the model can predict the field variables in the future.

In contrast to the original differential equations which describe the whole spectrum of atmospheric motions, the discretized equations describe only processes with certain spatial and temporal scales. Since subgrid-scale processes are not included in models, only their statistical effects on the mean flow are taken into account. The statistical contributions by the different processes must be expressed in terms of the large-scale parameters themselves. The mathematical procedure involved is called parametrization.

Numerical models calculate the heat transfer due to short wave solar radiation and long wave emissions produced by atmospheric gases, clouds and the surface of the Earth. This results also in a provision of global and direct irradiation at the surface.

The AROME/HARMONIE (Hirlam Aladin Regional/Meso-scale Operational NWP In Europe) and ECMWF (European Centre for Medium-Range Weather Forecasts) models will be employed in this project as deterministic models. The AROME/HARMONIE mesoscale model produces hourly-outputs up to 48-hour forecast length. The deterministic ECMWF global model produces 3-hour outputs up to ten days forecast length.

3.2 Metrics to evaluate deterministic forecasts

The validation methodologies and statistical performance indicators proposed for validating the deterministic Direct Normal solar Irradiance (DNI) forecast are briefly depicted below. They can be grouped into five general categories according to their nature:

- Dispersion
- Distribution similitude indicators
- Variability
- Temporal distortion

For the sake of clarity, in the following, the \( i \)th observed data point will be noted \( o_i \) and the corresponding predicted (or modeled) data point will be noted \( p_i \); the mean values of both series, each totaling \( N \) points, are noted \( O_m \) and \( P_m \) respectively.

**Requirement:** Only daytime observations between sunrise and sunset are taken into account. Sunrise/sunset is defined by solar elevation being above 5°. Time intervals (e.g. hour or 15 min period) including sunrise/sunset are included.

The value of 5° has been chosen based on own experience. 5° is relevant for trough technology, while 10° is perhaps more appropriate for tower technologies. Kistler (1986) in the DELSOL3 handbook even choose 15° for tower technologies. However, we chose 5° to allow maximum information and take the pre-warming phase into account. Users of forecasts may decide differently.

Error measures are usually normalized, so the relative metrics are commonly used.

**Requirement:** Normalized or also called relative MAE, bias, and RMSE is divided by the mean of all daytime observations.
Comparing the relative advantages of different forecasting methodologies is not straightforward because different studies use different evaluation criteria, and also because the solar radiation data sets are dependent on geographic location and climate. The second issue is particularly significant as it is relatively easy to forecast the solar irradiance during clear day periods, and therefore an over-simplistic forecast model can yield very good conventional statistical metrics for those conditions when solar irradiance is highly predictable. The same over-simplistic forecast model would certainly fail under different geo-climatic conditions.

Therefore, error metrics can be visualized in a heat map as a function of solar zenith angle (SZA) and cloud conditions. Cloud conditions are described by the beam clear sky index $k_{cb}$ being defined as the ratio of DNI and the clear sky value (e.g. in Ineichen, 2016). The clear sky value can be taken out of the CAMS McClear service.

**Recommendation for the dispatch optimizer development:** Knowing the bias as function of SZA and $k_{cb}$ can be used as static bias correction for individual values as an estimate. The RMSE (again resolved by SZA and $k_{cb}$) can then be split into standard deviation (STDV) and bias, and the STDV can be further used as uncertainty by the optimizer.

**Requirement:** Suitable metrics have to be provided as overall values, as heat maps and as a database being dependent on SZA and $k_{cb}$.

In the following we use and extend the classification of metrics as suggested by Gueymard (2004). The $i$th observed data point will be noted $o_i$ and the $i$th predicted (modelled) data point will be noted $p_i$. The daytime mean values of the two distributions (each totalling $N$ points) are noted $O_m$ and $P_m$ respectively. The $i$th modelled-measured difference in the distribution is $p_i-o_i$.

### 3.2.1 Class A – Indicators of dispersion

The **bias** is the arithmetic average of the values of the differences:

$$bias = \frac{1}{N} \sum_{i=1}^{N} p_i - o_i$$  \hfill (1)

The relative **bias** is the arithmetic average of the values of the differences, normalized by the daytime observational mean:

$$relbias = \left( \frac{100}{O_m} \right) \frac{1}{N} \sum_{i=1}^{N} p_i - o_i$$  \hfill (2)

The **Mean Absolute Difference** (MAD) (also referred to as Mean Absolute Error, MAE) is:

$$MAD = \frac{1}{N} \sum_{i=1}^{N} |p_i - o_i|$$  \hfill (3)

The relative MAD is normalized by the daytime observational mean and defined as following:

$$relMAD = \left( \frac{100}{O_m} \right) \frac{1}{N} \sum_{i=1}^{N} |p_i - o_i|$$  \hfill (4)

The **Root Mean Square Difference** (RMSD) is

$$RMSD = \left[ \frac{\sum_{i=1}^{N} (p_i - o_i)^2}{N} \right]^{1/2}$$  \hfill (5)
The Relative Root Mean Square Difference (relRMSD) is normalized by the daytime observational mean and defined as following:

\[
\text{relRMSD} = \left( \frac{100}{O_m} \right)^2 \left( \frac{\sum_{i=1}^{N} (p_i - o_i)^2}{N} \right)^{1/2}
\]  
(6)

The Relative standard deviation of the residuals (relSD) is normalized by the daytime observational mean and defined as following:

\[
\text{relSD} = \left( \frac{100}{O_m} \right)^2 \left( \frac{\sum_{i=1}^{N} (p_i - o_i)^2 - \left( \sum_{i=1}^{N} (p_i - o_i) \right)^2}{N} \right)^{1/2}
\]  
(7)

Uncertainty at 95% (U95%) is defined as following:

\[
U95\% = 1.96(\text{relSD}^2 + \text{relRMSD}^2)^{1/2}
\]  
(8)

### 3.2.2 Class B – indicators of distribution similarity

The statistical metric used to check the similarity between generated and measured high-frequency DNI values are KSI, OVER and CPI indexes.

**KSI (Kolmogorov–Smirnov test Integral).** This index is defined as the integrated differences between the Cumulative Distribution Functions (CDF) of two data sets (Beyer et al., 2009).

\[
\text{KSI} = \int_{x_{\text{min}}}^{x_{\text{max}}} D_n dx
\]

As \(D_n\) is a discrete variable and the number of integration intervals is identical in all cases, trapezoidal integration is used over the whole range of the independent variable \(x\). A relative value of KSI, KSI(%), is obtained by normalizing the critical area, \(a_{\text{critical}}\), as in Eq. (5).

\[
\text{KSI} (%) = 100 \frac{\int_{x_{\text{min}}}^{x_{\text{max}}} D_n dx}{a_{\text{critical}}}
\]

(9)

where \(a_{\text{critical}}\) is calculated as shown in the following equation:

\[
a_{\text{critical}} = V_c(x_{\text{max}} - x_{\text{min}})
\]

(10)

and \(V_c\) can be determined from the number of values \(N\) in the sample:

\[
V_c = \frac{1.63}{\sqrt{N}}, \quad N \geq 35
\]

(11)
OVER. This index is derived from KSI index, and describes the relative frequency of exceedance situations when the normalized distribution of modeled data points in specific bins exceeds the critical limit that would make it statistically undistinguishable from the reference distribution. OVER (in percent) is obtained as follows (Beyer et al., 2009):

\[
OVER = 100 \frac{\int_{x_{\text{min}}}^{x_{\text{max}}} \max(D_n-D_c,0)dx}{a_{\text{critical}}}
\]  

(12)

OVER is 0 if the normalized distribution always remains below Dc.

Combined Performance Index (CPI). This indicator combines conventional information about dispersion and bias with information about distribution likeness (Gueymard 2012). CPI is defined as follows:

\[
CPI = \left( KSI + OVER + 2RMSD \right) / 4
\]

(13)

3.2.3 Class C – indicators of variability

For quantifying and comparing measured and forecasted DNI variability, the statistical metrics proposed is the ramp rate (RR). For applying this parameter, DNI irradiance under clear sky conditions, DNI\(^{CS}\), must be calculated (for example using REST2 model (Gueymard, 2008) or CAMS McClear (Lefèvre et al., 2013) and substracted from the measured and forecasted DNI, such that the remaining value is the variation from expected irradiance (Lave and Kleissl, 2010). Thus, the RRs of these variations are calculated as the difference between successive data points, using the following equation given in units of W/m\(^2\)min (eq. 1):

\[
RR = \frac{DNI(t)-DNI^{CS}(t)-[DNI(t-1)-DNI^{CS}(t-1)]}{\Delta}
\]

(14)

where \(\Delta\) is the time step of the series, expressed in minutes.

For RR characterization, the following statistical parameters can be calculated: mean, standard deviation, skewness and kurtosis. Mean and standard deviation characterizes, respectively, both the average and dispersion of datasets. Skewness is a measure of the degree of asymmetry of a distribution: if the left tail is more pronounced than the right tail, the distribution skewness is negative. If the reverse is true, the distribution skewness is positive. If the two tails are equal, the distribution skewness is zero. Kurtosis is a measure of whether the data are heavy-tailed or light-tailed with respect to a normal distribution: high kurtosis is indicative of pronounced peaked distribution, or outliers, whereas low kurtosis is indicative of flat distributions.

3.2.4 Class D – indicators of temporal distortion

To complement the traditional verification measures previously shown, it is desirable to evaluate the impact that time misalignments produces in the forecast accuracy. This new measure of error allows a deeper knowledge of the prediction model behavior, besides a bi-criteria perspective to the problem of comparing different forecasts.
The process used to characterize the temporal component of the prediction error is based in Dynamic Time Warping (DTW) principles which obtain the optimal alignment of two time series by applying dynamic optimization to a shortest path problem [Sakoe 1971]. The rationale behind DTW is the following: given two time series the method stretches and/or compresses them locally in order to make one resemble the other as much as possible. So, we work with two time series, the test (or prediction) series \( T = T(T_1, T_2, ..., T_N) \) and the reference (or measured) series \( R = R(R_1, R_2, ..., R_N) \). The DTW processing produces a set of time modifications in the test series to align it into the reference series.

The modifications of the time axis of the test series are offered by mean of the solution of an optimal path problem. So, the start point is the definition of a local distance \( d \) between pairs of points \( T_i \) and \( R_j \) of both series. After that, the concepts of path between series as a sequence of points \( w_l = (i_l, j_l) \in [1:N] \times [1:N] \) for \( l \in [1:k] \) and the total cost associated to a path \( c_w(T, R) = \sum_{i=1}^{k} d(T_i, R_j) \) where \( (i_l, j_l) = w_1 \) is used.

Accordingly, an optimal path between the series \( T \) and \( R \) is a path \( w^* \) that presents a minimum total cost taking into account all the possible paths. The optimal path \( w^* \) is obtained by using dynamic programming (DP), a procedure that solves optimization problems by breaking them down into simpler ones. The basis of DP is Bellman’s principle of optimality (Bellman 1956) which includes a recursive formula (RF) which collects the kind of allowed time modifications. Hence, a global measure of the temporal distortion carried out in the test series, in order to obtain the aligned series, is provided by the area between the resulting optimal path and the identity path (one formed by the following points: \((1,1), (2,2), ..., (N-1,N-1), (N,N)\)), which is the so-called TDI (Temporal Distortion Index, Frias-Paredes et al., 2016). So, this parameter will serve to describe the temporal component of the error and it is expressed as follows:

\[
P_l = \int_{i_l}^{i_{l+1}} \left( x - \frac{(x - i_l) \cdot (j_{l+1} - j_l)}{i_{l+1} - i_l} + j_l \right) dx
\]

\[
TDI = \frac{2 \sum_{l=1}^{k-1} |P_l|}{N^2}
\]

Note that TDI is a dimensionless number varying in the interval \([0,1]\), where \(0\) corresponds with the null temporal distortion and \(1\) with the maximum one.
4. Probabilistic forecast evaluation metrics

4.1 Principles of probabilistic weather forecasts

The atmospheric movements can be described by non-linear differential equations that unfortunately have no analytical solution. The numerical methods to solve them have been developed in different stages. The evolution on numerical weather prediction is a direct consequence of the increase of computer resources, the spatio-temporal high resolution of the models and the improvement in observational networks and assimilation methods. All of this contributed to extend the knowledge on the dynamics and microphysical processes in the atmosphere.

For a long period until the 90ies, the numerical weather prediction (NWP) philosophy was based on the deterministic atmospheric behaviour. That means, given an initial state of the atmosphere, its time evolution can be numerically predicted to give a final state, which is unique. Accordingly to this premise, a deterministic system is one in which the chance is not involved in any future state of the system. As a consequence, a deterministic model will always lead to the same final state from identical initial conditions. Consequently, the efforts of the scientific community on NWP were focused on producing the most accurate forecast (Tracton and Kalnay, 1993).

Lorenz (1963) showed the concepts of Chaos Theory using a simplified model of fluid convection to numerically represent a dynamical system that exhibits most of the properties of other more complex chaotic systems. Lorenz demonstrated that small variations on the model initial conditions (ICs) do not produce a single final solution but a set of different possible solutions. Errors are amplified as the forecast period grows and will evolve into spatial structures shaping the flow of the day. The inherent atmospheric predictability is thus state-dependent and that is why the predictability of the future atmospheric states is also limited in time (Lorenz, 1963 and 1969).

Observational methods, assimilation strategies and the own characteristics of numerical models have inherent limitations that introduce uncertainty in the estimation of the possible future atmospheric states. This uncertainty misleads the forecast and is amplified when the forecast period grows and when the spatio-temporal resolution of the model increases.

Therefore, the atmospheric state cannot be exactly known because the forecast chain always contains errors, which only can be estimated. The inaccurate determination of the real atmospheric state leads to the existence of a set of initial conditions compatible with it. A single model only provides a single solution of all possible future atmospheric states.

The traditional deterministic approach gave way to a new paradigm with richer information than a single solution of the future state of the atmosphere. The new paradigm includes quantitative information about the uncertainty (errors) of the predictive process. The atmospheric non-linear behaviour, consequently chaotic, must be treated now in a probabilistic way by means of the generation of multiple forecasts starting from slightly different but equally probable initial conditions in order to characterize the uncertainty of the prediction (Leith, 1974).

To capture these sources of uncertainty, many operational and scientific centres worldwide produce ensemble forecasts (e.g. NCEP, ECMWF, etc.) since the early of 1990s. The basic idea behind ensemble forecasting is to run multiple (ensemble) forecast integrations from slightly perturbed ICs (ICs forecast error source) coming from multiple models and/or perturbing model formulation (model formulation forecast error source).

The ensemble prediction system (EPS) is a tool for estimating the time evolution of the Probability Density Function (PDF) viewed as an ensemble of individual selected atmospheric states. Each of these initial different states is physically plausible. The spread of the states is representative of the prediction error (Toth and Kalnay, 1997).
If an idealized EPS could be generated that just properly captures all sources of forecast error (uncertainty), then the forecasted PDF would be reliable and skillful, that is, sharper than the climatological PDF. No further information is needed to become trustworthy the forecast-error predictions since a perfect PDF is a complete statement of the actual forecast uncertainty.

These errors are particularly pronounced when dealing with mesoscale forecast of near-surface weather variables leading to large under-dispersion results because of the insufficient ensemble size, inadequate parameterization of sub-grid scale processes and inaccurate knowledge of land-surface boundary conditions (Eckel and Mass, 2005). Even so, real-ensemble forecast distributions often represent a substantial portion of the true forecast uncertainty although they were generated from an incomplete representation of weather forecast error sources.

### 4.2 Parameters needed and their temporal characteristics

The parameters DNI, wind speed, wind direction, wind gust, and ambient temperature remain the same as discussed in section 3.2.

Temporal resolution and temporal availability is also the same as discussed in section 3.3 and 3.4.

### 4.3 Uncertainty representation

In probability weather forecasting, uncertainty is represented with the help of the Probability Density Function (PDF). Let us consider a random variable $x$ that we do not know anything about its nature. The question is whether we can infer something from that variable. If we take $n$ different values of $x$ that belong to the same population, we can obtain an approximation of the PDF of the random variable $x$ when constructing the histogram of these values.

As an example, we could think of the variable $x$ as the mean monthly temperature of April at a surface observation station. Then, the population corresponding to the variable $x$ would be the mean monthly temperatures of April at that station. If we restrict the study period from 1981 to 2010, the $n=30$ values of $x$ would form the sample space.

The PDF gives information about the behaviour of the random variable $x$. Taking the normal or Gaussian PDF of $x$, the analytical formula is represented as follows:

$$PDF(x) = \frac{1}{\sqrt{2\pi \sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$  \hspace{1cm} (17)

where $\sigma$ is the standard deviation and $\mu$ is the mean.

Plotting the PDF of the random variable $x$ as it is shown in Figure 13, it can be noticed there is a value of $\mu$ around which all the random values of $x$ are distributed symmetrically. Likewise, $\sigma$ is a measure of the standard deviation of $x$ from its mean. We can think of $\sigma$ as the mean error or uncertainty we would have if we approximated any possible values of $x$ by $\mu$.

Summarizing, the PDF gives a depiction of all possible values of the random variable $x$ and their associated probabilities of occurrence, resulting this approach in a quantitative way to represent the uncertainty of a random variable.

EPS tries to estimate the uncertainty of the forecast by discretizing the forecast PDF for each model variable at each grid point in $N$ values corresponding to $N$ ensemble members. As an example, Figure 14 shows the forecasted PDF 2m-temperature at forecast length H+60 by the AEMET-SREPS (García-Moya et al., 2011) for the closest grid point to Sevilla (Spain). This ensemble was formed by 25 members, but for this particular case, there was only one ensemble member that did not integrate properly, so $N=24$. It can be noticed the more ensemble members are the higher is the resolution of the PDF.
From the Numerical Weather Prediction (NWP) perspective is computationally unfeasible to integrate the Liouville Equation (also referred as Fokker-Plank Equation) that describes the time evolution of a PDF when random processes are included to account for model error. A practical approximation to resolve this computational problem is to use an Ensemble Prediction System (EPS).

Figure 14: Histogram of 2m-temperature forecasted at H+60 by AEMET-SREPS at the closest grid point to Sevilla (Spain). Valid time: 30th June 2011 at 12:00 UTC.
A simplified approach to represent the uncertainty of the forecast from EPS is the spread (Toth and Kalnay, 1997) measured as the standard deviation $\sigma$ of the PDF. The standard deviation quantifies how much the ensemble members deviate from the mean, on average, and it is often used as a measure of the spread (Wilks, 2006).

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (f_i - \text{em})^2}$$  (18)

Where $f_i$ is the forecasted variable by ensemble member $i$ on each grid point and $\text{em}$ is the ensemble mean, e.g. the mean of the $N$ forecasts. In the aforementioned case study of 2m-temperature forecast (Fig. 14), $\text{em}$ and the spread are 32.3°C and 1.5°C, respectively. The spread measured, as standard deviation is an estimation of the error of the deterministic forecast so that the higher the spread is the higher uncertainty is associated to the forecast.

Ensembles are composed of individual members that are deterministic predictions and allow providing individual deterministic information about the forecast (García-Moya et al., 2011). This information supports the probabilistic interpretation of the EPS when it is considered as an intrinsically probabilistic prediction system.

**Recommendation for the dispatch optimizer developers:**

Alternative A: Use the spread (measured as 1 sigma) of an EPS PDF as uncertainty in the sense of a standard deviation while the forecast is the mean of the ensemble. This needs to be done on individual DNI forecasts, on temperature forecasts and on wind speed forecasts separately.

Alternative B (preferred and agreed on in PreFlexMS): Use all ensemble members and apply this approach on the electric and thermal power parameters further downstream in the optimizer. The latter approach allows to incorporate a further weighting when repeating the optimizer run with more recent information about real weather occurring (e.g. by observations made at the power plant or by satellites).

**Note:** As a recommendation what to do later with the probabilistic optimizer result, we now give information about ensemble mean, ensemble median, spread, and exceedance of probability thresholds or percentiles as alternative use of an EPS.

The ensemble mean ($\text{em}$) forecast is a simple but effective product. The averaging serves as a filter to reduce or remove atmospheric features that varies amongst the members and is therefore likely to be regarded as less predictable at the time. Such non-predictable features are effectively removed from the $\text{em}$. Significant high-impact events are often weakened or absent in the $\text{em}$. Use of probabilities is therefore essential in conjunction with the $\text{em}$.

**Ensemble mean** is most suited to parameters like temperature and pressure, which usually have a rather symmetric Gaussian distribution. It is less suitable for DNI, wind speed and precipitation because of their skewed distributions. For these parameters, the median might be more useful. It is defined as the value of the middle ensemble member, if the members are ordered according to rising (ranked) values. Due to the anti-symmetry of the initial perturbations, the $\text{em}$ is very similar to the Control (or HRES) in the short range, but will differ on larger forecast ranges.

**Ensemble mean** tends to weaken gradients: all members might forecast an intense low-pressure system with 15-20 m/s winds in different positions. These differences in position lead to a rather shallow low in the $\text{em}$, which gives the impression of weak average winds.

The ensemble spread is a measure of the difference between the members and is represented by the standard deviation with respect to the $\text{em}$. On average, small spread indicates high a priori forecast accuracy and large spread low a priori forecast accuracy. The ensemble spread is flow-dependent and varies for different parameters. It usually increases with the forecast range, but there can be cases when the spread is larger at shorter forecast
ranges than at longer. This might happen when the first days are characterized by strong synoptic systems with complex structures but are followed by large-scales ‘fair weather’ high-pressure systems.

The **spread** around the *em* as a measure of a priori accuracy applies only to the *em* forecast error, not to the **median**, the Control or the HRES, even if they happen to lay mid-range within the ensemble.

The ensemble spread should reflect the diversity of all possible outcomes, in particular when the deterministic forecasts are “jumpy”, which might indicate that different weather developments are possible.

**Note:** For several geophysical parameters, it may happen that the spread is not sufficiently large. This applies especially to our case when using the first 48 hours of the ECMWF ENS ensemble, which is optimized for a good spread in the 5 day forecast range. Having sufficient spread may be different for various geophysical parameters. Some may be acceptable, others may have too small or too large spread. Therefore, an assessment has to be made for each parameter under concern. And DNI is not among those the meteorological services assess routinely. Therefore, AEMET is doing this as part of WP4. In case of a spread being too small, a calibration can be performed (section 4.6).

Generally, whether the ensemble spread is small or large, the *em* (or median) will, beyond the short range, exhibit higher accuracy than the Control (and high-resolution forecast). This is particularly true for “dry” parameters, such as mean surface layer pressure (MSLP), temperature and wind speed.

With increasing spread, the predictability of the meteorological situation is very low. Therefore, one should more rely on the use of probabilities. If the spread is low, the deterministic run may be more reliable. This is a rule of thumb. This is particularly true for “wet” parameters, such as precipitation and cloud amounts and therefore DNI. The *em* will also display a higher degree of day-to-day consistency.

Although *em* and, similarly, averages of forecasts from the same or different models provide more accurate and considerably less “jumpy” forecasts than deterministic forecasts, meteorologists are somewhat apprehensive about using them. This reluctance derives mainly from three reasons:

- **Ensemble averages** do not constitute genuine, dynamically three-dimensional representations of the atmosphere.
- Ensemble averages are less able to represent extreme or anomalous weather events.
- Ensemble averages might lead to inconsistencies between different parameters as the averaging is done on each parameter separately. General principles as energy balance or momentum balance may be lost.

To avoid over-interpreting the *em*, it should preferably be presented together with a measure of the ensemble spread or event probabilities; these will convey an impression complementary to the *em*.

The most consistent way to convey forecast uncertainty information is by the probability of the occurrence of an event. The event can be general or user-specific representing the exceedance of a threshold. The event threshold often corresponds to the point at which the user has to take some action to mitigate for the potential damage of a significant weather event.

Probabilities can be instantaneous, such as 10m wind speed probabilities. They can also be calculated over a time interval, for instance precipitation, because the values are themselves originally computed as values accumulated over some time interval.
Note: Comments made for the traditionally more investigated parameter precipitation are most likely relevant for DNI as well.

Probabilities give no indication of the physical reason of the uncertainty. A 25% probability of precipitation > 5 mm/24h might be related to a showery regime or to the uncertainty of the arrival of a frontal rain band. A 25% risk forecast for temperatures < 0°C might be related to the possible early morning clearing of low cloud cover or the possible arrival of artic air. A 25% probability of DNI < 600 Wm⁻² might be related to the development of local convective clouds due to the daily radiative forcing or to the arrival of a frontal system led by synoptic-scale forcing.

Moreover, probability forecasts cannot be linearly extrapolated into the future. Thus, if an event was assigned a 10% probability in the forecast two days ago, 20% yesterday’s forecast and 30% today’s, there is no reason that it will necessarily further increase in tomorrow’s forecast; it could equally well remain at its current level or decrease.

4.4 EPS verification

With verification, we assess the quality and value of forecasts (Murphy, 1993) by using metrics or measures often called scores. Comprehensive descriptions of standard verification methods can be found in Wilks (2006) and in Jolliffe and Stephenson (2012), whereas in Candelil and Talagrand (2005) and in Stensrud and Yussouf (2007) there is a thorough study of probabilistic forecasts, including ensemble forecasts.

Despite their probabilistic interpretation, ensemble forecasts are a priori not probabilistic forecasts, but rather only finite sets of deterministic forecast realizations. To derive probabilistic forecasts from the ensemble members, further assumptions concerning their statistical properties are required. The question as to how the quality of ensemble forecasts can be assessed therefore depends on how they are interpreted, i.e. whether they are seen as finite random samples from underlying forecast distributions, or whether they have been converted into probability distributions and are interpreted probabilistically. For both interpretations specific metrics exist.

Those metrics that are designed for the sample interpretation aim at characterizing the collective behaviour and statistical properties of the ensemble members. They do not require that some form of probability distribution has been derived prior to verification, but often represent the first step towards a probabilistic interpretation to follow. On the other hand, once a probabilistic forecast has been derived from an ensemble, verification is in principle straightforward by applying standard probabilistic skill metrics.

The assessment of the deterministic quality of the ensemble members is a first requirement in the development of an EPS. When the quality of the ensemble members is similar, then any of them can be equally weighted in the computation of a probabilistic forecast, as they are assumed to be equally probable. In addition, the ensemble mean should show a better deterministic performance than any individual member in terms of Root Mean Squared Error (RMSE) as it is shown in Figure 15 (Leith, 1974; Murphy, 1988; Whitaker and Loughe, 1998; Ziehmann, 2000).

One of the most important skill attributes of probabilistic forecasts is reliability, which characterizes the degree to which forecast probabilities are consistent with the relative frequencies of the observed outcomes. This implies that the ensemble members and observed outcomes are consistently sampled from the same underlying probability distributions, i.e. that they are statistically indistinguishable from each other. This property, also referred as ensemble consistency, is commonly assessed through two visual methods: the rank histogram and the spread-error diagram.
Figure 15: Pressure at mean sea level (PMSL), BIAS (lower group) and RMSE (upper group) of each ensemble member (thin continuous line) and ensemble mean (thick continuous line) as the forecast length. Continuous lines representing the operational ensemble members are highlighted with circles (García-Moya et al., 2011).

Figure 16: Examples of rank histograms.
### 4.4.1 Rank histogram

The **rank histogram**, also known as **analysis rank histogram**, **Talagrand diagram** or **binned probability ensemble** (Anderson, 1996; Hamill and Colucci, 1997; Talagrand et al., 1998) is one of the most widely used tests for ensemble reliability (Figure 16). The basis of this test is an analysis of how the observed outcomes rank with respect to the corresponding ensemble members. An integer number according to the position of the observation value in the sorted list of forecast values defines the rank of an observation. Thus, determining the ranks of all observations, and then counting how often each of the possible rank values is taken, and finally displaying these numbers in the form of a histogram construct the rank histogram.

As stated before, if an EPS is reliable, then the ensemble members and observations should be statistically indistinguishable. Consequently, it would then be equally likely for an observation to take any of the rank values and a reliable EPS would, apart from deviations due to sampling variability, yield a flat histogram.

On the other hand, rank histograms significantly deviating from flatness imply that the forecasts are unreliable. In this case, the histogram shape may give hints about the nature of the forecast deficiencies. For instance, if the forecast are subject to a systematic unconditional negative (positive) bias, there is an enhanced probability that an observation exceeds (is exceeded) by the majority of the ensemble members, leading to an overpopulation of the higher (lower) ranks and hence a positively (negatively) sloped histogram. If both the lowest and the highest ranks are overpopulated and the central ranks depleted, this indicates the ensemble is under-dispersive: since the spread is on average too low, there is an enhanced probability that an observation is not captured by the ensemble. Conversely, ensemble forecasts that are over-dispersive capture the observation too often and result in a peaked histogram.

### 4.4.2 Spread-error diagram

The **ensemble spread** (often measured by the standard deviation with respect to the ensemble mean control forecast, if possible) and the **error of the ensemble** (measured by the RMSE with respect to the analysis for either the control forecast or the ensemble mean) should show a linear relationship and a similar growth rate with respect to forecast step (Buizza and Palmer, 1997; Whitaker and Loughe, 1998). An EPS is expected to sample the uncertainties of NWP models (ensemble spread), as well as to give explicit and quantitative information about the predictability of the atmosphere (ensemble error). According to this, an ensemble can be statistically consistent (calibrated) or, on the other hand, can be under-dispersive (quite common in operational ensembles) or over-dispersive (Figure 17).
Resolution and discrimination are two further central attributes of probabilistic prediction skill. Both attributes are related in that they measure whether there is an association in the sense of a correlation between what is predicted and what is observed. Resolution takes the perspective of the forecasts, quantifying the degree to which the observed outcomes change as the forecasts change. Discrimination takes the perspective of the observations, asking whether different observed outcomes could be correctly discriminated by the forecasts. Resolution and discrimination of ensemble forecasts are typically assessed by means of standard probabilistic approaches, requiring that the ensemble members, which are considered as finite random samples have been converted into probability distributions and interpreted probabilistically prior to verification.

Once a probabilistic forecast has been derived from an ensemble, standard probabilistic scores can be applied. The choice of the score is thereby mainly related by the prediction context and by the skill attribute of interest. Common summary measures of ensemble prediction skill are the Brier Score (BS) for binary events; the Ranked Probability Score (RPS) for categorical events and the Continuous Ranked Probability Score (CRPS) for outcomes measured on a continuous scale. Here, a short summary of these most frequently used scores in the context of probabilistic ensembles is provided. A more recent score, the user-oriented quantile score (QS), is also explained.

### 4.4.3 Brier Score

For a binary outcome (event and not event), the Brier Score (BS) is given by (Candille and Talagrand, 2005; Jolliffe and Stephenson, 2012):

$$ BS = \frac{1}{n} \sum_{t=1}^{n} (f_t - o_t)^2 $$  \hspace{1cm} (19)
Being \( f_i \) the probability assigned to the event by the \( t^{th} \) forecast, and defining \( o_i = 1 \) (respectively \( o_i = 0 \)) if the \( t^{th} \) observation corresponds to an event (respectively, non-event). The BS is often formulated as skill score by relating it to the score obtained from a reference forecast strategy. If climatology is used as a reference, and if \( c \) is the climatological probability of the event, the Brier Skill Score (BSS) is given by:

\[
BSS = 1 - \frac{BS_{cl}}{BS_{cl}} = 1 - \frac{1}{n} \sum_{i=1}^{n} (c - o_i)^2
\]  

(20)

Positive values of the BSS indicate forecast benefit with respect to the climatological forecast. The BS can be decomposed into three components measuring reliability, resolution and uncertainty (expectation value of the score if climatological was used as a forecasting strategy):

\[
BS = \frac{1}{N} \sum_{k=1}^{K} n_k(f_k - \bar{o}_k)^2 - \frac{1}{N} \sum_{k=1}^{K} n_k(\bar{o}_k - c)^2 + c(1-c) = BS_{rel} - BS_{res} + BS_{unc}
\]  

(21)

With \( N \) being the total number of forecasts issued, \( K \) the number of unique forecasts issued, \( n_k \) the number of forecasts with the same probability category and \( o_k \) the observed frequency, given forecasts of probability \( f_k \).

The reliability term \( BS_{rel} \) measures the difference between the probabilistic forecasts \( f \) and the mean observation associated over all of the forecasts; the resolution term \( BS_{res} \) measures how much the conditional probabilities given the different forecasts differ from the climatic average; the uncertainty term \( BS_{unc} \) measures the inherent variability in the event.

### 4.4.4 Ranked Probability Score

The Ranked Probability Score (RPS) can be interpreted as a multi-categorical generalization of the BS (Candille and Talagrand, 2005; Jolliffe and Stephenson, 2012). It is applicable to discrete probabilistic forecasts issued for categorical events. In the following, \( K \) denotes the number of categories; \( c_k \) denotes the climatological probability that the observed outcome is in category \( k \). For a set of \( n \) forecast-observation pairs, \( f_{i,k} \) is the probability assigned by the \( i \)th forecast to the \( k \)th category. Further, we define \( o_{i,k} = 1 \) if the observation is in category \( k \) and \( o_{i,k} = 0 \) otherwise. Finally, let \( \bar{F}_{i,k} \) and \( O_{i,k} \) denote the \( k \)th component of the \( i \)th cumulative forecast and observation vectors, and \( C_k \) the \( k \)th category of the cumulative climate distribution:

\[
\bar{F}_{i,k} = \sum_{t=1}^{k} f_{i,t} \quad ; \quad O_{i,k} = \sum_{t=1}^{k} o_{i,t} \quad ; \quad C_k = \sum_{t=1}^{k} c_t
\]  

(22)

Using this notation, the RPS is given by:

\[
RPS = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} (\bar{F}_{i,k} - O_{i,k})^2
\]  

(23)

Like the BS, the RPS is often formulated as skill score to assess forecast benefit with respect to a reference strategy. If climatology is used as a reference, the ranked probability skill score (RPSS) is given by:

\[
RPSS = 1 - \frac{RPS_{cl}}{RPS_{cl}} \quad ; \quad RPS_{cl} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} (C_k - O_{i,k})^2
\]  

(24)

Note that also for RPS formulations exists of decomposition into three terms representing reliability, resolution and uncertainty (Candille and Talagrand, 2005).
4.4.5 Continuous Ranked Probability Score

The continuous ranked probability score (CRPS) is defined as the integrated squared difference between the cumulative forecast and observation distributions. It is defined on a continuous scale and therefore does not require reduction of ensemble forecasts to discrete probabilities of binary or categorical events. It can be interpreted as an integral over all possible BS values and also allows decomposition into three terms representing reliability, resolution and uncertainty (Candille and Talagrand, 2005). Furthermore, the deterministic limit of the CRPS is identical to the mean absolute error and thus has a clear interpretation.

4.4.6 Quantile score

The quantile function assigns to each probability \( \tau \) attained by a certain probability density function a value \( f_\tau \) defined by:

\[
f_\tau = \{ x : Pr(X \leq x) = \tau \}
\] (25)

It is the inverse function of the cumulative distribution function.

The quantile score (QS), is a measure derived from this function which gives information about the quality of a quantile forecast. It is based on an asymmetric piecewise linear function, the check function (Koenker and Bassett, 1978):

\[
\rho_\tau(u) = u[\tau - I(u < 0)] = \begin{cases} 
\tau u & \text{if } u \geq 0 \\
(\tau - 1)u & \text{if } u < 0 
\end{cases}
\] (26)

where \( I(u<0) \) is an indicator function having value 1 if the condition in parenthesis is true, and zero otherwise. QS results of taking the mean of the check function applied to the difference between the observation and a quantile forecast:

\[
QS = \frac{1}{n} \sum_{i=1}^{n} \rho_\tau(o_i - f_{\tau,i})
\] (27)

This score is specially indicated when the loss function is asymmetric (the cost of overestimating and underestimating solar power is not the same; Bouallegue, 2015).

It enables to define a relative user characteristic (RUC) curve, equivalent to the event-oriented relative operating characteristic (ROC) curve (Bouallegue, 2015).

The traditional ROC curve plots the hit rate (proportion of occurrences correctly forecast) versus the false alarm rate (proportion of non-occurrences incorrectly forecast) for a specific event with a base rate, using a set of probability thresholds to make the decision. Different decision makers will have different optimal strategies and use different probability thresholds to determine which action to take. For example the events “2m temperature higher than 0 C” and “2m temperature higher than 20 C” are associated to two different base rates, and we can calculate from our ensemble forecast the hit rate and false alarm rate for different probabilities and plot the ROC curves (Fig. 18). For some users low probabilities (plotted at the top right corner of the curve) will be more meaningful, for others high probabilities (plotted at the bottom left corner) will. If we want to characterise one particular user’s behaviour, we will need to draw many ROC curves for all possible events and select for each one the point most suitable to the user.
Conversely, the RUC curve plots the hit rate versus the false alarm rate for a specific user defined by a cost-loss ratio, using a set of events with different base rates to plot the curve. It is appropriate for user-focused situations where the decision criterion is known for a range of events. For example, a user with a high cost-loss ratio will only consider high quantile forecasts significant. We can draw the RUC curve for that quantile, plotting the hit rate versus the false alarm rate for events with varying base rates. Events with a low base rate have low hit and false alarm rates, and appear at the bottom left corner of the graph. As with the ROC curve, forecasts with no skill follow the diagonal line and perfect forecasts consist of two lines coincident with the left and the upper boundaries.

Both ROC and RUC diagrams give information about potential skill: they are not sensitive to bias, so it is possible to improve the forecast through calibration. In both diagrams the area under the curve is commonly used as a measure of discrimination (for probability and quantile forecasts, respectively).
4.5 **Statistical post-processing for calibrating the EPS**

The main objective of statistical post-processing techniques consists in perform a calibration using the former prediction’s skill to correct the current probabilistic forecast.

Different methodologies have been proposed to build calibrated probabilistic forecasts from ensembles, including Bayesian Model Averaging (Raftery et al., 2005), Logistic Regression (Hamill et al., 2008) and Extended Logistic Regression (Wilks, 2009).

### 4.5.1 Bayesian Model Averaging

Bayesian Model Averaging (BMA) is a statistical post-processing method that generates calibrated and sharp predictive PDFs from EPS (Raftery et al., 2005). The BMA predictive PDF of a weather variable is a weighted average of PDFs centred on the individual bias-corrected forecasts. The weights reproduce the predictive skill of the ensemble member over a training period.

BMA can be applied when forecasts error are approximately Gaussian distributed (e.g. surface temperature or sea level pressure). For non-Gaussian distributions the use of a mixture of skewed PDFs allows to extend the BMA methodology (e.g. a combination of point mass zero and a power-transformed gamma distribution can be applied to quantitative precipitation (Sloughter et al., 2007)).

The BMA predictive PDF is a summation of weighted PDFs of each individual ensemble member (Kass and Raftery, 1995; Hoeting et al., 1999):

\[
PDF(o|f_1, ..., f_m; \theta_1, ..., \theta_m) = \sum_{i=1}^{m} w_i PDF_i(o_i|f_i, \theta_i)
\]  

Where \( f_i \) is the ensemble member forecast, \( o \) represents the forecasted variable and \( \theta_i \) are the characteristic parameters of the \( i \)th individual PDF from the \( i \)th ensemble member. Each of the individual PDFs associated to each ensemble member is weighted based on its relative performance during a training period. The weights \( w_i \) are probabilities, i.e. non-negative and add up to 1. Both the weights \( w_i \) and the parameters \( \theta_i \) are estimated by maximum likelihood (Wilks, 2006) using the training data.

### 4.5.2 Logistic Regression and Extended Logistic Regression

The logistic regression (LR) and extended logistic regression (ELR) techniques are described in detail in Wilks (2009). LR approximates the Cumulative Distribution Function (CDF) of the predicted parameter \( o \) by the following equation:

\[
CDF(q) = PDF(o \leq q) = \frac{e^{f(x)}}{1+e^{f(x)}}
\]  

Where \( q \) is a selected prediction threshold and:

\[
f(x) = b_0 + b_1 x_1 + ... + b_n x_n
\]
Being \( \{x_1, \ldots, x_n\} \) the regression predictors and \( \theta = \{b_0, b_1, \ldots, b_n\} \) the unknown coefficients to be estimated during the training process. The name logistic comes to the fact that the regression equation is linear on the logistic scale:

\[
\ln \left[ \frac{CDF(q)}{1-CDF(q)} \right] = f(x) \tag{31}
\]

Typical predictors for LR when calibrating ensemble predictions are ensemble mean, ensemble spread or a function of them (Hamill et al., 2008). As thresholds \( q \), using the representative climatological quantiles of the meteorological parameter \( y \) ensures a statistical uniformity in the process of regression. According to Wilks (2009) \( \theta \) coefficients are generally estimated using maximum likelihood (Wilks, 2006), but other estimation techniques could give better performance, for example the minimization of the continuous ranked probability score (Hersbach, 2000).

By construction Equation (31) is fitted separately for every threshold \( q \) and this fact involves several problems. We consider for example the parameter precipitation and two thresholds, \( q_1 = 2mm \) and \( q_2 = 10mm \). After the training we have two different regression equations for each threshold, \( f_1(x) \) and \( f_2(x) \), which in general are not parallel. The non-parallelism of the functions implies that for some values of the predictors \( \{x_1, \ldots, x_n\} \) these curves will cross leading to the unrealistic result of \( CDF(q_1) > CDF(q_2) \). Another problem arises when we want to estimate the CDF of a threshold for which regressions have not been fitted. This process requires some kind of interpolation of CDFs, which is not statistically coherent. Finally, the more equations are to be fitted the more unknowns have to be estimated.

To overcome these problems, Wilks (2009) proposed a new approach that consists of including a function \( g(q) \) in the exponent which increases with threshold \( q \):

\[
CDF(q) = PDF(o \leq q) = \frac{e^{f(x)+g(q)}}{1+e^{f(x)+g(q)}} \tag{32}
\]

Thus, unique regression estimation for any value of \( q \) is needed, which implies the parallelism of the functions \( f(x) \) for the different thresholds (the unknowns \( \{b_0, b_1, \ldots, b_n\} \) are always the same). This approximation is known as Extended Logistic Regression (ELR).

A relevant feature of LR (and ELR) is that it has no statistical restriction to be used with non-Gaussian parameters such as precipitation or wind. At the same time this technique can be applied to ensembles whose members are non-distinguishable.
5. References


ECMWF, as obtained in 2016 from https://software.ecmwf.int/wiki/display/IFS/CY41R1+Official+IFS+Documentation?preview=/48104914/48234772/IFSPart4.pdf


