PUBLIC CONSULTATION

Baseline methodology assessment

September 24 2021

Contents

Executive summary

In recent years, Elia has developed baseline methodologies for the participation of different technologies such as active demand response, distributed storage and distributed generation to different products. These baseline methodologies are required to determine the volume of flexibility actually delivered and used as part of the activation control and/or the Transfer of Energy mechanism. Currently, the following baseline methodologies are available:

- Last QH and High X of Y for mFRR;
- High X of Y for strategic reserves;
- High X of Y* for participation to the day-ahead and intraday markets via the Transfer of Energy mechanism (ToE DA/ID);
- Declarative baseline methodology for aFRR, in which the BSP needs to determine and submit the baseline at the latest one minute ahead of real time.

In this study, Elia analyzes the performance of the baseline methodologies currently in place and the possible opportunities for improving the existing baseline methodologies or introducing new baseline methodologies.

Performance of baseline methodologies

The first part of the study focuses on the performance of different baseline methodologies for the different products under consideration (i.e., mFRR, ToE DA/ID, the CRM and strategic reserves)**.** To this end, the return of experience from market parties active in Belgium is gathered and an analysis of international practices, a literature review and an assessment of the performance of different baseline methodologies has been performed. In these different steps, the characteristics of the products that are relevant for the choice and design of baseline methodology have been taken into account (notably, the typical activation duration, the time between the activation request/decision and the start of the activation, and the frequency of activation).

The main conclusion from this first part is that the current baseline methodologies applied for the different products under consideration are in line with best practices:

- mFRR: both Last QH and High X of Y combine a good performance in terms of simplicity and accuracy for a wide range of assets and therefore achieve the best overall performance. In addition, the analysis of international practices reveals that these types of baseline methodologies are used most commonly for mFRR or similar products.
- ToE DA/ID / CRM / strategic reserves: The High X of Y(*) baseline methodology currently applied combines simplicity with a good performance in terms of accuracy and therefore achieves the best overall performance. The literature furthermore highlights that this baseline methodology can achieve a high accuracy for a large variety of assets, also during periods with the highest loads/prices for which activations of these products is most likely. In addition, the analysis of international practices reveals that this type of baseline methodology (generally referred to as historical baseline methodology) is used most commonly for these types of products.

This conclusion is also reflected in the fact that market parties indicated to be generally satisfied with the current baseline methodologies.

Regarding historical baseline methodologies such as High X of Y(*), it can be observed that different variants exist, differing mainly in the selection of the representative days (e.g., the parameters X and Y) and whether or not a so-called same-day adjustment is applied (i.e., whether or not the baseline is calibrated to the offtake/injection measured shortly before the start of the activation). In this regard, different quantitative assessments can be found in the literature, all concluding that applying a same-day adjustment significantly increases the accuracy, whereas other parameters (such as the values of X and Y) have a minor impact on the overall performance (in particular in case a same-day adjustment is applied). Considering that a same-day adjustment is already (allowed to be) applied, there are no indications that the current baseline methodologies could be (significantly) improved.

Opportunities for alternative baseline methodologies

Although market parties were positive regarding the existing baseline methodologies, they did express concerns that the current baseline methodologies might not be sufficiently accurate for all types of assets they might want to add to their pool in the future, and in particular variable renewable generation. The literature and the performance assessment in this regard confirm that MBMA baseline methodologies (e.g., Last QH) and historical baseline methodologies (e.g., High X of Y(*)) might not be sufficiently accurate for a minority of assets that have a (highly) variable and irregular offtake/injection pattern, such as wind and solar PV generation, but also variably operated units or demand facilities.

The second part of the study therefore looks in more detail in the need and opportunities for introducing new baseline methodologies to further remove barriers for participation for assets with a (highly) variable and irregular offtake/injection pattern. The assessment in this regard indicates that declarative baseline methodologies, in which the FSP can determine its own baseline and submit it to the TSO, are the most inclusive as they enable full flexibility to FSPs to select and use a suitable methodology for calculating their baseline. In addition, calculated baseline methodologies, in which the baseline is calculated based on real-time external data, could form an alternative for variable renewable generation (e.g., calculation of the active power a wind farm would have generated in case no downward activation would have taken place based on the wind speeds measured in real time).

Elia acknowledges that adding a declarative baseline methodology may reduce barriers for participation for assets with a highly variable offtake/injection pattern for which the current baseline methodologies might not be sufficiently accurate as well as more generally to enable more flexibility to FSPs to use a suitable method for calculating their baseline. For these reasons, Elia is willing to consider developing a declarative baseline methodology for the products ToE DA/ID, CRM and mFRR in case there are clear indications of the need and added value. In this regard, although adding a declarative baseline methodology has certain potential/theoretical benefits, Elia questions the current need for and added value of introducing a declarative baseline methodology for these products. Therefore, Elia invites market parties to provide indications and specific use cases for which a declarative baseline methodology would be needed to reduce entry barriers.

In case an implementation would be targeted, the proposed design of the declarative baseline methodology differs for mFRR on the one hand, and the products ToE DA/ID / CRM / strategic reserves on the other hand. This because for mFRR, requests for activation happen close to real time and are difficult to anticipate, whereas for ToE DA/ID / CRM / strategic reserves, the request or decision to activate can happen a significant time in advance, therefore possibly providing opportunities to manipulate the baseline during activations. In this regard, Elia proposes that for mFRR the declarative baseline would need to be submitted shortly before the activation request (e.g., 45 minutes before the start of the quarter hour), and that a monitoring of the bias (i.e., systematic deviations of the baseline) in moments

where no activation has taken place would be sufficient to remove the opportunities for manipulating the baseline. In contrast, for ToE DA/ID / CRM / strategic reserves, Elia proposes that the baseline needs to be submitted minimally two days prior to the start of the activation. In addition, for these products Elia proposes to monitor systematic deviations and to request transparency of the process used by the FSP for calculating the baseline, including the inputs used for the calculation. Finally, both for mFRR and for ToE DA/ID / CRM / strategic reserves, Elia proposes to foresee a validation/prequalification of the baseline before an FSP can make use of the declarative baseline methodology. Such a baseline prequalification would imply that, during a test period, the baseline submitted by the FSP needs to be more accurate than the default baseline methodology(y)(ies).

Regarding the challenge of baselining for wind and solar PV, which is considered mainly relevant for mFRR and aFRR, the study concludes that for mFRR, a declarative baseline methodology is deemed sufficient to enable the participation of wind and solar PV. This because it is considered that short-term forecasts for a product with a 15-minute resolution can be sufficiently accurate.

In contrast, for aFRR, the combination of the high resolution of the product (i.e., the baseline needs to be submitted for every 4-second interval) and the short-term fluctuations of wind power (and possibly solar PV) output can make it difficult to submit an accurate baseline one minute in advance, as needed for the current declarative baseline methodology for aFRR. Therefore, if the need is confirmed, Elia is prepared to enable BSPs to calculate their baseline based on real-time data and to submit the baseline in real-time under the following conditions (in addition, the criteria and monitoring applied for the current declarative baseline would also apply for a baseline submitted in real time):

- The BSP provides a sound justification indicating why there is no viable way to submit a baseline one minute in advance while meeting the accuracy requirements and why submission of the baseline in real time would allow achieving a sufficient accuracy.
- The BSP provides a clear description of the method and inputs used to calculate the baseline and can make these inputs available to Elia upon request. The calculated baseline should be traceable to these different inputs.
- The BSP can prove that the calculation of the baseline is fully independent of the aFRR reference set point and independent of the operating conditions (e.g., curtailed or non-curtailed conditions)[1].

In order ensure that a possible implementation would effectively be used, and considering that the above conditions would in any case need to be met before a BSP would be allowed to submit his baseline in real time, Elia invites stakeholders to provide indications on the volumes for which the current baseline is considered to form a barrier for participation and to provide indications of the feasibility of meeting the proposed criteria applying for a baseline that can be submitted in real time.

[1] For instance, there are different methodologies that are being used to calculate the power a wind farm would have generated in case it would not be activated downwards/curtailed (e.g., to provide downward reserves). Some of these methodologies are accurate under normal conditions but tend to (significantly) overestimate the potential power in periods of downward activations as a result of not considering wake effects. In this regard, it is important that the calculation of the baseline is accurate regardless of the operating conditions.

Terminology

Automatic Frequency Restoration Reserves or **aFRR**	As defined in Article 3(99) of the SOGL.
Balancing Services	As defined in Article 2(3) of the EBGL.
Balancing Service Provider or **BSP**	As defined in Article 2(6) of the EBGL.
Balance Responsible Party or **BRP**	As defined in the EBGL.
CRM Functioning Rules	Functioning rules for the capacity remuneration mechanism in accordance with Articles 7undecies, § 12, of the law of 29 April 1999 concerning the organization of the electricity market.
DA	Day-ahead
Delivery Point	A point on an electricity grid or within the electrical facilities of a Grid User where a volume of flexibility is delivered.
Delivery Point DP_{PG} or **DP_{PG}**	Delivery Point for which ELIA does not receive daily schedules.
Delivery Point DP_{SU} or **DP_{SU}**	Delivery Point for which ELIA receives daily schedules (in MW), in accordance with the T&C Scheduling Agent.
EBGL	Commission Regulation (EU) 2017/2195 of 23 November 2017 establishing a guideline on electricity balancing.
Frequency Restoration Reserves or **FRR**	As defined in Article 3(7) of the SOGL.
Frequency Restoration Reserves with Manual Activation or **mFRR**	Frequency Restoration Reserves (FRR) that can be activated manually.
Flexibility Service Provider or **FSP**	As defined in Article 2, 64° of the Electricity Act.
ID	intraday
Rules for the Organization of the Transfer of Energy or **ToE Rules**	The set of rules, as defined by Article 19bis §2 of the Electricity Act, proposed by Elia and approved by the CREG, that lay down the principles for Transfer of Energy.
SOGL	Commission Regulation (EU) 2017/1485 of August 2nd, 2017, establishing a guideline on electricity transmission system operation.
Strategic Demand Reserve or **SDR**	As defined in section 2 of the Strategic Reserve Functioning Rules.
Strategic Reserve Functioning Rules	Rules governing the functioning of the strategic reserve established by Elia and, following consultation of Grid Users, approved by CREG and published on Elia's website pursuant to Article 7 of the Electricity Act.

Supplier	As defined in Article 2 15°bis of the Electricity Act.
T&C Scheduling Agent or T&C SA	Terms and Conditions for scheduling agents pursuant to pursuant to Article 46, Article 49 and Article 52 of SOGL and Article 249 of Federal Grid Code.
Transfer of Energy	As defined in Article 19bis section 2 of the Electricity Act.

1. Introduction

In recent years, Elia has enabled market access for different technologies including active demand response, distributed storage and distributed generation. In this regard, Elia has developed baseline methodologies, i.e., methodologies to estimate the offtake/injection at a Delivery Point if no activation would have taken place. These **baseline methodologies are used to calculate the volume of flexibility that is effectively delivered** at a certain Delivery Point, and are used as part of:

- the activation and availability control of the service (if applicable); and
- the Transfer of Energy mechanism (if applicable).

Currently, the **following baseline methodologies are available for the different products**:

- Last QH and High X of Y for mFRR;
- High X of Y for strategic reserves;
- High X of Y* for participation to the day-ahead and intraday markets via the Transfer of Energy mechanism;
- Declarative baseline methodology for aFRR, in which the BSP needs to determine and submit the baseline at the latest one minute ahead of real time.

The current baseline methodologies have been chosen in a pragmatic way and frequently based on suggestions of stakeholders, with the idea that these baseline methodologies could be reviewed in a later stage.

In this context, the aim of this study is to analyze the performance of the baseline methodologies currently in place and to analyze possible opportunities for improving the existing baseline methodologies or introducing new baseline methodologies. The products in scope are restricted to mFRR, strategic reserves, the Transfer of Energy mechanism for day-ahead and intraday markets (ToE DA/ID) and the CRM, but findings and recommendations applicable to aFRR are also considered.

The **remainder of this document is structured in three parts**. First, **Part A assesses the performance of different baseline methodologies**. This part consists of the following sections:

- **Section 2** analyzes the **best practices** with respect to baselining. This section consists of:
  - an overview of the different baseline methodologies encountered;
  - an assessment of international practices with respect to baselining;
  - a literature review;
  - a return of experience based on discussions organized with market parties active in Belgium.
- **Section 3** presents the methodology and results of the **assessment of the performance** of the different types of baseline methodologies for each of the products in scope. Sections 2 and 3 have been developed with the support of the consulting firm DNV.

- **Section 4** provides an overview of the **main conclusions with respect to the baseline methodologies currently applied in Belgium**.

Next, **Part B** looks in more detail in the **need and opportunities for introducing new baseline methodologies**. This part consists of the following sections:

- **Section 5** analyzes the possibilities for market parties to choose their own baseline methodology.
- **Section 6** analyzes the options for baselining for wind and solar PV.

Finally, **Part C** contains **an analysis of the implementation impacts and summarizes the main conclusions**:

- **Section 7** contains the **impact assessment** of the recommendations.
- **Section 8** provides an overview of the **conclusions and recommendations** of the study.

The present document is the first version of a **report** that will be submitted to the CREG **on December 23, 2021**. As such, the present public consultation is an opportunity to **collect stakeholders' views** on the conclusions and recommendations of the study. The stakeholders' feedback will be considered for the finalization of the study.

Part A: Baseline methodology performance

2. Best practices

This section aims to assess the best practices with respect to baselining. To this end, an overview of the different baseline methodologies encountered internationally or in the literature is first provided. Next, a detailed assessment of the international practices is presented, followed by a review of the literature. Finally, an overview is provided of the return of experience from market parties active in Belgium. This section has been developed with the support of the consulting firm DNV.

2.1 Overview of baseline methodologies

Baseline methodologies can be classified in different groups depending on characteristics such as the type of data that is used and the method governing how the selected data is processed to calculate a baseline for a given period.

Table 1 provides an overview of different types of baseline methodologies that are used in different systems across the globe and/or known from the literature.

Table 1: Overview of baseline methodologies

Methodology group	Example	Characteristics
Meter Before - Meter After (MBMA)	Last QH used by Elia for mFRR	MBMBA baseline methodologies calculate the baseline by taking a single meter reading before the period of activation or by taking the average/median/maximum/minimum value of several meter readings before the period of activation. As such, MBMA baseline methodologies result In baselines that are constant during the entire period of activation. The terminology Meter Before - Meter After originates from the fact that meter readings during the period of activation ("Meter After") are compared against meter readings prior to the activation ("Meter Before") to calculate the delivered flexibility. MBMA baseline methodologies are widely used for accurately estimating the level of service delivered under real-time dispatch conditions and short activation periods. It is also a preferred baseline for services with frequent activations as it does not require the use of significant amounts of undistorted historical data. For these reasons, MBMA baseline methodologies are very common in balancing products. MBMA baseline methodologies are most suitable for Delivery Points with relatively flat offtake/injection profiles during the activation period.
Historical baseline methodology	High X of Y used by Elia for mFRR and/or High X of Y* used by Elia for	Historical baseline methodologies make use of historical offtake/injection measurement data (usually recent data of several days prior to the day of the activation) to calculate the baseline for the period of activation. In contrast to the MBMA baseline methodologies, historical baseline methodologies provide baselines that are not constant during the period of activation.

	ToE DA/ID and the CRM	There are many different variants of historical baseline methodologies. A first important aspect in this regard is the data selection and exclusion rules that determine which data will be used to calculate the baseline. Historical baseline methodologies usually use data of 5 to 10 previous days. In terms of data exclusion rules, historical baselines methodologies typically exclude prior-event days (i.e., days in which activations of flexibility have taken place) and non-similar days (e.g. excluding weekends and holidays when determining the baseline for an activation that takes place on a week day, and vice versa). Moreover, many historical baselines exclude additional days based on the offtake/injection characteristics. For example, the methodology may drop days with the lowest offtake if they fall below a threshold related to the mean of the selected days (e.g. the low load day is less than 20% of the mean load during the selected days during dispatch window hours). Other baselines rank the included days based on load and exclude a subset of those chosen days. This could be either extreme days at both ends or the day(s) with the lowest load. For example, a "mid 8 of 10" historical baseline methodology excludes the days with the highest and lowest load out of the 10 selected days, and a "high 4 of 5" historical baseline methodology excludes the day with the lowest load. These exclusions are designed to target the recent days that are most likely to accurately represent the load on the day of the activation. A second aspect relates to the calculation of the baseline based on the selected days. Historical baselines are predominantly calculated using a simple mean for each interval (e.g., quarter hour) of the period of activation across the final set of chosen days. A median approach or a weighted moving average[2] are also applied for some historical baselines. Finally, historical baseline methodologies frequently make use of a same-day adjustment to calibrate the baseline to the measured offtake during the day of the activation. Different approaches for performing a same-day adjustment can be found, and are described below this table.
Declarative baseline methodology (also known as "nomination baseline methodology")	Baseline used by Elia for aFRR / Nomination baseline used In France for NEBEF	Declarative baseline methodologies rely on the estimation, as provided by the Flexibility Service Provider (FSP) to the TSO, of the offtake or injection pattern of the asset or portfolio if no flexibility activation would take place. This forecast is sent before gate closure or at another predefined deadline. In general, the choice of (a) method(s) used by the FSP to calculate the baseline is left to the discretion of the FSP. Since FSPs can apply different, technology-specific, forecasting methodologies, the declarative baseline methodology can support different technologies, provided a reasonable forecast can be produced at the time of declaration.
Regression-based baseline methodology	Regression-based baseline methodology used in the ERCOT system	Regression-based baseline methodologies use a regression model and historical data to estimate the offtake/injection during the period of activation. Regression-based baseline methodologies can use, among others, schedule, weather and other variables to forecast offtake/injection profiles during the period of activation. Regression-based baseline methodologies require substantially more data in comparison to MBMA and historical baseline methodologies. This because more types of data are used

[2] Using a weighted moving average (WMA) puts more emphasis on more recent days' loads by assigning larger weights to these days. KPX (Korea) uses this type of estimation method, among others.

		(i.e., not only measured offtake/injection) and larger sets of data are needed to train the regression model. Most regression-based baseline methodologies require at last a full year of data. In addition, regression-based baseline methodologies are substantially more complex in terms of the calculation of the baseline. For example, a regression-based baseline methodology may include calendar, weather and daylight variables in multiple forms. The specification controls for heat build-up over days, heat gain within the day, hour of light and fraction of dark as well as a range of temperature-time interactions. The regression uses such specification to calculate a baseline that reflects the calendar, weather and daylight characteristics of the period of activation. The regression produces both weekday and weekend baselines.
Calculated baseline methodology	The baseline that can be used for aFRR In the Netherlands	Calculated baseline methodologies involve a calculation of the baseline based on external parameters, such as weather conditions, typically without relying directly on historical data of measured offtake/injection. Calculated baseline methodologies are not commonly applied as they are applicable only to technologies and assets for which the offtake/injection profile can be calculated based on external parameters. Some examples include the estimation of the potential generation of a wind turbine/wind park based on measured wind speed and power curves or the estimation of the potential generation of a solar PV farm based on solar panels characteristics (peak capacity, orientation), and measured solar irradiance and temperature.
Control group baseline methodology (also known as peer group baseline methodology)	Control group baseline methodologies used in the ERCOT system	Control group baseline methodologies determine the baseline by taking the average/median of measurements of the offtake/injection of similar customers/assets that do not participate in the flexibility service.

Same-Day Adjustment Method and Period

In addition to the baseline methodologies presented in Table 1, different baseline methodologies (particularly historical baseline methodologies) can also include a so-called same-day adjustment (SDA). This refers to an adjustment made to the initial calculation of the baseline in order to calibrate the baseline to the measured load/generation on the day of the activation. The calculation of the same-day adjustment is based on data and information of the adjustment period, which is typically a period prior to the start of the activation. The result is that the baseline is better calibrated to the actual (or counterfactual) load/injection at the start of the activation.

Figure 1 illustrates how this mechanism works in practice. In this figure, a customer with a weather-sensitive load profile is shown with the meter data displayed in green. The baseline calculated with a historical baseline methodology without a SDA (in this example a mid 3 of 10 historical baseline methodology) is displayed in blue. As can be seen, in this illustration, the baseline without SDA is constantly below the measured values. During the activation period, the baseline is adjusted according to the difference between the historical baseline without SDA and the measured offtake during the adjustment period. This results in the adjusted baseline displayed in pink, which, based on a visual inspection, approaches the counterfactual more closely.

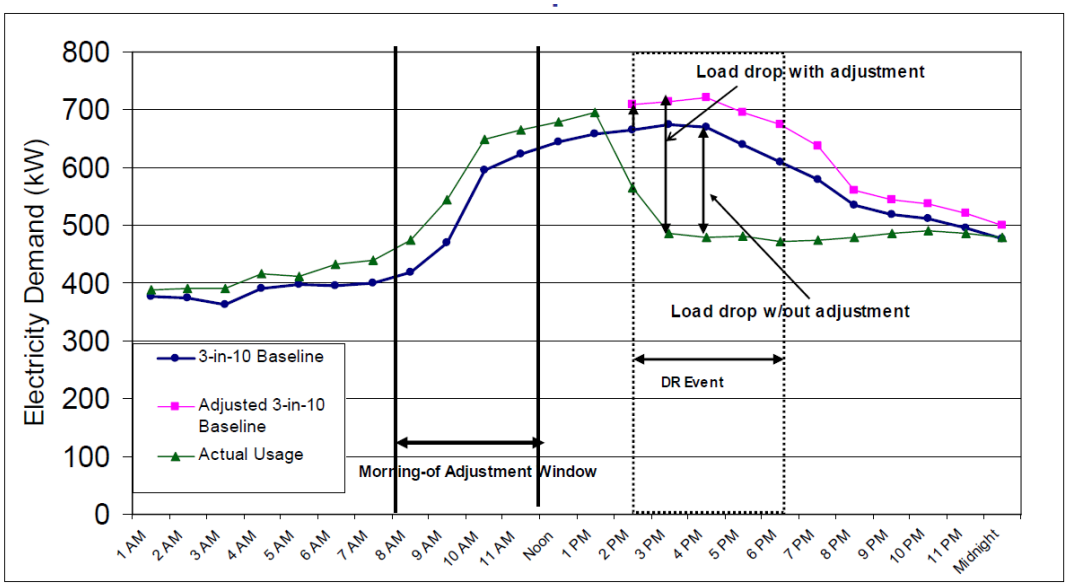

Figure 1: Illustration of Same Day Adjustment mechanism for a hypothetical weather-sensitive load profile. Reference: CAISO (https://www.caiso.com/Documents/Presentation-Baselines_RetailDemandResponsePrograms.pdf)

The literature review (see Section 2.3) suggests that same-day adjustments improve the accuracy of baselines. However, same-day adjustments also carry a degree of risk because the adjustment calculation typically relies on data from a relatively short time period. This implies first of all that SDAs can be more susceptible to gaming as load variations during the adjustment period can have a significant effect on the calculated baseline. Second, situations can occur that may also affect the baseline in an undesirable way. As an example, pre-heating can be applied on electric boilers to compensate the impact of the later activation. When pre-heating occurs during the adjustment period, the activated volume will be over-estimated. For these reasons, the adjustment period is typically a period before the activation request or a longer time interval before the activation period.

Adjustments are made either using an additive or scalar (multiplicative) approach. Both approaches have the effect of calibrating the baseline. However, the adjustments differ during the period of activation dependent on the approach used:

- In case of an additive adjustment, the baseline is shifted by the same amount for the entire activation period. For example, the average difference between the measurement and the baseline value (before applying the SDA) during the adjustment period is added to the baseline calculated without the SDA.

- For the scalar adjustment, the baseline calculated without the SDA is multiplied by a constant factor for the entire activation period. For example, the average ratio of the measurement during the adjustment period and the baseline value (before applying the adjustment) is used as a multiplication factor (also known as proportional adjustment coefficient). As a result, the value of the adjustment in nominal terms is dependent on the nominal value of the original baseline.

One other element of the SDA is whether adjustments are applied symmetrically or asymmetrically. The symmetric approach adjusts the baseline in both directions whereas the asymmetric approach only adjusts the baseline upwards (i.e., increase of the offtake/decrease of the injection).

2.2 International practices

This section assesses the international practices with respect to baselining in order to draw conclusions on common practices internationally and how these practices compare against the current set of baseline methodologies used for the different products in the Belgian system.

The overview of common practices is based on an assessment performed by DNV based on data from 9 countries (namely, GB, France, Finland, the Netherlands, Switzerland, Belgium, USA, Canada and Australia), 16 different TSOs and 94 different products. The assessment covers products and baseline methodologies applied since 2017. The study focuses on countries with a relatively advanced use of distributed and demand-side flexibility in balancing, adequacy and wholesale markets. In addition, markets similar to the Belgium market have been chosen where this was considered useful and insightful (e.g. the Netherlands and France). Within the balancing market, a differentiation is made between FCR, aFRR and mFRR products[3]. Figure 6 shows an overview of the product type and countries/TSOs that have been assessed.

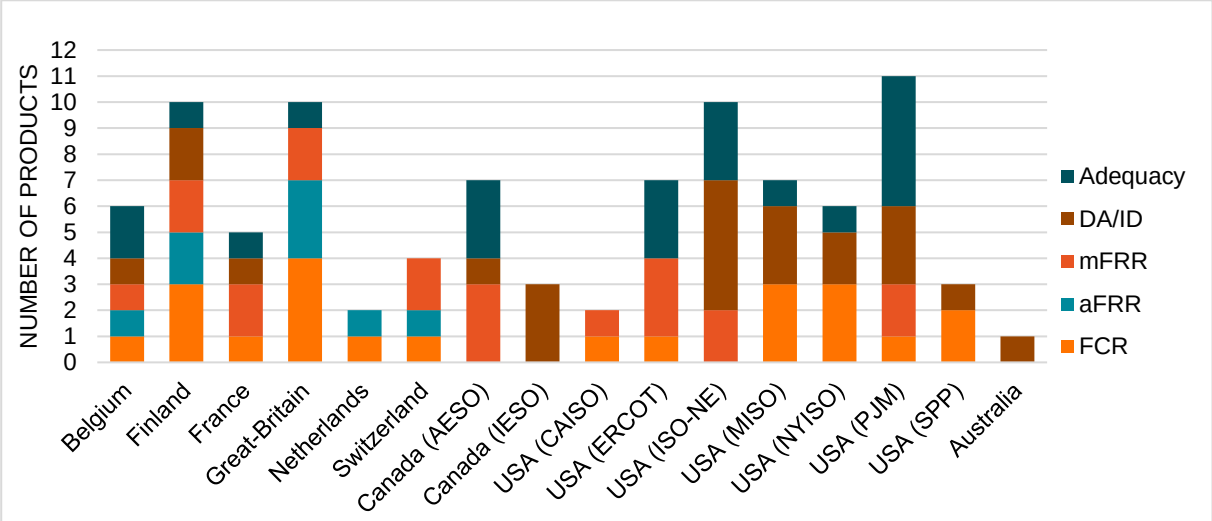

Figure 2: Overview of countries/TSOs and products covered in the benchmark of international practices

2.2.1 Use of different types of baseline methodologies

Figure 3 provides an overview of the baseline methodologies encountered per product and TSO. In addition, Figure 4 and Figure 5 respectively provide a mapping of the baseline methodologies encountered relative to the activation du-

[3] The study classified the products based on the classification of the relevant TSO. For products outside the EU, the classification is based on the parameters of the service and their mapping to European standards. The mapping of these products is based on the design purpose of the product, the activation period and the response time. DNV has applied the best of its knowledge and experience to provide a sufficient mapping to put GB, US, Canadian and Australian products in the right context.

ration and the period between the activation request and the start of the activation. The assessment looks at the occurrence of different types of the baseline methodologies and does not distinguish between different variants of a certain baseline methodology type (e.g., different variants of historical baseline methodologies).

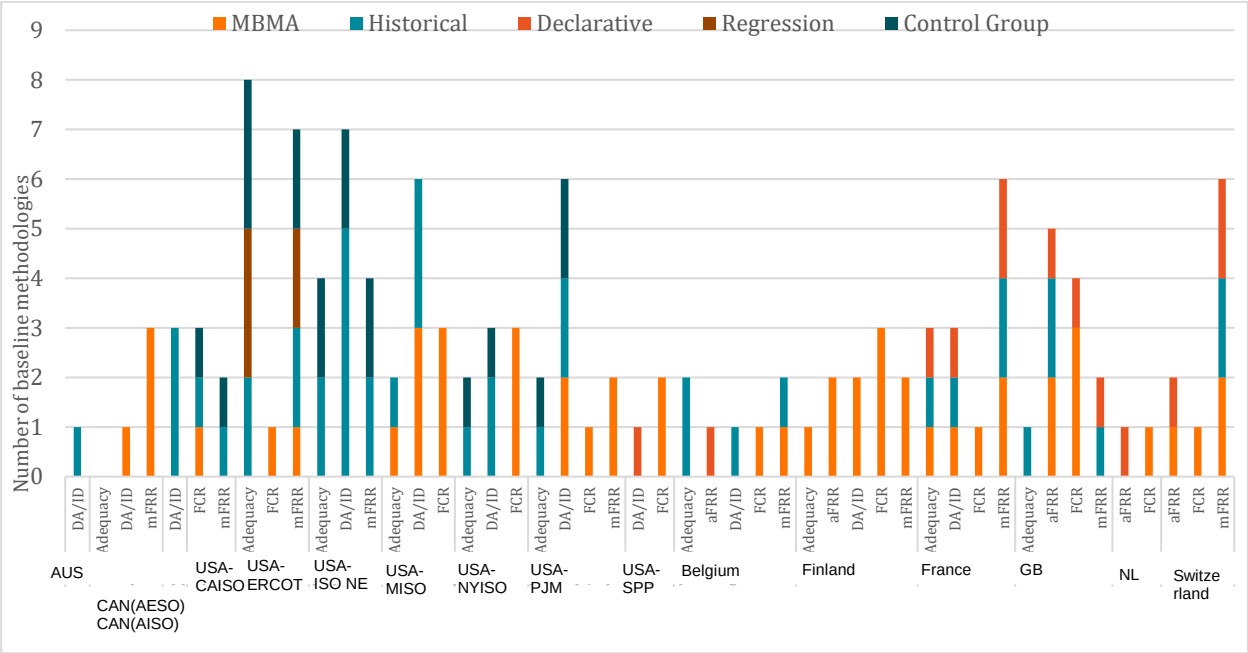

Figure 3: Overview of baseline methodologies encountered per product and TSO

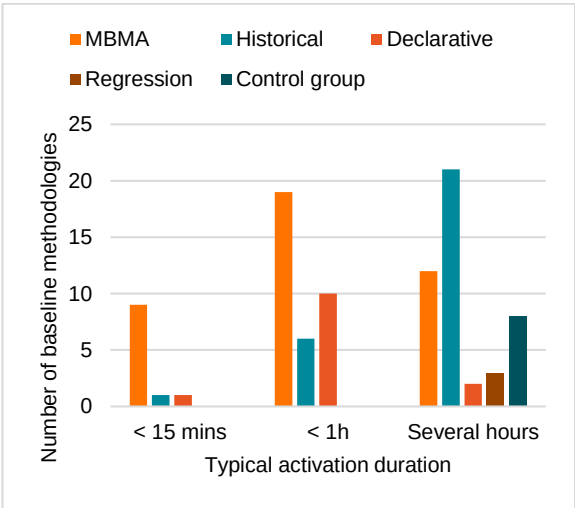

Figure 4: Mapping of baseline methodologies to the activation duration

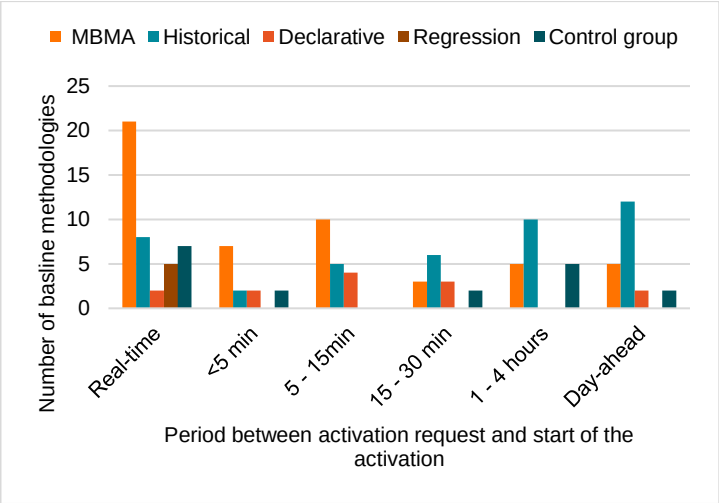

Figure 5: Mapping of baseline methodologies to the duration of the period between the activation request and the start of the activation

With respect to the use of **historical baseline methodologies**, it can be observed that these methodologies are commonly used for nearly all products (in particular for mFRR, adequacy and DA/ID products) and systems. Particularly for adequacy and DA/ID products, which are characterized by longer activation periods and longer times between the activation request and the start of the activation, historical baseline methodologies tend to be preferred compared to other (under these conditions) less accurate baseline methodologies such as MBMA methodologies, or more complex baseline methodologies such as regression-based baseline methodologies. In contrast, the use of historical baseline methodologies is limited for products such as FCR and aFRR, which are characterized by short activation periods and a short period between the request for activation and the start of the activation.

Although not visible in the figures, a second observation from the assessment related to the use of historical baseline methodologies is that the majority (around 70%) of the historical baseline methodologies assessed involve some form of SDA. In general, different versions of the historical baseline methodologies can be found that all attempt to address different potential shortcomings of the approach. Versions in which not all reference days are selected (e.g. middle 8 of 10 or high 4 of 5) decrease the likelihood of outliers affecting the baseline. Moreover, for activations that are more likely to take place during days with extreme weather conditions, high X of Y historical baseline methodologies produce a baseline that has the shape characteristics better reflecting the most extreme days.

With respect to the use of **MBMA baseline methodologies**, it can be observed that these baseline methodologies are most commonly used for products that have a short activation periods and a short time between the activation request and the start of the activation, such as FCR, aFRR and mFRR. The short activation period and the short time between the activation request and the start of the activation make the simple, flat MBMA baseline similarly accurate as an historical baseline methodology with SDA. In contrast, MBMA baseline methodologies are only to be found for adequacy and DA/ID products in certain US systems and this generally as part of a suite of baseline methodologies. Given the longer activation periods for these products, MBMA baseline methodologies will only be appropriate for a subset of loads. For example, MBMA baseline methodologies would still provide accurate results for longer activation periods for loads that are not variable (e.g. industrial customers with flat operational profile or back-up generators).

With respect to **declarative baseline methodologies**, it can be observed that these baseline methodologies are not as widely used as historical and MBMA baselines. Nevertheless, the use of declarative baseline methodologies has become more common, particularly in Europe, with the increased participation of variable load in flexibility services.

In particular, declarative baseline methodologies are found to be rather used for products with short periods between the activation request and the start of the activation as this improves the ability of the FSP to provide an accurate forecast (considering that the baseline typically needs to be submitted before the activation request to limit gaming opportunities). In practice, this corresponds mainly to balancing products. The only instance where a declarative baseline is used for services with a longer period between the activation request/decision and the start of the activation is in France (NEBEF).

Declarative baseline methodologies provide the freedom and allocate the responsibility for calculating the baseline to the FSP. As such, different methods could be used for the baseline calculation (including common baseline methodologies such as MBMA or historical baseline methodologies but also more complex or tailor-made methodologies). In terms of accuracy and inclusivity of this baseline methodology, it is hence important to consider the forecasting ability of the FSP. In terms of complexity, declarative baseline methodologies have the advantage of enabling the use of complex baseline calculations while limiting the implementation complexity, but also require measures to mitigate potential gaming opportunities (e.g., requesting submission of the baseline before the activation request).

Regression-based, calculated and control group baseline methodologies are not commonly applied, mainly due to their complexity or limited applicability. As shown in Figure 3, no regression or control group baselines were found in the European countries part of the assessment. Regression-based baseline methodologies are only found in a minority of US systems. The best-known example of a regression-based baseline methodology is the ERCOT regression baseline approach, which was an attempt to provide an accurate baseline for the widest possible set of customers under extreme weather variability. As per the literature review, the regression-based baseline methodology was initially the only option for ERCOT. However, following stakeholders' feedback on the complexity of this baseline, ERCOT decided to add additional baselining methodologies with a lower complexity and now offers the regression-based baseline methodology as one of the options.

Similar to regression-based baseline methodologies, **control group baseline methodologies** are only found in US systems and are always offered as part of a set of baseline methodologies. Control group baseline methodologies have been mainly used for residential customers (e.g. in ERCOT). In some other cases, control group baselines were introduced to estimate electricity usage of an aggregated demand resource where interval metering was not available for the participating population. However, the market analysis confirms that effective usage of control group baseline methodologies is limited and that historical baseline methodologies are always the default option due to the balance between accuracy and simplicity.

Calculated baseline methodologies (as per the definition of this study) are not used by the TSOs for which the assessment was performed. However, in some cases such as in the baseline methodology applied by Tennet for aFRR, declarative baselines are used in which the FSP (under certain conditions) is allowed to submit the baseline until real-time. This enables the FSP to use similar methodologies as those that could be used for a calculated baseline methodology, for instance the calculation of the Actual Available Power (AAP) of a wind turbine/wind farm based on real-

time wind-speed measurements. In this regard, the difference between declarative baselines and calculated baselines may become small. A more in-depth analysis of declarative and calculated baseline methodologies can be found in Sections 5 and 6.

2.2.2 Multiple baseline methodologies for a single product

Figure 6 provides an overview of the occurrence of multiple types of baseline methodologies for the different types of products. Enabling multiple types of baseline methodologies comes with an increase in complexity, but gives the FSP greater flexibility to choose a baseline that is appropriate for their assets, while the TSO maintains control and visibility of all underlying baselines.

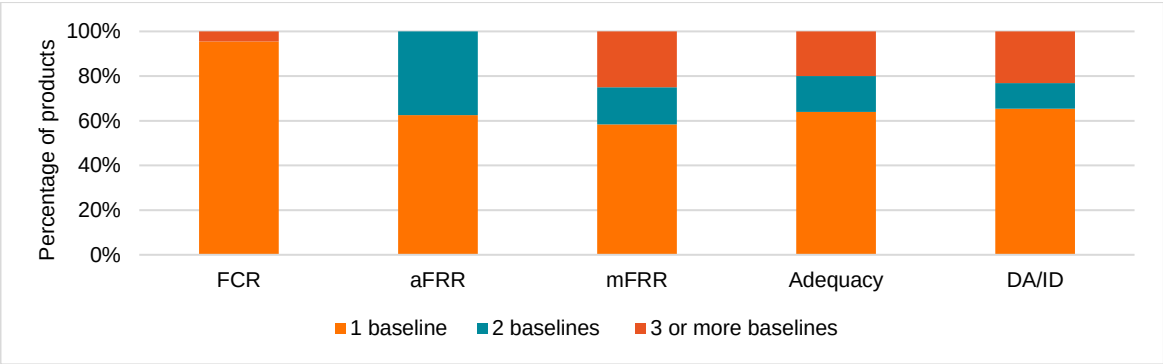

Figure 6: Occurrence of multiple baseline methodologies per product

In Europe, the possibility to choose between two or more baseline methodologies is the result of the evolution of the flexibility products and the participation of different loads. However, for the majority of assessed products only a single baseline methodology is applied. This is particularly the case for FCR products, where, in the exception of CAISO, a single type of baseline methodology is used in all systems. For aFRR, more than one type of baseline methodology is observed to be used in Switzerland and Great Britain. In these countries, a combination of declarative, historical and/or MBMA baseline methodologies can be used. For mFRR, several combinations are used both in the US and in Europe (e.g., Switzerland, France and Belgium). In France and Switzerland, declarative, historical and MBMA baseline methodologies are used. In North America, the available combinations are historical with control group and/or regression-based baseline methodologies. For adequacy products and DA/ID services, several combinations of baseline methodologies are available which do not indicate a clear trend. In Europe only in France a combination of historical, declarative and MBMA baseline methodologies is available. In North America, several combinations with historical baseline methodologies are in place.

In general, the gained experience with regard to complexity and accuracy and the participation of more diverse customers (e.g., variable loads and generation, residential customers) has led to the introduction of more baseline methodologies for one product. However, it must be noted that there is no clear trend towards the introduction of more complex baseline methodologies. For example, experiences in ERCOT have led to the introduction of simpler baseline methodologies than the regression-based baseline methodology. A different example can be found in GB, where

declarative baseline methodologies have been used for many years for most of their products whilst historical baseline methodologies were added when the participation of DSR assets increased.

2.3 Literature review

A review of the literature with respect to baseline methodologies has been performed in order to capture best practices. This section provides an overview of the key messages and most relevant reference works.

The majority of the studies focusing on baseline methodologies aim to provide insights regarding the accuracy of different baseline methodologies. In this regard, various studies have performed elaborated assessments of the accuracy of common baseline methodologies based on large data sets covering multiple years and a variety assets or customer segments. For instance:

- a 2019 study for the Australian Renewable Energy Agency[4] analyzed the accuracy of different variants of historical baseline methodologies based on data of a three-year trial by the California Independent System operator (CAISO) for industrial, commercial and residential participants to the Reliability and Emergency Reserve;
- a 2011 study from KEMA for PJM[5] performed an empirical analysis of the accuracy of different variants of MBMA, historical as well as regression-based baseline methodologies based on more than 2 years of data of over 4500 commercial and industrial demand-response customers;
- a 2018 study of DNV-GL[6] performed an evaluation of the accuracy of MBMA, different variants of historical baseline methodologies as well as regression-based baseline methodologies for a two-year demonstration project in Massachusetts for 99 commercial and industrial customers across 9 different sectors;
- a 2019 study from Lee[7] assessed the performance of different variants of historical baseline methodologies with a focus on the residential sector.

A first conclusion is that historical baseline methodologies are systematically assessed to be accurate for a wide variety of assets[4,5,6,7,8]. Different studies also investigated different variants of historical baseline methodologies. These different variants of historical baseline methodologies tend to differ in five different aspects: 1) the amount of representative days selected, 2) the criteria used for the selection of the representative days (e.g., take the Y days prior to the activation, the Y previous days that are the same day of the week, etc.), 3) the approach taken to select X reference days out of these Y representative days (e.g., take all representative days, exclude days with the highest/lowest offtake, etc.), 4) the method used to calculate the baseline based on the data from the X reference days (e.g., median, average or weighted average where a higher weight is given to the most recent reference days), and last but not least 5) whether or not a same-day adjustment is applied (and which type). In this regard, multiple studies conclude that applying a same-day adjustment can significantly increase the accuracy[5,6,7,8,9]. The type of adjustment (i.e., scalar or additive adjustment) is shown not to have a significant impact[5,6,7], although additive adjustments tend to be

[4] Oakley Greenwoord. (2019). Baselining the ARENA-AEMO Demand Response RERT Trial, **baselining-arena-aemo-demand-response-rert-trial.pdf**
[5] KEMA. (2011). PJM Empirical Analysis of Demand Response Baseline Methods, **pjm-analysis-of-dr-baseline-methods-full-report.ashx**
[6] DNV-GL. (2018). Evaluation of 2017 Demand Response Demonstration: C&I Connected Solutions.
[7] Lee, Seungman. (2019). Comparing Methods for Customer Baseline Load Estimation for Residential Demand Response in South Korea and France: Predictive Power and Policy Implications.
[8] Goldberg, Miriam L. and Agnew, G. Kennedy. (2013). Measurement and Verification for Demand Response. **Measurement and Verification for Demand Response (energy.gov)**
[9] EnerNOC. (2011). The Demand Response Baseline (White Paper). **CEE_EvalDRBaseline_2011.pdf (cee1.org)**

preferred[5,8], mainly for the reason that scalar adjustments are more prone to producing extreme outliers (in particular when the offtake/injection is close to zero). Without applying a same-day adjustment, historical baseline methodologies tend to underestimate the load during emergency situations. In this regard, historical baseline methodologies, such as a High 4 of 5, that select the days with the highest offtake tend to perform slightly better[5,6]. However, it must be noted that different studies indicate that in case a same-day adjustment is applied, the amount of representative days that are used and the criteria applied to select the reference days from these representative days only have very limited impact on the accuracy[6,7].

With respect to other baseline methodologies, the literature indicates that for products with short activations and with activation requests close to real time, simple MBMA baseline methodologies are well suited. In contrast, for products with longer activations, the literature provides clear evidence that MBMA baseline methodologies tend to provide a significantly lower accuracy due to the fact that these baseline methodologies cannot reflect expected changes in the offtake/injection taking place during the period of activation[5,6].

On the other hand, more complex, and sometimes site-specific regression-based baseline methodologies (considering e.g., the daily temperature, the day of the week, schedule variables etc.) tend to achieve a good accuracy. More specifically, different studies evaluated regression-based baseline methodologies to be more accurate than historical baseline methodologies that do not apply a baseline adjustment, and equally or slightly less accurate than historical baseline methodologies that do apply a baseline adjustment[5,6,8].

For other types of baseline methodologies, there is little evidence available. For certain types of baseline methodologies such as declarative baseline methodologies, assessing the accuracy in a general way is difficult as the accuracy is dependent on the forecasts (and the methods used to make these forecasts) of the FSP. Again for other types of baseline methodologies, such as calculated or control-group baseline methodologies, there is limited experience available.

Although historical baseline methodologies are found to perform well for a large variety of assets, and in particular for larger commercial and industrial loads[4,5], the literature also indicates that for assets with a volatile offtake/injection pattern, historical baseline methodologies have challenges in achieving a high accuracy[4,8]. Volatile offtake/injection patterns are for instance weather-sensitive loads, loads influenced by on-site renewable generation (in case no submetering would be possible) or loads that follow specific processes that vary from day to day without a consistent pattern. This irregular volatility is also observed for individual residential households, for which historical baseline methodologies can be insufficiently accurate[4,5,8]. However, on a more aggregated level, historical baseline methodologies are shown to provide good results for the residential sector[7,8]. For irregular loads, allowing customers to provide their own forecast of their load/generation (i.e., using a declarative baseline methodology) has been suggested[8,10].

Finally, in addition to assessments and considerations regarding accuracy, different studies also recognize and emphasize the importance of considering other aspects, such as simplicity/replicability, inclusivity, scalability and integrity[8,9,10]. In this regard, the increased data requirements and processing complexity of regression-based baseline methodologies is considered to make this baseline methodology less practicle[8]. For this reason, multiple studies also

[10] Ena energy networks association. (2020). Open Networks Project – Baseline Methodologies. **ON20-WS1A-P7 Baseline Principles-PUBLISHED.23.12.20.pdf (energynetworks.org)**

recommend historical baseline methodologies over regression-based baseline methodologies[5,9] (although they can achieve a similar performance in terms of accuracy). The importance of simplicity and replicability has also been noticed in practice in ERCOT, where initially a regression-based baseline methodology was developed, but following the feedback from customers indicating the difficulty of replicating the regression-based baseline methodology, more simple historical baseline methodologies have been implemented[4]. Considering the importance of different aspects, the literature also indicates that implementing only one baseline methodology might be the simplest option, but not necessarily the best option[11].

2.4 Return of experience from market parties active in Belgium

Several market parties active in Belgium were interviewed to gather their feedback on experiences with the use of the baseline methodologies for the considered products. In order to preserve anonymity of the participants and the confidentiality of the discussions, only a summary of the interviews is presented in this section. The key findings based on these interviews have been presented and confirmed during an open workshop on the 16th of March 2021.

2.4.1 Overview of key messages

The key messages of the stakeholders were:

1. Accuracy, simplicity and interaction with other products (i.e. possibilities for value stacking) are considered as key criteria for baseline methodologies.
2. Market parties are generally satisfied with the baseline methodologies adopted by Elia for their current portfolios of assets for mFRR, the CRM, ToE in DA/ID and SR).
3. Though experiences are positive, some questions and potential points of attention have been raised.
4. Market parties have expressed important concerns related to the application of the current baseline methodologies for technologies they might want to add to their pool, such as variable renewables. Market parties indicated that these assets could potentially require new baseline methodologies.

2.4.2 Current experience per product and baseline methodology

Table 2 summarizes the key feedback and remarks provided by the stakeholders regarding the use of the current baseline methodologies for their current portfolio of assets.

[11] DNV KEMA. (2013). Development of Demand Response Mechanisms – Baseline Consumption Methodology – Phase 1 Results. **2013 DNV KEMA report-proposal template (aemo.com.au)**

Table 2: Overview of stakeholder feedback on the current baseline methodologies

Product	Baseline methodology	Key remarks
mFRR	Last QH	Market parties recognize that Last QH works very well for almost all delivery points and the methodology is generally accepted by all stakeholders. The key drivers for market parties to use this baseline is accuracy and simplicity. Potential points of attention were related to the fact that in specific situations/moments, the Last QH baseline methodology can involve certain risks: • When an activation is requested in the quarter hour following a quarter hour in which a planned change in the offtake/injection has taken place, the baseline will reflect the load profile prior to the planned change. As a result, the calculation of the volume of flexibility that is delivered could be either over- or underestimated (depending on the direction of the planned change and the direction of the requested activation). • In case of subsequent activations for which there are 2 quarter hours between the first and the second activation, the reference QH that is used to determine the baseline for the second activation can be polluted due to the ramp down of the first activation.
	High X of Y	Market parties are generally satisfied with this baseline methodology although the experience with using this baseline is limited. This results from the fact that this baseline methodology was introduced later than the Last QH baseline methodology, is more complex to implement and most market parties did not perceive a need to use a different baseline methodology than Last QH. Feedback furthermore suggests that this baseline works well for loads with regular daily/weekly schedules/load profiles. No specific issues were mentioned.
Strategic reserves	High X of Y	No particular comments or suggestions for improvements were provided, but it has to be recognized that experience is limited with using this baseline for SR. Questions were raised by one party regarding the accuracy of this baseline during extreme situations/cold spells as well as regarding the applicability of this baseline methodology for all types of assets.
ToE DA/ID	High X of Y*	No particular comments or suggestions for improvements were provided, but it has to be recognized that there Is not yet experience with the use of this baseline methodology for this product. Market parties were generally satisfied to have the possibility to apply a Same Day Adjustment (SDA) for delivery points where it can be proven that the SDA increases accuracy. Questions were raised by one party regarding the accuracy of this baseline during extreme situations/cold spells as well as regarding the applicability of this baseline methodology for all types of assets (i.e., the inclusivity of the baseline methodology)[12].

2.4.3 Need for future developments

Market parties were also invited to provide their view on the possible need for future developments in terms of baseline methodologies to enable the participation of technologies that are not part of their current portfolios but that are envisaged to be added in the coming years.

A first key message is that market parties have expressed clear concerns regarding the accuracy of the current baseline methodologies for certain technologies that are expected to provide Balancing Services (FCR, aFRR and mFRR) in the future. This concerned mainly variable renewables (in particular wind and solar PV), but also, to a smaller extent, batteries and residential flexibility (household batteries, EV chargers, heat pumps and e-boilers).

Designing baseline methodologies for variable renewables was considered a priority by stakeholders, considering that there are significant volumes of wind generation and solar PV available that could participate to the balancing markets. The primary target was generally considered to be in first instance the provision of downward aFRR, although the provision of downward mFRR was also not excluded. The current baseline methodology for aFRR (declarative baseline that needs to be submitted on a 4-second basis one minute before real-time) and mFRR (Last QH and High X of Y) are considered blocking by stakeholders[13]. In this regard, stakeholders did express an interest for calculated baseline methodologies for wind and solar PV.

Some market parties furthermore expressed concerns regarding baselining for batteries which are operated not exclusively to provide system services (but also, e.g., to optimize self-consumption) or to provide multiple system services. In this regard, market parties suggested that declarative baselines could be feasible for participation of batteries in aFRR (although issues could remain in combination with self-consumption).

Regarding residential flexibility, some market parties expressed doubts regarding the accuracy of the current baseline methodologies. Market parties indicated that current baseline methodologies could potentially work, but that experience is limited. Market parties furthermore suggested that declarative baselines could be a feasible solution if historical baselines prove to be insufficiently accurate. However, in general, addressing baselining for residential flexibility was considered by interviewed market parties to be a lower priority for the moment[14].

[12] In the context of the **study on ToE DA/ID** in 2019 (prior to its implementation), different market parties did express concerns regarding the suitability of the High X of Y* baseline methodology for all types of assets and expressed a desire to have the possibility to choose between different baseline methodologies and/or to be able to propose their own baseline methodology. However, these concerns have not been repeated in the interviews organized with market parties in the context of this study.

[13] Note that, as discussed in Section 5.3.2, for assets providing MW schedules (i.e., the DP_{SU}) these MW schedules are used for the activation control of mFRR (i.e., one could consider that there is de-facto a declarative baseline methodology for these assets). The MW schedules are already today used for the participation of Delivery Points that have an obligation to provide MW schedules, such as large wind farms. The isues raised by stakehoilders during the interviews related to mFRR focused on the Last QH and High X o f Y baseline methodologies.

[14] Although there is not a sufficient return of experience on the use of the current baseline methodologies for residential flexibility that can be considered as part of this study, Elia is in parallel with this study experimenting with the possibilities for residential flexibility via the Flexity project.

3. Assessment of the baseline performance

This section presents the assessment of the performance of the different types of baseline methodologies for each of the products in the scope of the study. First, Section 3.1 and Section 3.2 respectively present the methodology for the performance assessment and the products characteristics relevant for the assessment. Next, Section 3.3 presents the result of the performance assessment for the different baseline methodologies for each of the considered products.

The assessment has been performed by the consulting firm DNV. The assessment criteria and weights have been based on findings from the literature survey, the experience of DNV as well as the interactions with stakeholders (including the CREG). The methodology and the results have been presented during a workshop on June 17th 2021 to gather feedback.

3.1 Assessment criteria and methodology

The assessment of the performance of the different baseline methodologies for each product is based on the criteria described in Table 3.

Table 3: Overview of assessment criteria and weights

Criterion	Description	Weight
Accuracy	This factor refers to whether or not the baseline is sufficiently accurate for the settlement/activation control of the considered product as well as, if applicable, for the calculation the delivered volume of flexibility that forms the basis for Transfer of Energy.	35%
Simplicity	This criterion reflects the level of effort and the complexity of implementing and operating/using the baseline methodology, including but not limited to collecting the right data, performing the calculation as well as possible data exchanges. This criterion also considers the ease and possibilities for FSPs to replicate the baseline calculation by the FSP.	30%
Inclusivity	This criterion reflects to what extent a baseline methodology can provide accurate results for a variety of different types of assets. Some baselines are technology-agnostic and suitable for (almost) all technologies, whilst some other baselines may be applicable or accurate only for certain technologies, consumer types and/or portfolios of assets. For instance, the Last QH baseline methodology may prevent a wind park from participating in a flexibility service due to low baseline accuracy as its generation profile may show high volatility.	15%
Integrity	This criterion relates to the gaming opportunities and possibilities for strategic behavior when using a given baseline methodology.	15%
Facilitation of value stacking	This criterion refers to opportunities for value stacking that a baseline facilitates. Value stacking refers to the ability of an FSP to participate in different markets and services at the same time with the same resource, allowing them to stack revenues from various streams.	5%

Different weights are given to the different assessment criteria. The literature and experiences with respect to baseline methodologies show that priorities vary across flexibility products, markets, organizations and countries. Whilst accuracy of the baseline is a key criterion for markets with large participation across all sectors and technologies, simplicity and inclusivity may be the main priority for markets that are less mature. Therefore, the assessment criteria were prioritized based on the following evidence:

- Feedback from market parties active in Belgium suggested that simplicity is one of the main drivers for using the existing baselines. This is also observed in other markets. For instance, in a recent assessment of baselines for DSO products in GB, simplicity was prioritized above all other criteria following engagement with stakeholders of GB market and GB network companies[10].
- Accuracy is key in order to ensure appropriate compensation of flexibility activation, and to ensure the effectiveness of the product where the flexible asset is used. The assessment of the accuracy of each baseline for each product in this section is based on multiple studies that have performed quantitative assessments of the accuracy of different baseline methodologies as well as stakeholder's feedback. As presented in Section 2.3, the literature describes clear conclusions on the accuracy of baseline methodologies that are confirmed in multiple studies. In addition, many of these studies build on large data sets covering multiple years and many participating sites/assets. Performing a statistical analysis of the accuracy for the Delivery Points participating in the Belgian market is in this regard not considered to provide an added value given the current limited market participation (e.g. no participation yet to the ToE in DA/ID or the CRM, very limited use of the high X of Y baseline in mFRR). In contrast, a statistical analysis based on Belgian data could involve a risk of providing skewed or non-representatitve outputs.
- Inclusivity was considered important in order to remove barriers to participation to existing products. It must be noted that a baseline methodology that is only applicable/accurate for a limited set of assets would not perform well in terms of inclusivity, and would not be a good choice as a unique baseline methodology available for a given product. However, a low score on inclusivity can be mitigated by enabling the choice between multiple baseline methodologies. For this reason, this criterion was given a lower weight than simplicity and accuracy.
- Integrity was also considered an important criterion. However, as mitigation measures could be put in place in order to remove/reduce the possibilities and/or incentives for strategic behavior, this criterion was also given a lower weight compared to accuracy and simplicity.
- Facilitation of value stacking was given the lowest weight given that the possibilities for value stacking are not primarily driven by the baseline methodologies for the considered products; both market design principles and product design principles are key for enabling value stacking.

The identified baselines are assessed for each considered product against the assessment criteria based on the assessment framework shown in Table 4. The total scoring of each baseline uses the sum-product of the weight and the corresponding score for each criteria.

Table 4: Assessment framework

Criterion	5 - Excellent	4 – Good	3 – Average	2 – Fair	1 – Poor
Accuracy	Variance: Small variance throughout the service window. Bias: No bias throughout the service window	Variance: Small variance in general, medium variance in certain parts of the service window. Bias: Small bias in certain parts of the service window	Variance: Small, medium and high variances depending on the time within the service window. Bias: Small or no bias on average, yet higher bias at specific moments of the day / service window	Variance: Some small and medium variances, mostly high variance during the service window. Bias: Some bias on average, yet higher bias most moments of the day / service window	Variance: High variance for larger part of the service window. Bias: Strong bias throughout the service window
Simplicity	Simple, straightforward calculation	Relatively simple calculation	Slightly complex calculations	Rather complex calculations	Highly complex calculations
Inclusivity	Fully technology agnostic	Technology agnostic with few exceptions	Favoring specific technologies, not excluding others	Favoring specific technologies, excluding some others	Favoring specific technologies, excluding others
Integrity	No gaming options foreseeable	Some gaming options, with sufficient mitigations options	Likely gaming options with good mitigation / some gaming options with little mitigation	Likely / obvious gaming options with some mitigation	Obvious gaming options with little/no mitigation
Facilitation of value-stacking	Stacking is always possible	Stacking is possible in most situations	Stacking is sometimes possible	Stacking is possible in exceptional situations	Stacking is never possible

3.2 Products and characteristics

There are four product characteristics that should be considered when evaluating the performance of a baseline methodology:

- The typical activation duration: this product characteristic mainly impacts the accuracy of the baseline. Typically, long activation periods reduce the accuracy of certain baseline methodologies, particularly for assets with a variable offtake/injection pattern. For example, the MBMA baseline methodology is not very accurate for long activation periods and therefore it is less frequently used in products with long activation periods.
- The time between the activation request/decision to activate and the start of the activation: this product characteristic mainly affects the integrity and inclusivity of a given baseline methodology. This in particular for declarative baseline methodologies, MBMA baseline methodologies and historical baseline methodologies with same-day adjustments. This product characteristic refers to the moment that the FSP has knowledge of the (probability of an) activation, and hence could have incentives to impact, or even manipulate the baseline (if possible). For example, in case of an activation related to a DA trade for which a historical baseline with a same-day adjustment is used, the FSP could manipulate their consumption on the activation day to manipulate the baseline adjustment.
- Frequency of activation: this product characteristic mainly impacts the data selection, and hence the performance, of historical baseline methodologies. For example, due to the fact that days with activations are typi-

cally excluded, it may prove challenging for products with high frequency of activation to obtain a set of reference days sufficiently close to the day of the activation. As such, the frequency of activation could impact the accuracy of historical baselines.

- Metering resolution: the metering resolution used for the validation of the service.

Table 5 summarizes the key characteristics for the considered products. These characteristics are considered in the assessment of the performance of the different baseline methodologies for the different products under consideration.

Table 5: Product characteristics relevant for the assessment of the baseline performance

Characteristic \ Product	**mFRR**	**Strategic reserves**	**ToE DA/ID**	**CRM**
Typical activation duration	< 30 minutes[15]	1 to several hours	1 to several hours	1 to several hours
Time between activation request/decision and start of the activation	0-15 minutes[16]	ID, a few hours in advance	Day-ahead after 15:00; For ID deals can also be until 5 minutes before real time	Day-ahead after 15:00[17]
Frequency of activations	Low	Low	In theory, activations could happen on a daily basis, but the current expectation is that the activation frequency will be low (e.g., during near-scarcity moments or exceptional balancing events, cf. study on ToE in DA/ID)[18]	Low
Metering resolution	15 minutes	15 minutes	15 minutes	15 minutes

3.3 Results

Using the assessment criteria and framework (presented in Section 3.1) and taking into account the relevant product characteristics for each of the considered products (presented in Section 3.2), the performance of each baseline methodology is assessed.

[15] This is based on observed activations of Delivery Points of the type DP_{PG} in 2020. The duration of a typical activation might change once MARI is operational and depending on activation prices of DP_{PG}.
[16] Currently, the time between the request for activation and the start of the activation is between 0 and 15 minutes in case of a scheduled activation, and 0 minutes in case of a direct activation. As of the connection to the MARI platform, there will be no time between the request and the start of the activation (although BSPs will have 2,5 minutes after the activation request to prepare before the ramp starts.
[17] The Availability Monitoring Trigger Moments are determined depending on the day-ahead market clearing price.
[18] The study on Transfer of Energy for DA and ID markets can be found on the **Elia website**.

It has to be noted that the evaluation of the performance of a given baseline methodology in terms of simplicity and inclusivity is considered to be independent of the product. Hence, the score of a specific baseline methodology on these criteria is identical across all products.

3.3.1 mFRR

Table 6 provides an overview of the performance of the different baseline methodologies for mFRR. A justification for the scoring on all assessment criteria is provided below.

Table 6: Evaluation of the performance of different types of baseline methodologies for mFRR

Criteria	Final Scoring	Simplicity	Accuracy	Inclusivity	Integrity	Facilitation of value stacking
Weighting factor		30%	35%	15%	15%	5%
MBMA	4.2	5	4	3	4	4
Historical without same-day adjustment	3.5	4	3	2	5	3
Historical with same-day adjustment	4.2	4	5	3	4	4
Declarative	3.8	3	4	4	4	5
Calculated	3.4	2	5	1	5	3
Regression-based	3.4	1	5	3	5	3
Control group	3.1	2	4	1	5	3

Simplicity

Big differences can be observed regarding the simplicity of the different baseline methodologies. Looking at each baseline methodology individually, the following observations can be made:

- MBMA baseline methodologies are simple to use for both the FSP and the TSO as it involves no calculation and requires only the metering value prior to the activation.
- Historical baseline methodologies without same-day adjustment are not as simple as Last QH (MBMA) but the baseline calculation has a limited complexity and only requires the use of measurement data and data related to prior activations.

- Historical baseline methodologies with same-day adjustments require an additional step in the calculation compared to historical baseline methodologies without a same-day adjustment. Nevertheless, it is considered that this does not significantly affect the level of complexity.
- Declarative baseline methodologies have a higher complexity compared to MBMA and historical baseline methodologies. From the perspective of the FSP, the complexity of the baseline calculation depends on the tools and data used by the FSP. In addition, in contrast to MBMA and historical baseline methodologies, declarative baseline methodologies requires additional communications with the TSO on a regular and possibly close to real-time basis. From the perspective of the TSO, declarative baseline methodologies require additional monitoring processes to ensure sufficient accuracy and to mitigate potential gaming opportunities.
- Calculated baseline methodologies are rather complex as they involve calculations based on external parameters. For example, to calculate the baseline for a wind turbine/park, a SCADA system may be required to calculate the available power using wind speed data and the theoretical power curve(s). In some cases, further calculations may be required to take into considerations grid losses and/or wake effects[19]. The methodology for the baseline calculation is not always standardized. The FSPs might need to develop their own processes and therefore, the complexity of these baselines is considered to be high. It has to be noted that the complexity of calculated baseline methodologies depends on the calculation requirements of the concerned TSO, the presence of readily available software systems capable of performing the calculations (e.g. SCADA systems) and the availability of the data required to perform the calculations.
- Regression-based baseline methodologies require more complex calculations as well as large amounts of historical data (usually one year worth of data) that is not restricted to the measured power offtake/injection at the Delivery Point. Frequently, regression-based methodologies consider site-specific parameters, therefore further increasing the complexity.
- Control group baseline methodologies are relatively complex as they require measurements of similar customers (or assets) that do not participate to the flexibility service. Appropriate selection of the peer group and management of large data volumes is required; This complexity can lead to a lack of transparency and/or difficulties for third parties to calculate/verify the baseline calculation.

Accuracy

All baseline methodologies perform well for this product in terms of accuracy as a result of the short activation period and short time between the activation request and the start of the activation. Looking at each baseline methodology individually, the following observations can be made:

- MBMA baseline methodologies tend to perform well in terms of accuracy for products with short activation periods (as discussed in Section 2.3).

- Historical baseline methodologies without same-day adjustment perform reasonably well in terms of accuracy. The relatively low frequency of activations of most Delivery Points improves the accuracy of historical baselines. This because a low frequency of activations allows selecting non-event days closer to the day of

[19] A more detailed decription of calculated baseline methodologies can be found in Section 6.

the activation. Using data of recent days improves the accuracy as days that are close to each other tend to have more similar conditions (e.g. weather conditions or operational conditions). Moreover, for longer activations, historical baseline methodologies have the benefit of being capable of forecasting the variable offtake/injection pattern of the Delivery Point (at least, in case there is a regular daily/weekly pattern). However, historical baseline methodologies without same-day adjustment score less well in terms of accuracy compared to the other baseline methodologies because the conditions on the day of the activation are not considered.

- Historical baseline methodologies with same-day adjustment are typically one of the most accurate baseline methodologies, as they can forecast the possibly variable offtake/injection pattern, while also taking the conditions on the day of the activation into account via the calibration with the same-day adjustment. As shown in the literature, the adjustment of the baseline significantly improves the accuracy.

- Declarative baseline methodologies perform well in terms of accuracy. Due to the fact that the activation periods tend to be short and that the activation request (and hence baseline submission[20]) takes place close to real time, FSPs should be able to accurately predict the offtake/injection at the Delivery Point. Declarative baseline methodologies enable maximal flexibility to FSPs for constructing an accurate baseline using an appropriate methodology. For certain Delivery Points, more simple methodologies, such as MBMA and/or historical baseline methodologies might suffice, whereas for certain types of loads, achieving high accuracy forecast might require more sophisticated methods to be implemented by the FSP (e.g. regression-based or machine-learning methods). The actual accuracy of the baseline eventually depends on the methodology that the FSP uses, the ability of the FSP to predict its offtake/injection pattern and potential accuracy requirements imposed by the TSO. As such, declarative baseline methodologies may not necessarily always be as effective as other baseline methodologies that have received the highest score in terms of accuracy.

- Calculated baseline methodologies use external conditions and asset conditions to provide the expected offtake/injection. Under the assumption the methodology that is used is suitable (and in some cases tested during pre-qualification), the accuracy of this baseline is considered to be high.

- Regression-based baseline methodologies predict offtake/injection patterns based on the relationship between load and weather and/or other external parameters. As a result, these methodologies allow to consider more factors impacting the offtake/injection at the Delivery Point, and therefore are capable of achieving a high accuracy. The literature review also indicated that regression baselines have a similar performance in terms of accuracy as historical baseline methodologies with same-day adjustment.

- Control group baseline methodologies typically achieve a high accuracy as it is based on measurements during the day of the activation of similar assets that do not participate in the product.

Inclusivity

[20] Note that in the assessment of the performance of declarative baseline methodologies, it is assumed that the baseline needs to be submitted prior to the activation request.

Big differences exist in terms of the inclusivity of the different types of baseline methodologies. Looking at each baseline methodology individually, the following observations can be made:

- In line with the literature review and the interactions with stakeholders, MBMA baseline methodologies are generally considered to perform very well for assets with that do not have high variations in the offtake/injection within a the period of activation. Given the relatively short duration of mFRR activations, MBMA baseline methodologies are considered to perform well for a variety of assets, but may be insufficiently accurate for assets that have a highly variable offtake/injection pattern.
- Historical baseline methodologies without same-day adjustment can have difficulties in achieving a high accuracy for assets with an offtake/injection pattern that does not follow a systematic daily/weekly pattern. For instance, this can be the case for renewable generation, weather dependent loads or loads/generation with a general volatile profile (e.g. residential load).
- Historical baseline methodologies with a same-day adjustment face similar difficulties as historical baseline methodologies without a same-day adjustment for Delivery Points with an irregular offtake/injection pattern. However, historical baseline methodologies with a same-day adjustment ensure a higher accuracy for a wider range of assets as the adjustment enables to consider to some degree the injection/offtake that is dependent on external conditions on the day of the activation.
- Declarative baseline methodologies are considered the most inclusive as it enables FSPs to provide a forecast of the offtake/injection pattern before the activation request based on any calculation methodology and taking into account all relevant and available data. The reason this baseline does not receive the maximal score is that real-time data, such as real-time wind speed or solar irradiance measurements, cannot be considered in the baseline calculation[20]. As a result, the accuracy that can be obtained could depend on the ability to forecast the relevant parameters for the baseline calculation as the baseline needs to be submitted prior to the activation request.
- Calculated and control group baseline methodologies can only be applied to specific types of Delivery Points and therefore they have the lowest score in terms of inclusivity. Calculated baseline methodologies are typically only applied to specific technologies such as wind parks. Control group baseline methodologies can only be applied to specific technologies for which there are high numbers of identical or similar assets. However, as already indicated in Section 3.1, when combined with other baseline methodologies, the introduction of such baseline methodologies could still contribute to removing barriers to participation.
- Regression-based baseline methodologies generally allow to consider more external parameters, therefore making them more broadly applicable compared to historical or MBMA baseline methodologies. However, one important drawback of regression-based baseline methodologies is that they are not suitable for new or other assets for which limited historical data is available.

Integrity

All baseline methodologies perform well for this product in terms of integrity. Considering that it is difficult to anticipate activations, and the limited time between the activation request and the start of the activation, there are little opportunities for gaming. Looking at each baseline methodology individually, the following observations can be made:

- MBMA baseline methodologies in theory allow manipulation by changing the offtake/injection pattern in the quarter hour(s) used to calculate the baseline. However, given that the period used to calculate the baseline is typically a period prior to the activation request. This would imply that the FSP should be capable of (i) anticipating mFRR activation requests and (ii) changing its load/generation in the opposite direction of the activation request without excessive additional costs. For mFRR, activations are considered difficult to predict.

- Historical baseline methodologies without a same-day adjustment cannot easily be manipulated because they are based on measured data well before the actual activation. The baseline could in theory be manipulated by modifying the offtake/injection pattern during non-event days. However, this would again imply that the FSP can anticipate activation requests well in advance and is capable of changing its offtake/injection pattern in the opposite direction of the activation request without excessive additional costs.

- Historical baseline methodologies with a same-day adjustment are slightly more prone to gaming compared to historical baseline methodologies without same-day adjustments. This is due to the additional gaming option of manipulating the adjustment value. When a same-day adjustment is used, the baseline is adjusted to the offtake/injection measured at the Delivery Point during the adjustment period (which is typically defined a relatively short period prior to the activation period). As such, the FSP could modify its offtake/injection during the adjustment period in order to gain a strategic advantage. However, the same arguments as for MBMA baseline methodologies hold also for historical baseline methodologies with a same-day adjustment. Namely, manipulation of the adjustment still requires that the FSP (i) can anticipate mFRR activation requests significantly in advance, and (ii) is capable of changing its load/generation in the opposite direction of the activation request without excessive additional costs.

- Declarative baseline methodologies generally face more gaming opportunities compared to historical or MBMA baseline methodologies as the baseline can be modified without having to manipulate the actual offtake/injection patterns. Nevertheless, the gaming options for declarative baseline methodologies are considered limited as mFRR activations are difficult to anticipate. Not knowing when an activation will be requested, an FSP could still benefit from providing structurally biased nomination (e.g., overestimating the actual offtake when offering upward mFRR). However, this type of abuse can be easily mitigated by monitoring the bias of the baseline in moments without activations.

- Calculated baseline methodologies provide little options for gaming as they are based on external parameters and/or well-defined calculation algorithms that cannot be manipulated. It has to be noted that the integrity of calculated baseline methodologies is assessed with the assumption that transparency is provided o the TSO regarding the calculation method during pre-qualification of the baseline.

- Regression-based baseline methodologies face little gaming opportunities as they use methodologies and processes that are based on large volumes of data and parameters that cannot be easily manipulated. Regression-based baseline methodologies do not typically perform same-day adjustments as they already provide high levels of accuracy. A pre-qualification process of the baseline may be required to ensure that a valid method is used.

- Control group baseline methodologies face limited options for strategic behavior since the FSP has limited influence/control on the behavior of the control group.

Facilitation of value stacking

The ease with which different baseline methodologies could support value stacking is discussed across all products in Section 3.3.5.

Conclusions

The key conclusions of the assessment of the different baseline methodologies for mFRR are listed below:

- MBMA baseline methodologies and historical baseline methodologies with a same-day adjustment outperform all other baseline methodologies for mFRR as they combine a good score in terms of both simplicity and accuracy. Declarative baselines also perform well in terms of accuracy (and inclusivity), but score average in terms of simplicity. Regression-based, control group and calculated baseline methodologies can achieve a high accuracy but are complex. Moreover, control group and calculated baseline methodologies have a limited inclusivity.
- Declarative baseline methodologies are the most inclusive baseline methodologies, as they enable the FSP to use a suitable methodology for providing their own forecast. Historical baseline methodologies and MBMA baseline methodologies perform slightly worse on this criterion. As such, barriers for participation to mFRR could in theory be reduced in case more baselining options are available. For example, adding the option of a declarative baseline could remove possible barriers to participation for assets that show insufficient accuracy using the current baseline methodologies.
- Integrity can be ensured at a sufficient level for all baseline methodologies for mFRR.

3.3.2 ToE DA/ID

Table 7 provides an overview of the performance of the different baseline methodologies for ToE DA/ID. A justification for the scoring on all assessment criteria is provided below.

Table 7: Evaluation of the performance of different types of baseline methodologies for ToE DA/ID

Criteria	Final Scoring	Simplicity	Accuracy	Inclusivity	Integrity	Facilitation of value stacking
Weighting factor		30%	35%	15%	15%	5%
MBMA	2.9	5	1	3	3	3
Historical without same-day adjustment	3.4	4	3	2	4	4
Historical with same-day adjustment	3.7	4	4	3	3	3
Declarative	3.1	3	3	4	2	4
Calculated	3.0	2	4	1	4	4
Regression-based	3.0	1	4	3	4	4
Control group	3.0	2	4	1	4	4

Simplicity and inclusivity

As indicated earlier, the performance of a baseline methodology in terms of simplicity and inclusivity are considered to be independent of the product for which the baseline methodology would be used. For this reason, the scores for these assessment criteria and the corresponding justification are identical to those for the other products under consideration (see Section 3.3.1 for a more detailed justification of the scoring of the different baseline methodologies in terms of simplicity and inclusivity).

Accuracy

As argued in Section 3.2, the accuracy of certain baseline methodologies is dependent on the typical activation duration, the frequency of activation and the time between the activation request/decision to activate and the start of the activation. Considering that for ToE DA/ID, there is no activation request issued by the TSO, the assessment of the accuracy is performed under the assumption that the decision to activate is taken at the moment of the DA market clearing. In addition, activations in the context of ToE DA/ID are assumed to take place mainly in a limited number of periods where high prices are observed in the DA and/or ID markets, for instance as a result of near-scarcity situations. This is in line with experiences in the NEBEF mechanism. Under these considerations, the following observations can be made:

- MBMA baseline methodologies achieve the lowest score in terms of accuracy, considering that for all assets except those with a flat injection/offtake profile, this baseline methodology achieves insufficiently accurate results for products with (potentially) long activation periods. The literature review and analysis of international practices confirms that MBMAs baseline methodologies are not used for products with long activation periods.

- Historical baseline methodologies without same-day adjustment perform rather well in terms of accuracy but can in general have difficulties in achieving a high accuracy for assets with an offtake/injection pattern that does not follow a systematic daily/weekly pattern. As the accuracy of this type of baseline methodology is not impacted by the time between the decision to activate and the start of the activation, the score is the same as for mFRR. One particular element is that DA/ID activations are more likely to happen in periods with more extreme prices (e.g. elevated prices), possibly resulting from extreme weather conditions such as cold spells. As a result, for weather-sensitive (e.g. temperature dependent loads) or price-sensitive load/generation, historical baselines without same-day adjustment could produce baselines that underestimate the load/overestimate the generation because they are based on less extreme weather/prices in the days prior to the activation. In this regard, selecting the higher load days (high x of y instead of mid x of y) can increase accuracy as evidenced in the literature.

- Historical baseline methodologies with same-day adjustment are typically one of the most accurate baseline methodologies and typically provide more accurate results than historical baseline methodologies without same-day adjustment. This has been also evidenced in the literature. The accuracy of this baseline methodology is nevertheless considered lower than for mFRR because the activations are expected to be longer. As a result, the correction/calibration via the same-day adjustment has a smaller impact.

- Declarative baseline methodologies enable the FSP to choose an appropriate methodology for forecasting the offtake/injection at the Delivery Point. However, achieving a high accuracy might be challenging for variable loads that are inherently difficult to predict well in advance[21]. In addition, offtake/injection patterns might be more difficult to predict under extreme weather conditions (e.g., extreme cold), which might be the cause of the high DA or ID prices.

- Calculated baseline methodologies use external parameters and asset conditions to calculate the baseline. Under the assumption that the methodology used for the calculation is suitable (and tested during pre-qualification), the accuracy of this baseline is high.

- Regression-based baseline methodologies will predict load shapes during extreme conditions based on the relationship between load and weather or other external parameters in less extreme periods. As such, the accuracy of this baseline methodology tends to be high also during more extreme situations.

[21] Although the details of implementing a declarative baseline should be properly considered, the current assessment has taken the assumption that the baseline will be submitted at least day ahead and prior to the day-ahead market clearing.

- Control group baseline methodologies typically achieve a high accuracy as it is based on measurements during the day of the activation of similar assets that do not participate in the product. As such, the impact of possible extreme (weather) conditions can be reflected well.

Integrity

Compared to mFRR, ensuring integrity forms a bigger challenge for ToE DA/ID. As a result, bigger differences can be observed in the assessment of the performance of the different baseline methodologies with respect to this criterion. Looking at each baseline methodology individually, the following observations can be made:

- MBMA baseline methodologies and historical baseline methodologies with same-day adjustment are susceptible to manipulation of the baseline. The baseline could be manipulated by an FSP by changing the offtake/injection during a relatively short period on the day of the activation. For ToE DA/ID, an FSP could already have decided to perform an activation day-ahead (based on the DA market clearing) or several hours in advance (depending on ID market prices). In both cases, an FSP could possibly take advantage by adjusting its offtake/injection pattern according to their activation plans (e.g., increase the load prior to an activation that involves a reduction of the load). It has to be noted that the possibility for manipulation still presumes that the FSP is capable of manipulating the Delivery Point's offtake/injection in the opposite direction of the activation. In addition, the incentives for manipulating the baseline would need to be rather high, as a physical change in the offtake/injection prior to the activation tends to come with an additional cost. Particularly for historical baseline methodologies with a same-day adjustment that use a rather long adjustment period, the incentives for manipulation of the baseline might be considerably lower.
- Declarative baseline are in general susceptible to manipulation. However, as indicated before, the assessment assumes that the baseline would need to be submitted at least day ahead (prior to the DA market clearing). Under this assumption, the FSP can have an incentive to manipulate their submitted forecast in case activations (and thus, DA/ID market prices) could be anticipated prior to the moment the baseline needs to be submitted. The potential for manipulating the baseline can be mitigated by requesting the BSP to submit its baseline earlier (in combination with monitoring the bias of the baseline to avoid that an FSP systematically over/underestimates its offtake/injection). For instance, in the NEBEF mechanism in France, the baseline needs to be submitted minimally 2 days and up to one week in advance (and can only be changed up to 2 days in advance). However, increasing the performance with respect to integrity by requesting baseline submission multiple days in advance can go at the expense of a lesser performance in terms of inclusivity and accuracy, as it requires that the injection/offtake profile can be forecasted sufficiently well several days in advance.
- Historical baseline methodologies without same-day adjustment, calculated baseline methodologies, regression-based baseline methodologies and control group baseline methodologies score in general very high in terms of integrity due to the fact that the baseline calculation is difficult to manipulate. The reason that these baseline methodologies score slightly lower for ToE DA/ID results from the fact that the earlier trigger for DA/ID activations provide slightly more opportunities for manipulation compared to the mFRR product.

Facilitation of value stacking

The ease with which different baseline methodologies could support value stacking is discussed across all products in Section 3.3.5.

Conclusions

The key conclusions of the assessment of the different baseline methodologies for ToE DA/ID are listed below:

- In terms of the overall performance, it can be observed that historical baseline methodologies achieve the highest score for ToE DA/ID. Applying a same-day adjustment increases accuracy and inclusivity but comes at the expense of a reduced performance in terms of integrity. Compared to the historical baseline methodologies, declarative baseline methodologies again perform better in terms of inclusivity but are more complex and perform (significantly) less in terms of integrity. Although calculated, regression-based and control-group baseline methodologies achieve a similar performance as historical baseline methodologies with same-day adjustment in terms of accuracy; other criteria such as simplicity (and inclusivity) decrease the total performance of these baseline methodologies. Finally, due to a lack of accuracy for products with longer activations, MBMA baseline methodologies are not considered suitable for this service.
- Ensuring Integrity is challenging for ToE DA/ID, in particular for historical baseline methodologies with same-day adjustment and declarative baseline methodologies. Appropriate mitigation measures should be taken to limit opportunities for baseline manipulation and/or to detect baseline manipulation. In this regard, there are more mitigation opportunities for historical baseline methodologies with a same-day adjustment compared to declarative baseline methodologies. The main mitigation option to prevent a targeted manipulation of the baseline during activations when using declarative baseline methodologies is to request the baseline to be submitted well in advance. This however comes at the expense of the performance in terms of accuracy and inclusivity.

3.3.3 CRM

Table 8 provides an overview of the performance of the different baseline methodologies for the CRM.

The assessment of the performance of the different baseline methodologies for the CRM is identical to the performance assessment for ToE DA/ID. This results from the fact that these two products have similar product characteristics that could impact the performance the different baseline methodologies. First, the time between the request for activation/decision to activate and the start of the activation is assumed to be similar for both products as day-ahead prices form the trigger for activations for the CRM and can also form the trigger for DA/ID activations[22]. Second, the typical activation duration for both products is expected to be one to several hours (being significantly longer than typical balancing activations). Finally, for both products, the frequency of activations is expected to be low and focused on periods with high market prices. As such, we refer to Section 3.3.2 for a more detailed justification of the scoring of the different baseline methodologies.

[22] One difference exists between both products as for ToE DA/ID, the moment of activation can in principle be chosen freely by the FSP, whereas they cannot be chosen when participating to the CRM (activation is monitored in case the DA price exceeds the declared price). However, this did not impact the final assessment of the performance of the different baseline methodologies.

Furthermore, it has to be noted that the calculated baseline methodology is included here for completeness. However, such baseline methodologies are typically used for variable renewable generation assets in order to calculate the power that would have been generated from these assets in case they would not have been regulated downwards as part of an activation. As these assets would participate to the CRM based on capacity credits/derating factors, and downward activations (e.g., wind curtailment) cannot be valorized as part of a CRM, calculated baseline methodologies which are typically used for downwards activations of generation assets are not considered to be applicable for this product.

Table 8: Evaluation of the performance of different types of baseline methodologies for the CRM

Criteria	Final Scoring	Simplicity	Accuracy	Inclusivity	Integrity	Facilitation of value stacking
Weighting factor		**30%**	**35%**	**15%**	**15%**	**5%**
MBMA	2.9	5	1	3	3	3
Historical without same-day adjustment	3.4	4	3	2	4	4
Historical with same-day adjustment	3.7	4	4	3	3	3
Declarative	3.1	3	3	4	2	4
(Calculated)	3.0	2	4	1	4	4
Regression-based	3.0	1	4	3	4	4
Control group	3.0	2	4	1	4	4

3.3.4 Strategic reserves

Table 9 provides an overview of the performance of the different baseline methodologies for strategic reserves.

The assessment of the performance of the different baseline methodologies for strategic reserves is identical to the performance assessment for ToE DA/ID and the CRM. This results from the fact that these products have similar product characteristics that could impact the performance the different baseline methodologies. First, the time between the request for activation/decision to activate and the start of the activation is assumed similar for these products. Although the activation request comes later for strategic reserves in comparison to the CRM, there are still several hours between the activation request and the start of the activation. Second, the typical activation duration for

both products is of the same order of magnitude. Finally, for both products, the frequency of activations is expected to be low and focused on periods with high market prices (following near-scarcity situations). As such, we refer to Section 3.3.2 for a more detailed justification of the scoring of the different baseline methodologies.

One difference with the assessment of the performance of the baseline methodologies for the CRM and ToE DA/ID is that opportunities for value stacking is not considered to be a relevant assessment criterion for the performance of baseline methodologies for strategic reserves since this product targets assets that are outside the market and are not allowed to provide other services. The weight given to the other assessment criteria is therefore increased proportionally. This difference does not impact the conclusions.

Finally, it has to be furthermore noted that the calculated baseline methodology is included here for completeness. However, such baseline methodologies are typically used for variable renewable generation assets. However, given that these assets cannot participate to Strategic Demand Reserves, calculated baseline methodologies are not considered applicable for this product.

Table 9: Evaluation of the performance of different types of baseline methodologies for strategic reserves

Criteria / ***Weighting factor***	Final Scoring	Simplicity **32%**	Accuracy **37%**	Inclusivity **16%**	Integrity **16%**
MBMA	2.9	5	1	3	3
Historical without same-day adjustment	3.3	4	3	2	4
Historical with same-day adjustment	3.7	4	4	3	3
Declarative	3.0	3	3	4	2
Calculated	2.9	2	4	1	4
Regression-based	2.9	1	4	3	4
Control group	2.9	2	4	1	4

3.3.5 Considerations on facilitation of value stacking

Creating a market design that enables value stacking is far from straightforward. A full analysis of the market design for enabling value stacking is out of the scope of this study. However, when making recommendations on appropriate baseline methodologies, it is relevant to consider the possible impact of the choice of baseline methodology on the opportunities for enabling value stacking.

As will be illustrated below, most baselines methodologies will at least in certain situations (depending on the timing of the respective activations and the corresponding activation requests/decisions) have limitations resulting in inappropriate activation control/Transfer of Energy when different products/services are activated simultaneously or sequentially. However, an erroneous baseline calculation or settlement could be resolved via coordinated settlement/activation control/ToE mechanisms. The design and implementation of such coordinated mechanisms would require further effort and may form a motivation to prefer baseline methodologies that lead to fewer distortions and/or requiring less coordination. Taking into account these considerations, a relatively low weighting factor is taken for the assessment of the performance of a baseline methodology in terms of its ability to facilitate value stacking.

Illustration 1: mFRR activation during a longer activation in the context of DA/ID trades

In this example, it is assumed that the decision to perform an activation in the context of DA and/or ID trades precedes the mFRR activation request, and that the start of the DA/ID activation precedes the start of the mFRR activation[23]. In addition, the mFRR activation is assumed to be in the other direction than the activation in the context of DA/ID trades. This situation is illustrated in Figure 7, which shows the resulting load profile of a Delivery Point that has been activated for two different services at the same time: selling energy on the DA/ID market for a period of 4 quarter hours by reducing the net load of the Delivery Point, and providing downwards mFRR balancing energy during a single quarter hour by increasing the net load of the Delivery Point. In the absence of activations, the Delivery Point is assumed to be operated continuously at a level of 10 MW. As a result of DA/ID trading, the offtake is reduced to the level of 8 MW during 4 quarter hours, but is returned to the level of 10 MW during the third quarter hour of the DA/ID activation in order to provide downward mFRR balancing energy.

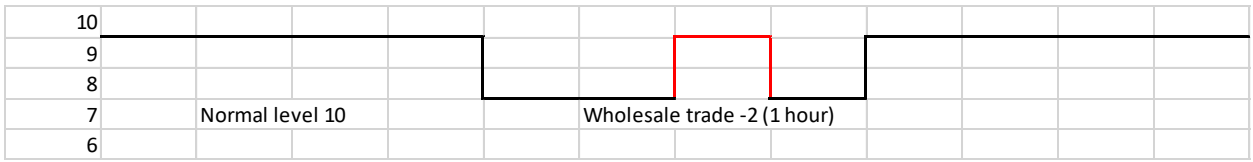

Figure 7: Illustration of an mFRR activation during an ongoing activation in the context of a DA/ID trade. In this illustration, the calculation of the baseline for the second activation (mFRR activation) should take into account the ongoing activation (DA/ID).

In this illustration, for the mFRR activation control, the baseline should in principle be set to a level of 8 MW since this would have been the operating level if no request for a downward mFRR activation would have taken place. MBMA (last QH), declarative baseline methodologies and historical baseline methodologies with a same-day adjustment (depending on the period used to calculate the baseline adjustment) are capable of taking the impact of the ongoing DA/ID activation into account in this specific example and hence could lead to a baseline of 8 MW. In contrast, historical baseline methodologies without SDA, calculated, regression and control group baseline methodologies would typically not take the ongoing DA/ID activation into account and would therefore in principle set the baseline at 10 MW.

In addition, for the ToE mechanism, adding up the delivered volumes of flexibility calculated for the DA/ID service and the mFRR service would be incorrect for certain baseline methodologies. Indeed, for the ToE mechanism, the total volume to be settled for the quarter hour of the combo should equal 0 MWh. Assuming that the baseline for the DA/ID

[23] A similar example can be made with an mFRR activation during an ongoing activation in the context of the CRM.

activation equals 10 MW (e.g., in case a historical or declarative baseline methodology would be used), the delivered volume of flexibility for the DA/ID service in the quarter hour of the combo would equal 0 MWh. For mFRR, the delivered volume of flexibility would equal -0.5 MWh (= -2MW * 0.25h) in case the baseline for mFRR would equal 8 MW (e.g., in case a MBMA (Last QH) or declarative baseline methodology would be used), and 0 MWh in case the baseline for mFRR would equal 10 MW (e.g., in case the historical baseline methodology would be used). As such, the baseline methodologies that provided the correct result for the activation control of mFRR in this illustration do not result in a correct ToE settlement and vice versa.

Based on this illustration, it can be concluded that some form of a coordinated/corrected settlement/activation control/ToE mechanism would be needed in order to have a correct activation control and ToE settlement.

Illustration 2: mFRR activation immediately following an activation in the context of DA/ID trades

Figure 8 shows a very similar situation as in the previous illustration, with the only difference that the mFRR activation only occurs in the quarter hour directly following the end of the DA/ID activation. In this situation, it seems reasonable that the baseline for the mFRR activation should be set at a level of 10 MW since this would have been the operating level if no request for a downward mFRR activation had taken place.

Figure 8: Illustration of an mFRR activation following an activation in the context of a DA/ID trade. In this illustration, the calculation of the baseline for the second activation (mFRR) should not be distorted by a preceding activation (DA/ID).

In this case, historical baseline methodologies without same-day adjustment, declarative, calculated, regression and control group baseline methodologies are all capable of setting the level of the baseline at 10 MW, since these are not impacted by the earlier DA/ID activation (either intentionally or coincidentally). In contrast, in this specific example, MBMA baseline methodologies and historical baseline methodologies with same-day adjustment (depending on the period used to calculate the baseline adjustment) would incorrectly set the level at 8 MW.

Conclusions

These two illustrations highlight that a given baseline methodology may result in a correct baseline value for one situation of combo/sequential activations and result in an incorrect baseline value for another situation. None of the considered baseline methodologies seems to be correct in all possible circumstances. In addition, the first illustration also highlighted that a baseline methodology achieving a correct value for the activation control does not necessarily lead to a correct value for the settlement of the Transfer of Energy. A possible option could be to use a unique baseline in case of a combo (e.g., impose to use the same baseline for both products or have a “master baseline” that is applied in case of a combo) that reflects the offtake/injection in case no activations would have taken place at all, in combination with a joint/coordinated activation control in which the total volume is allocated to the different products. However, an in-depth analysis of the possible designs of such coordinated mechanisms is out of the scope of this study.

Nevertheless, based on the examples provided, it can be concluded that declarative baseline methodologies perform good in almost all situations, MBMA and historical baseline methodologies with a same-day adjustment perform well in most situations, whereas historical baseline methodologies without same-day adjustment, calculated, regression and control group baseline methodologies are not directly capable of considering ongoing activations, and would require a correction via a coordinated settlement/activation control in the majority of situations.

4. Conclusions regarding the assessment of baseline methodologies

This section provides an overview of the main conclusions with respect to the baseline methodologies currently applied by Elia. These conclusions are based on the analysis of best practices (including the analysis on international practices, the literature review, and the interactions with market parties active in Belgium) and the assessment of the performance of the different baseline methodologies for the different products as performed by the consulting company DNV. The key conclusions are summarized below per product or group of products.

mFRR

Both the **High X of Y and the Last QH baseline methodology current applied are in line with best practices**. The analysis of international practices confirms that these types of baseline methodologies are used most commonly for these types of products. Combining a good performance in terms of accuracy and simplicity, MBMA and historical baseline methodologies are assessed to perform well for these products. This was also reflected in the stakeholder interactions, where market parties active in Belgium indicated to be generally satisfied with the Last QH and High X of Y baseline methodologies for their current portfolios of Delivery Points.

However, market parties did express concerns that the current baseline methodologies may not be sufficiently accurate for technologies they might want to add to their pool in the future, such as variable renewables. The literature and the performance assessment in this regard confirm that MBMA and historical baseline methodologies can be insufficiently accurate for a minority of assets that have a (highly) variable and irregular offtake/injection pattern, such as wind and solar PV generation, variable operated batteries/generation units or demand facilities. In this regard, the assessment also indicates that declarative and calculated baseline methodologies could be relevant to consider in order to reduce barriers to participation and to provide more options for FSPs.

Based on the above observations, the following **recommendations** are made for mFRR:

- **Maintain the Last QH and High X of Y baseline methodologies**.
- **Analyze the possibilities** to reduce barriers for participation of Delivery Points with a variable and irregular offtake/injection pattern in general and wind and solar PV in particular. These possibilities are further investigated in Section 5 and 6 respectively.

ToE DA/ID / CRM / Strategic reserves

The **High X of Y(*) baseline methodology currently applied is in line with best practices**. The literature provides clear evidence that historical baseline methodologies can achieve a high accuracy for a large variety of assets, also during periods with the highest loads/prices for which activations of these products is most likely. In addition, the analysis of international practices reveals that historical baseline methodologies are used most commonly for these types of products (that are characterized by relatively long activation durations and a relatively long period between the request/decision to activate and the start of the activation). Combining a good performance in terms of accuracy

and simplicity, historical baseline methodologies are assessed to perform well for these products. This was also reflected in the stakeholder interactions, where market parties active in Belgium indicated to be generally satisfied with this baseline methodology and did not provide suggestions for further improvements.

Different variants of historical baseline methodologies exist. Historical baseline methodologies differ mainly in the selection of the representative days (e.g., the parameters X and Y in the High X of Y(*) baseline methodology) and whether or not a same-day adjustment is applied. In this regard, different quantitative assessments can be found in the literature that conclude that **applying a same-day adjustment significantly increases the accuracy**, whereas other parameters (such as the values of X and Y) have a minor impact on the overall performance (in particular in case a same-day adjustment is applied). However, applying a same-day adjustment introduces additional opportunities for manipulating the baseline that need to be mitigated.

Although market parties were generally satisfied with the High X of Y(*) baseline methodology, one **market party questioned the inclusivity of this baseline methodology**. In this regard, the literature also indicates that for a minority of assets, namely those with a (highly) variable and irregular offtake/injection pattern, historical baseline methodologies can be insufficiently accurate, even if a same-day adjustment is applied. In stakeholder interactions during previous studies, market parties had also expressed an interest to obtain more freedom to choose between different baseline methodologies and/or to develop or propose their own methodology for calculating the baseline in case the existing baseline methodology would not be suitable. The assessment of the performance of different baseline methodologies in this regard indicates that **declarative baseline methodologies seem the best option for an alternative choice of baseline methodology for those assets for which a High X of Y* baseline would not perform well.** Moreover, **introducing an additional declarative baseline methodology could** more generally **be a suitable solution for offering FSPs full flexibility for the calculation of the baseline**.

Based on the above observations, the following **recommendations** are made for ToE DA/ID / CRM / strategic reserves:

- **Maintain the High X of Y and High X of Y* baseline methodologies**. Moreover, given the significant added value of applying a same-day adjustment, Elia proposes to continue to offer FSPs the choice to apply a same-day adjustment for the High X of Y* baseline methodology and to develop dedicated monitoring to mitigate the possibilities for baseline manipulation.
- **Analyze the possibilities** to reduce barriers for participation, in particular for assets with a highly variable and irregular offtake/injection pattern, by **providing flexibility to FSPs to propose or use their own baseline methodology**. These possibilities are further investigated in Section 5.

aFRR

Although aFRR was not in the planned scope of the study, some findings related to aFRR are considered following stakeholder interactions. In this regard, the international practices reveal that both MBMA and declarative baseline methodologies are used frequently for aFRR. As such, the current declarative baseline methodology applied in Belgium is in line with common practices. In addition, market parties indicated to be generally satisfied with the current baseline methodology for their current portfolios of assets.

However, market parties have also indicated that the current declarative baseline methodology may prove insufficiently accurate for renewable generation assets (notably wind turbines and solar PV), resulting in barriers for participation. In this regard, market parties suggested that calculated baseline methodologies could form a suitable alternative.

Based on these observations, the following recommendations are made for aFRR:

- **Maintain the current declarative baseline methodology**
- **Analyze the need and potential for suitable baseline methodologies for renewable generation assets**. These possibilities are further investigated in Section 6.

Part B: Opportunities for alternative baseline methodologies

5. Possibilities for market parties to choose or propose their own baseline methodology

As concluded in Section 4, the baseline methodologies currently applied for the products mFRR, ToE DA/ID, the CRM and strategic reserves are capable of achieving a high accuracy for the majority of assets. However, for a minority of assets that have a (highly) variable and irregular offtake/injection pattern, these baseline methodologies might not be sufficiently accurate. From the assessment of best practices and the interaction with market parties, declarative baseline methodologies, in which FSPs obtain full flexibility for calculating the baseline, are put forward as a suitable solution for such assets. In addition, in previous studies, market parties expressed a desire to have more flexibility in their choice of baseline or to propose their own baseline methodology.

In this context, this section analyzes the possibilities of reducing barriers to participation by enabling the FSP to propose or use alternative methodologies for calculating the baseline. First, Section 5.1 looks at different approaches to enable more flexibility to FSPs in terms of baselining. Next, Section 5.2 provides an overview of international experiences with the use of declarative baseline methodologies. The consulting firm DNV supported Elia on gathering the experiences from international practices. Finally, 5.3 presents the proposal for implementation.

5.1 Options for enabling flexibility for FSPs to choose or propose a baseline methodology

In general, two approaches can be imagined for enabling FSPs more flexibility in terms of the baseline calculation.

A first option, raised during discussions regarding the design and implementation of the Transfer of Energy mechanism for DA and ID markets, is to develop a dedicated process for the proposal, approval and introduction of new baseline methodologies. The key idea is that if this process would be regulated, FSPs could propose new baseline methodologies that can be validated and implemented within a relatively short period without having to adapt different regulated documents (e.g., the Rules on the organization of the Transfer of Energy and the T&C BSP mFRR) each time a new baseline methodology is implemented.

However, Elia is of the opinion that the baseline methodology is an essential part of a contract, as it impacts some of the most fundamental contractual elements including prequalification, settlement and activation control, and therefore should be unambiguously described in the (regulated) contract. In addition, any process that could be imagined will always consume a significant amount of time considering that the FSP would need to elaborate its proposal, the proposal would subsequently need to be analyzed and discussed between the system operator, market parties and the relevant regulatory authorities and finally implemented in the IT systems. As such, this will lead to significant delays for FSPs. These delays would be further reinforced in case the regulated contracts would need to be adjusted. More-

over, a process that involves an evaluation and possible an implementation on a relatively short term after the request of the FSP would imply that resources would need to be committed on the short term to analyze and possibly implement a proposed baseline methodology. Such a process would therefore make it impossible to do a prioritization and would thus involve a risk that efforts (from Elia and stakeholders) need to be focused on projects that may not be in line with the priorities in terms of market design evolutions commonly agreed between Elia, market parties and regulatory authorities.

The second option is to introduce a declarative baseline methodology, implying that the FSP calculates the baseline himself (following an approach of his choice) and submits the baseline to the system operator. In this option, the regulated documents would need to be changed once in order to unambiguously describe the conditions for FSPs to be able to use the declarative baseline methodology, the process of the baseline submission, etc. In addition, a one-time implementation effort is needed to enable the required data exchanges between the FSP and the system operator. However, after this one-time implementation, an FSP could use a method of its choice for calculating its baseline without significant delays. Moreover, this approach would provide more flexibility to FSPs, for instance to use tailor-made or complex baseline methodologies that would be difficult to generalize to a standard baseline methodology.

Considering the elements above, Elia is of the opinion that the best way to enable FSPs to have more flexibility in terms of the methodology to calculate the baseline is via the introduction of a declarative baseline methodology. As such, this could provide a solution for assets with a (highly) variable and irregular offtake/injection pattern while also providing opportunities for all assets to further improve the baseline quality.

It must be noted that a possible introduction of a declarative baseline methodology does not imply that no new baseline methodologies could be introduced in the future. For instance, if a return-of-experience would show that certain methodologies are used commonly by market parties to calculate their baseline that they submit to Elia through the declarative baseline methodology, a centralized implementation of such a baseline methodology could be considered.

5.2 International experience with declarative baseline methodologies

This section provides an overview of the relevant international experience on the use of declarative baseline methodologies, based on interviews with different system operators and market parties. The interviews were organized by the consulting firm DNV and were held with National Grid Electricity System Operator (NG ESO) (GB), CBS (experience in GB), RTE (experience with NEBEF in France), Scholt (experience with Alliander and TenneT in the Netherlands) and Enel X (overall experience with declarative baselines with a focus on GB). It should be noted however, that international experience with declarative baselines is still rather limited.

5.2.1 Application

In Great Britain, the declarative baseline methodology is the default baseline methodology[24] for flexibility providers with a so-called Primary Balancing Mechanism Unit (BMU)[25] to participate to the different balancing products, including the Balancing Mechanism[26], the dynamic containment (DC) product and the short term operating reserves (STOR) product[27].

In France, it is possible to choose between different baseline methodologies, including a declarative baseline methodology as well as MBMA and historical baseline methodologies. Declarative baseline methodologies[28] can be used within NEBEF (wholesale load reduction product) and in the balancing mechanism for telemetered sites that prequalify to use this baseline. For NEBEF, about 30% of the participating sites uses the declarative baseline methodology. For the balancing mechanism, the practical experience on the use of declarative baseline methodologies is limited because FSPs do not choose this method (possibly due to the high accuracy requirements for participation to balancing; see below).

In the Netherlands, Alliander uses a declarative baseline methodology for DSO congestion management and TenneT uses a declarative baseline for the aFRR product that is similar to the declarative baseline methodology used in Belgium for aFRR.

5.2.2 General return of experience

The overall return of experience on the use of declarative baseline varies. Most stakeholders (i.e., the TSOs and market parties interviewed regarding their experience with declarative baseline methodologies) suggested that the main reason to use a declarative baseline methodology is as an alternative to existing baseline methodologies that might not provide sufficient accuracy for certain assets. More specifically stakeholders mentioned that declarative baseline methodologies could be useful for:

- assets with a variable but irregular offtake/injection pattern, such as dispatchable generation (CHPs), dispatchable storage (batteries) and dispatchable load (electrolysers, e-boilers, possibly other load with predictable profiles), but also limitedly controllable assets (depending on the moment the baseline would need to be submitted and the ability to forecast the offtake/load)
- assets/products with very frequent and relatively long activations, for which historical baseline methodologies might not be suitable due to the exclusion of reference days with activations, and MBMA baseline methodologies might not be suitable due to the long duration of the activations.

When the interviewed stakeholders were asked to reflect on the performance of declarative baseline methodologies in relation to the criteria used for the assessment of the performance of the baseline methodologies (i.e. accuracy, inclusivity, integrity, simplicity and facilitation of value stacking, (see Section 3.1 for a more detailed description of

[24] This declarative baseline is called Physical Notification and is provided at the main meter level (BMU level).
[25] Primary Balancing Mechanism Units (BMUs) are the units used under the Balancing and Settlement Code (BSC) to account for all energy that flows in or out the Total System (the Transmission System and each Distribution System combined). A Primary BMU is the smallest grouping of generation and / or demand equipment that can be independently metered for settlement and all generation and demand equipment must be captured in a Primary BMU. https://www.elexon.co.uk/operations-settlement/balancing-mechanism-units/
[26] As of this year, providers of demand-side flexibility that participate without a BMU are provided the option to use an historical baseline methodology as an alternative to the declarative baseline methodology.
[27] Providers of demand-side flexibility that participate without a BMU used to use an historical baseline methodology, but recently had to start using the declarative baseline methodology.
[28] Referred to as “méthode par prévision de consommation”.

these criteria), the stakeholders indicated that declarative baseline methodologies are considered more complex (and costly) for the FSP to implement and operate compared to alternative baseline methodologies such as MBMA and historical baseline methodologies. In addition, declarative baselines were considered by the interviewed stakeholders to score relatively low in terms of integrity, although all interviewed stakeholders mentioned that there are ways to mitigate strategic behavior through monitoring and validation processes (see below). For these reasons, most stakeholders see declarative baseline methodologies therefore as an additional baseline methodology that can be used when the default baseline methodolog(y)(ies) cannot provide sufficient accuracy for specific assets.

5.2.3 Validation of the baseline

The practices with respect to prequalification/validation of the submitted baselines differ strongly between different systems and products.

In Great Britain, the accuracy of the so-called Physical Notifications that serve as a baseline is not tested as part of the prequalification nor during operations (although discussions on this topic are ongoing). Similarly, in the Netherlands, TenneT does not apply any prequalification requirements on the declarative baselines used for aFRR[29] and does not describe a systematic monitoring process on the accuracy of the submitted baselines. However, TenneT does reserve the right to check the baselines or to have an audit by an independent party to check the quality of the baseline.

In contrast, in France, for each site that requests declarative baselining, an accuracy testing process is conducted by RTE. The main accuracy requirement is identical for the declarative and historical baseline methodologies, but differ per product type. For NEBEF, the absolute error must be less than or equal to 40%[30], whereas for the balancing mechanism this must be less than or equal to 10%[31]. In addition, for declarative baseline methodologies, an additional accuracy requirement is introduced to avoid a bias of the submitted baselines. The metric used in this regard is the so-called centering error[32]. For NEBEF, the centering error must be less than or equal to 15%, whereas for the balancing mechanism it must be less than or equal to 3%.[33] In addition, the accuracy and bias of the submitted baselines is monitored on a monthly basis using the same criteria.

5.2.4 Process for submission of the baseline

The practices regarding the timing of the submission of the baseline vary significantly across different markets/products. Nevertheless, one common element is that the baseline needs to be submitted by the FSP before the activation request, or before the FSP has knowledge on the volume requested for activation.

In Great Britain, the baseline (Physical Notification) needs to be submitted one hour ahead. Although there are significant differences between the different products, this is the case for each of the balancing products. For example, for

[29] However, with the new aFRR requirements, prequalification of the baseline will be required when the FSP proposes a deviation from the preferred declarative baseline, i.e., the 1 minute ahead declarative baseline.

[30] RTE defines the absolute error as normalised mean error for 10 min time steps: $Absolute\ error = \frac{1}{Number\ of\ steps\ (N)} \times \sum_i^N \frac{|Baseline_i - Measurement_i|}{Flexible\ capacity_i}$

[31] The criteria for what is considered an “acceptable”, “good” or “very good” accuracy varies. The literature suggests that a relative root mean square error of 10 per cent or less is generally considered to be ‘good’, and a value between 10 and 20 per cent is considered ‘acceptable’.

[32] RTE defines the centering error as the normalised mean bias for 10min time steps: $Centering\ error = \frac{1}{Number\ of\ steps\ (N)} \times \left| \sum_i^N \frac{Baseline_i - Measurement_i}{Flexible\ capacity_i} \right|$

[33] See **link** and **link**.

the dynamic containment product, the activation requests happens only one second prior to the activation, whereas for the short-term operating reserves product, the activation request takes place 20 minutes ahead of the activation.

In the NEBEF and the balancing mechanism in France, the baseline needs to be submitted one week in advance[34] and FSPs are afterwards allowed to change their forecast up to two days in advance (and this maximally 4 times per month). The submission well ahead of real-time is introduced as mitigation measure to avoid manipulation of the baseline.

In the Netherlands, the FSP needs to submit the declarative baseline (referred to as “D-prognosis”) day-ahead to the DSO Alliander, whereas TenneT requires for aFRR that the FSP submits the baseline one minute ahead of real-time with a 4-second resolution.

5.2.5 Mitigation of baseline manipulation

As indicated before, the risk of baseline manipulation was considered by several of the TSOs and market parties interviewed regarding their experience with declarative baseline methodologies as an important point of attention when using declarative baseline methodologies. However, practices with respect to measures to mitigate these risks differ depending on the product characteristics and the design of the declarative baseline methodology (in particular, the moment the baseline needs to be submitted).

In Great Britain, considering that the baseline needs to be submitted one hour in advance and that balancing activations are difficult to predict, no further rules or mitigation actions are implemented. However, audits by independent technical engineers can be performed occasionally to verify that the software that has been used to make the forecasts has not been manually changed/manipulated. Similarly, TenneT does not prescribe any systematic monitoring process on the declarative baselines, but reserve the right to check the baselines or to set an audit by an independent party to check the quality of the baseline.

In France, the main measure to mitigate opportunities to manipulate the baseline is again to request the baseline to be submitted well in advance (minimally two days in advance). Considering that in the NEBEF mechanism, the moments of activations are chosen by the FSP based on market opportunities that can be known well ahead of real-time, the baseline is requested to be submitted significantly longer in advance in comparison to the hour-ahead baseline submission used in Great Britain and the 1-minute ahead baseline submission used by TenneT for aFRR. In addition to the moment of baseline submission, an additional check is performed to prevent FSPs to structurally bias their forecast to their benefit.

5.3 Proposal for implementation

[34] Specifically, the baseline needs to be submitted on Friday for the entire upcoming week.

Based on the sections above, Elia acknowledges that adding a declarative baseline methodology may reduce barriers for participation for assets with a highly variable offtake/injection pattern for which the current baseline methodologies might not be sufficiently accurate as well as more generally to enable more flexibility to FSPs to use a suitable method for calculating their baseline. For these reasons, Elia is willing to consider developing a declarative baseline methodology for the products ToE DA/ID, CRM and mFRR. However based on discussion with market parties, Elia does have questions regarding the current need for introducing a declarative baseline methodology. In this regard, Elia welcomes market parties to provide indications and specific use cases for which a declarative baseline methodology would be needed to reduce entry barriers.

Given that both the product characteristics[35] and the incentives for baseline manipulation significantly differ between mFRR on the one hand, and ToE DA/ID and the CRM on the other hand, the proposed design of a declarative baseline methodology for both products is also different.

5.3.1 Design of the declarative baseline methodology for ToE DA/ID and the CRM

Baseline submission and mitigation of baseline manipulation opportunities

Given that activations for these products are (or can be) triggered by the day-ahead electricity prices, and that periods with elevated day-ahead prices can to some extent be forecasted, there is a risk that the submitted baseline would be manipulated to the benefit of the FSP (e.g., by increasing the forecast of the offtake).

In order to minimize manipulation opportunities, Elia proposes to request the baseline to be submitted minimally two days prior to the start of the activation. In addition, Elia proposes to monitor systematic deviations between the submitted baseline and the measured offtake/injection (i.e., the bias) in moments for which the baseline has been submitted but no activations were requested or performed[36], and to apply penalties in case such systematic deviations are detected in the direction benefiting the FSP for which no sound justification can be provided. Finally, as part of the validation of the baseline (baseline prequalification), FSPs will need to provide a description of the process for calculating the baseline, including the inputs used for the calculation.

Validation of the baseline

Elia proposes to allow the FSP to select its preferred baseline methodology for each Delivery Point DP_{PG}. However, considering that when using a declarative baseline methodology the accuracy of the baseline submitted is dependent on the processes used by the FSP to calculate the baseline, and that there are potential opportunities for manipulating the baseline, Elia proposes that a validation/prequalification of the baseline is needed before an FSP can make use of the declarative baseline methodology. Such a baseline prequalification would involve that it should be proven during a test period that the baseline submitted by the FSP is more accurate than the default baseline methodology (High X of Y*). In addition, the FSP needs to provide a description of the method and inputs used for calculating the declarative baseline methodology.

[35] In particular, the time between the activation request/decision to activate and the start of the activation is much shorter for mFRR than for ToE DA/ID and the CRM, see Section 3.2.

[36] For ToE DA/ID, FSPs would not be requested to submit a baseline for each quarter hour of the year. However, in case the declarative baseline is chosen, it would not be possible to perform an activation for moments for which no baseline has been submitted.

5.3.2 Design of the declarative baseline methodology for mFRR

Baseline submission and mitigation of baseline manipulation opportunities

For mFRR, it has to be noted there is de-facto already a declarative baseline methodology in place for assets providing MW schedules (i.e., the DP_{SU}). Indeed, the MW schedules sent to Elia (in the framework of the T&C SA) are used as the basis for quantifying the delivered flexibility for mFRR (as specified in the T&C BSP mFRR) and can thus be seen as a declarative baseline methodology.

In this regard, an implementation of an additional declarative baseline methodology is only relevant for those assets that do not provide MW schedules in the framework of the T&C SA (and that do not have a scheduling obligation). More specifically, Elia proposes[37] that:

- for Technical Units that provide a MW schedule in the framework of the T&C SA, the MW schedule remains to be used for the calculation of the delivered flexibility for mFRR
- for Technical Units that do not provide a MW schedule, either the MBMA (Last QH), historical (High X of Y) or a declarative baseline methodology can be chosen (at least, under certain conditions). One exception relates to Technical Units that, as of ICAROS phase 2, will have an obligation to provide either ON/OFF or MW schedules in the context of the T&C SA (i.e., SPGM/PPM/ESD of Type B with a nominal power between 1 and 25 MW). These units either have the option to offer MW schedules (in this case, the MW schedule is used automatically as the baseline) or to submit ON/OFF schedules in which case either the MBMA (Last QH) or the historical (High X of Y) baseline methodology can be chosen for mFRR[38].

With respect to mitigation of baseline manipulations, it has to be noted that in contrast to ToE DA/ID, mFRR activations are requested by Elia and are difficult to anticipate as long as the baseline needs to be submitted a certain time before the activation request. As a result, a selective manipulation of the submitted baseline during moments of activations is not considered feasible. Manipulation of the baseline is therefore most likely to occur by submitting structurally biased declarations.

In order to avoid manipulation of submitted baselines, Elia therefore proposes that:

- the submission/final update of the baseline needs to happen prior to the activation request (e.g., 45 minutes before the start of the quarter hour).
- A monitoring is performed on systematic deviations between the submitted baseline and the measured offtake/injection during quarter hours for which mFRR balancing energy bids have been submitted but no activation has been requested. Penalties will apply in case systematic deviations are observed that are in the direction benefiting the BSP and for which no sound justification can be provided.

Validation of the baseline

[37] The current proposal is based on the current design proposals for mFRR/ICAROS. In case of evolving mFRR/ICAROS designs, the current proposal could need to be adapted.

[38] The exact design for DP_{PG} that provide ON/OFF schedules could change depending on the design that will be further elaborated for phase 2 of the ICAROS project.

Similar as for ToE DA/ID, Elia proposes to allow the FSP to select its preferred baseline methodology for each Delivery Point DP_{PG}, but to require a validation/prequalification of the baseline before an FSP can make use of the declarative baseline methodology. Again, the baseline prequalification would involve that during a test period, the submitted baseline needs to be more accurate than the default baseline methodologies for mFRR (i.e., Last QH and High X of Y). Given the limited risk of baseline manipulation for this product, it is not considered necessary to require transparency on the method used by the FSP to calculate the baseline.

6. Baselining options for wind and solar PV

As discussed in Section 2.4, market parties indicated that the current baseline methodologies for aFRR (i.e., the declarative baseline methodology in which the baseline needs to be submitted one minute before real-time for each 4-second interval) and mFRR (i.e., MBMA and historical baseline methodologies)[39] might not be sufficiently accurate for variable renewables such as wind and solar PV. In this context, this section addresses possible baselining options for wind and solar PV.

In general, two different approaches can be taken with respect to baselining for wind and solar PV:

1. forecasting the wind or solar PV generation in case no downward activation/curtailment would be requested hours to minutes ahead of real-time (declarative baseline methodology);
2. calculating the wind or solar PV generation that would have taken place in case no downward activation would have taken place (i.e., the so-called “Available Active Power” or “AAP”) based on real-time measurements such as wind speed or solar irradiance (calculated baseline methodology).

Section 6.1 looks into the accuracy that can be achieved via these two different approaches for the products aFRR and mFRR. Next, Section 6.2 provides an overview of insights from international practices with respect to baselining for wind and solar PV. The consulting firm DNV has supported Elia in gathering international experience and in performing the literature review. Finally, Section 6.3 describes the conclusion and the proposal for implementation.

6.1 Accuracy of forecasts and Available Active Power calculations

This section focuses on the accuracy of different forecasting methodologies and methodologies for calculation of the AAP for wind and solar PV. Elements impacting the accuracy that are relevant to consider for possible baseline methodologies, such as the calculation method, the time resolution of the forecast/calculation, the forecast lead time, etc. are considered.

In general, there are different methodologies to calculate the AAP for wind and PV. For instance, it is common practice that wind turbine manufacturers provide a ‘possible power’ or AAP calculation integrated in the wind turbine SCADA system that is executed in real-time. The methodology used by manufacturers is typically subject to confidentiality or copyrights and is therefore either unknown or cannot be easily replicated. The AAP is typically based on actual (measured) wind speed or solar irradiation (or measurements of reference inverters), and is therefore only available in real-time or ex-post. Under non-curtailed conditions, the AAP normally closely follows the actual power output. Under curtailed condition, the AAP indicates the active power that a wind farm/PV plant would have produced if it had

[39] it is relevant to highlight that baselining for wind and solar PV is only relevant for the balancing products (in this case aFRR and mFRR). This because:
- wind and solar PV can participate to the CRM directly via de-rating factors (capacity credits);
- wind and solar PV are not allowed to participate to Strategic Demand Reserves;
- Transfer of Energy (DA/ID) is restricted to Delivery Points with a net offtake on an annual basis.

not been curtailed (downward activation), and can therefore be used as the baseline for wind and PV providing downward flexibility.

6.1.1 Wind power

In general, the accuracy of wind power forecasts depends among others on the forecast horizon, the forecasting resolution, the level of aggregation (spatial scale) and the forecasting methodology[40]. In this regard, the literature provides clear evidence that increased spatial aggregation and shorter forecasting horizons improve the accuracy of the forecast.

For mFRR, a review of the literature indicates that the forecasting errors for forecasts with a lead time of about one hour and a resolution of roughly 15 minutes (i.e., parameters that are relevant for a possible declarative baseline methodology for mFRR) can have an order of magnitude in the range of less than 1% up to around 8-10% depending on the methodology used and the level of aggregation (i.e., single wind turbine versus the aggregate of one or several wind farms)[41]. Note that for participation to mFRR, the relevant geographical scale is that of a wind farm (PPM). Aside from forecasting methodologies based on advanced algorithms, even simple persistence forecasting methods, can achieve good results on these time scales. Persistence forecasting methods assume that the power/wind speed in the future is equal to the measured power/wind speed in the present. As such, the measured power output/wind speed is used to forecast the future power output/wind speed. Due to its simplicity and high accuracy for short time horizons, the persistence forecasting method has been widely used for forecasts with such horizons and is often used in the literature as an accuracy benchmark for more advanced forecasting methodologies[42]. In this regard, a recent study from a collaborative working group of International Energy Agency (IEA) on wind shows (qualitatively) that the persistence forecasting method is sufficiently accurate for a forecast length of around 60 minutes and time resolutions of 15 minutes[43]. Another study which has looked into the accuracy of the persistence method shows that for the first hour the normalized root mean square error of the persistence method is around 5% on a 10-minute resolution[44]. Elia therefore concludes that the literature provides clear evidence that a declarative baseline (short-term forecasts) allows achieving a sufficiently accurate baseline for a product such as mFRR[45].

The baselining challenge for wind generation participating to aFRR is quite different. On the one hand, even more close to real-time forecasts can be used given that the current declarative baseline methodology for aFRR requires the baseline to be submitted only 1 minute before real time. On the other hand, forecasts need to be made on a much

---

[40] Minute-Scale Forecasting of Wind Power—Results from the Collaborative Workshop of IEA Wind Task 32 and 36, Energies 2019, 12(4), 712; https://doi.org/10.3390/en12040712

[41] See for instance:

- Comparison of models for wind speed forecasting, J.C. Palomares-Salas, J.J. G. de la Rosa, J.G. Ramiro, J.Melgar, A. Agüera and A. Moreno, **http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.640.7900&rep=rep1&type=pdf**.
- Giebel, G., Brownsword, R., Kariniotakis, G., Denhard, M., & Draxl, C. (2011). The State-Of-The-Art in ShortTerm Prediction of Wind Power: A Literature Overview, 2nd edition. ANEMOS.plus. **https://doi.org/10.11581/DTU:00000017**
- Rishabh Abhinav et al. / Energy Procedia 142 (2017) 455–460, **https://www.sciencedirect.com/science/article/pii/S1876610217357995,** NN and WNN are two different neural network wind forecasting methods
- Yongqian Liu et al. Quantitative method for evaluating detailed volatility of wind power at multiple temporal-spatial scales. Volume 2 Number 4 August 2019 (318-327), 10.1016/j.gloei.2019.11.004. **https://www.researchgate.net/publication/337796756_Quantitative_method_for_evaluating_detailed_volatility_of_wind_power_at_multiple_temporal-spatial_scales**
- Available Active Power Estimation for the Provision of Control Reserve by Wind Turbines, **http://publica.fraunhofer.de/eprints/urn_nbn_de_0011-n-3010478.pdf**

[42] Energies 2020, 13, 3764; doi:10.3390/en13153764 **https://www.researchgate.net/publication/343140492_A_Critical_Review_of_Wind_Power_Forecasting_Methods-Past_Present_and_Future**

[43] Minute-Scale Forecasting of Wind Power—Results from the Collaborative Workshop of IEA Wind Task 32 and 36, Energies 2019, 12(4), 712; **https://doi.org/10.3390/en12040712**

[44] Available Active Power Estimation for the Provision of Control Reserve by Wind Turbines, **http://publica.fraunhofer.de/eprints/urn_nbn_de_0011-n-3010478.pdf**

[45] Recall from Section 5.3 that most wind farms have (>25 MW) or will have (1-25 MW) a schedule obligation and therefore can use the MW schedule to serve as a baseline.

higher resolution (i.e., for each 4-second period). The literature indicates that there are no mature forecasting techniques available for such high resolutions[46]. In addition, the literature does provide indications that the fluctuations of the wind speed and the power output of a wind farm can be relatively large within periods of several seconds to one minute[47]. Therefore, it can be concluded that the literature supports the indications provided by market parties that for (some) wind farms it may be difficult to achieve a high accuracy in case the baseline needs to be submitted one minute ahead of real time[48,]. This may be the case in particular for smaller wind farms as these short-term fluctuations are reduced (smoothed out) with an increasing size of the wind farm.

An alternative to the forecasts (declarative baseline) would be to use a calculated AAP signal as a baseline (calculated baseline). The accuracy levels that can be achieved by various methodologies to calculate the AAP is analyzed in the literature, where for 3-second intervals, normalized root mean square errors ranging from 2.8% to 5.3% are reported depending on the methodology that is used to calculate the AAP[49,50].

However, for certain methods that can be used to calculate the AAP, the accuracy can significantly decrease when the wind turbines are regulated down/curtailed in case so-called wake effects are not considered in the calculation of the AAP[51]. This is highlighted in Figure 9, which shows the accuracy of different methods to calculate the AAP both under normal operating conditions as well as under curtailed conditions.

[46] For example, a recent study from a collaborative workshop of the International Energy Agency (IEA) on wind provides a thorough summary of forecasting methods with lead times of one to several minutes but none of these forecasting methods provides forecasts on a resolution of one to several seconds. This study also highlights that forecasting methods with short lead times and/or high resolutions are still under development, as most of them are still in research, pilot or trial phase. Reference: Minute-Scale Forecasting of Wind Power—Results from the Collaborative Workshop of IEA Wind Task 32 and 36, Energies 2019, 12(4), 712; **https://doi.org/10.3390/en12040712**

[47] A study from the National Renewable Energy Laboratory (NREL) has looked into the ramping rates and step changes of the power output of three different wind farms varying in size between 100 MW and 220 MW. The distribution of ramping rates shows that for 15- second intervals, 96.8% of the ramps are within ±100 kW/s and for 5-minute intervals, 90% of the ramping rates are within ±780 kW/min (or 13 kW/s). This implies that for a wind farm of 100 MW, the power output would vary more than 1,5 MW within a 15-second time frame in 3,2% of the time, and more than 3,9 MW within a 5-minute time frame in 10% of the time. Reference: Wind Farm Power Fluctuations, Ancillary Services, and System Operating Impact Analysis Activities in the United States, **https://www.researchgate.net/publication/228909710**

[48] In the absence of mature advanced forecasting techniques on these time scales, the accuracy that a declarative baseline methodology in which the baseline needs to be submitted one minute ahead of real time could achieve can be estimated under the assumption that a simple persistence forecast technique would be used, i.e., assuming that the measured power of a wind farm in the current 4-second interval would be used as a forecast for the 4-second interval one minute later. In this regard, the accuracy depends on the volatility of the wind farm power output within these time frames. Moreover, it must be noted that it is difficult to use the measured power as a forecast during activations due to the fact that the measured power output is no longer representative of the power output in case no activation would have been requested.

[49] Available Active Power Estimation for the Provision of Control Reserve by Wind Turbines, D. Schneider et al., 2013, **http://publica.fraunhofer.de/eprints/urn_nbn_de_0011-n-3010478.pdf**

[50] ReWP - Regelenergie durch Wind- und Photovoltaikparks, Fraunhofer **https://www.iee.fraunhofer.de/content/dam/iee/energiesystemtechnik/de/Dokumente/Projekte/20170814_ReWP_Abschluss_final.pdf**

[51] More specifically, as a result of the downward activation, each wind turbine withdraws less energy from the wind and therefore the wind speed downstream of the wind turbines is reduced to a lesser degree. The result is that downstream wind turbines will measure higher wind speeds compared to a situation where no downward activation would have taken place in the upstream wind turbines. These higher measured wind speeds could lead to an incorrectly high estimation of the power that could have been generated on the downstream wind turbine (and hence the wind farm). This can typically occur in case the calculation of the AAP would be based on the measured wind speed/conditions instead of the wind speed/conditions that would have been measured if the upstream wind turbines would not have been activated downwards (i.e., if wake effects are not considered).

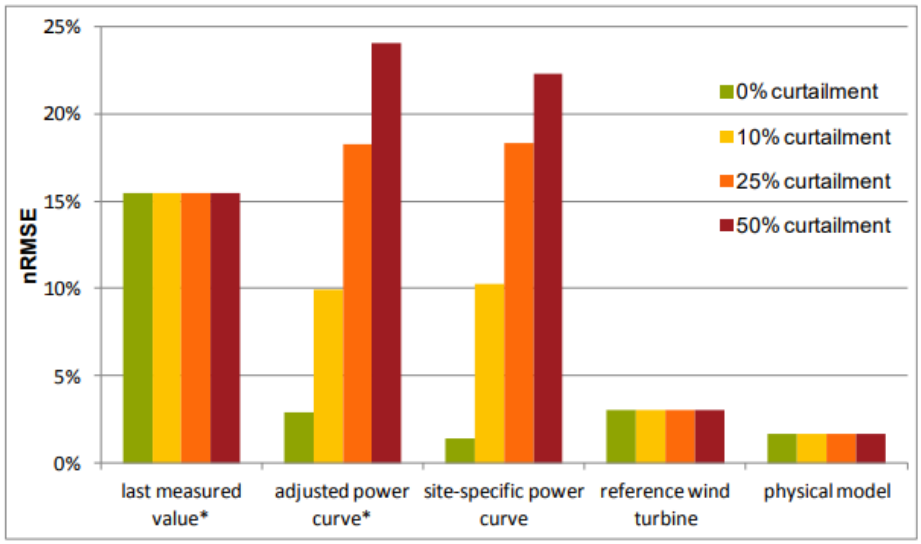

Figure 9: Accuracy of different methods to calculate the AAP[49]

Based on information provided in this section it can be concluded that it might be challenging to achieve a sufficiently high accuracy using short-term forecasts and that a calculated baseline methodology using an AAP signal could provide a sufficiently high accuracy. Achieving a sufficient accuracy at moments the wind farm is regulated downward (curtailed) is an important point of attention, but when appropriate methods are used, a sufficient accuracy can be achieved also during curtailment.

6.1.2 Solar PV installations

There is less research available regarding AAP calculations for solar PV. Nevertheless, the literature contains several examples of methods of AAP calculations, e.g., based on meteorological data or reference inverters[52]. The accuracy has been calculated for certain projects under normal operating conditions. Using methods based on real-time irradiance measurements, the accuracy reported for resolutions similar to that applicable for aFRR is that in 98% of the time, the errors are smaller than 10% of the rated capacity. When using a methodology based on reference inverters, standard deviations of the error have been reported in the range of 0.3-6.3% of the rated capacity, where errors tend to be smaller when more reference inverters are used and under clear sky conditions.

6.2 International experience with baselining for wind and solar PV

Although experience with the use of calculated baseline methodologies is more limited, this section aims to provide an overview of international practices with respect to prequalification and monitoring of the baseline. For this purpose, interviews with TenneT, Eirgrid, 50Hertz and Energinet have been organized by the consulting firm DNV.

[52] See among others:

- Advanced Grid-Friendly Controls Demonstration Project for Utility-Scale PV Power Plants, Vahan Gevorgian and Barbara O'Neill National Renewable Energy Laboratory, https://www.nrel.gov/docs/fy16osti/65368.pdf
- ReWP - Regelenergie durch Wind- und Photovoltaikparks, Fraunhofer https://www.iee.fraunhofer.de/content/dam/iee/energiesystemtechnik/de/Dokumente/Projekte/20170814_ReWP_Abschluss_final.pdf
- Highly Accurate Method for Real-Time Active Power Reserve Estimation for Utility-Scale Photovoltaic Power Plants, Vahan Gevorgian National Renewable Energy Laboratory , https://www.nrel.gov/docs/fy19osti/73207.pdf

6.2.1 Application

Table 10 provides an overview of the calculated baseline methodologies used by the different TSOs considered.

Table 10: Use of calculated baseline methodologies

TSO	Service	Baseline resolution	Description	Moment of submission
TenneT	aFRR	4 seconds	The FSP chooses the method to calculate the AAP and describes it during prequalification. The FSP performs the calculation and submits the baseline to the TSO.	Real time
50Hertz	mFRR	1 minute	The FSP chooses the method to calculate the AAP and describes it during prequalification. The FSP performs the calculation and submits the baseline to the TSO.	Real time
Energinet	Offshore emergency for downward regulation	5 minutes	The FSP chooses the method to calculate the AAP and describes it during prequalification. Energinet provides some guidelines. The FSP performs the calculation and submits the baseline to the TSO.	Ex-post
Eirgrid	Wind curtailment and wind constraint services	1 second	The FSP chooses the method to calculate the AAP. The FSP performs the calculation and submits the baseline to the TSO.	Real time

It can be observed that in all cases, the method for the calculation of the AAP is chosen by the FSP and the calculation of the baseline is performed by the FSP and submitted to the TSO. In this regard, calculated baseline methodologies are in practice implemented as (a variation of) a declarative baseline methodology.

6.2.2 Validation of the baseline

TenneT evaluates the baseline (i.e. reference signal as per TenneT's terminology) during the prequalification process. In this regard, the following requirements apply:

- An explanation is provided that justifies why a reference signal with a lead time of one minute (this is the default baseline for participation to aFRR) is insufficiently accurate and why a reference signal without a lead time provides a more accurate estimate of the reference signal.
- The design of, and inputs used for, the calculation of the baseline are described in detail.
- The calculation of the reference signal is completely independent of the TenneT aFRR delta-setpoint and of the portfolio output power.

- The reference signal is traceable to certain inputs, such that it is transparent how the inputs will lead to the calculation of the output.

In addition, TenneT reserves the right to check the data processing (including the reference signal) and relevant procedures of a BSP with an audit performed by an independent party.

50Hertz tests the accuracy of the baseline ("mögliche Einspeisung") outside activation periods. For mFRR, the accuracy of the baseline is tested specifically during the period of 15 minutes prior to the ramp period ("Vorhaltephase") by comparing the baseline to the measured data. In this period, the bias is calculated at minute-by-minute resolution. The absolute value of the mean of this bias should not be greater than 1% for the baseline to be qualified. In addition, the baseline is continuously monitored at pool level.

Energinet only uses the AAP as a baseline for the offshore wind emergency downward mechanisms (i.e. not in Balancing Services). At prequalification stage, Energinet will request a description of the applied methodology and results from the developed calculation. During operations, Energinet monitors the calculation of the non-supplied generation on a continuous basis by comparing the possible power (referred as Available Power Estimate by Energinet) with the measured power output. The accuracy threshold for Energinet is 5% of the actual produced power over 15-minutes intervals.

Eirgrid performs a test at the prequalification stage and tests the quality of the AAP based on the following criteria:

1. The AAP is limited to the "maximum export capacity" of the wind farm
2. The AAP signal is an accurate proxy of the active power the wind farm and should reflect the active power output under normal conditions (i.e. no curtailment or dispatch).

During operations, Eirgrid monitors the accuracy for non-event days by calculating the normalized root mean square error on a daily basis (based on errors calculated on 15-minute intervals). When the normalized root mean square error exceeds 6% for one day, the AAP is considered to be in error.

6.3 Proposal for implementation

For mFRR, Elia considers that a declarative baseline methodology in which the baseline needs to be submitted ex-ante by the FSP is the best solution to enable the participation of wind and solar PV considering that short-term forecasts on a 15-minute resolution can be sufficiently accurate[53]. Although the forecasting accuracy of PV assets has not been analyzed in this study, it is also expected that a sufficient accuracy could be achieved with a declarative baseline.

For aFRR, considering that i) there are certain indications suggesting that the power output of wind farms (and possibly also solar PV) can experience significant volatility on time scales of several seconds to one minute, ii) there are currently limited options in terms of forecasting such short-term power fluctuations, and iii) there are methods that could calculate the AAP with sufficient accuracy, **Elia is willing to allow BSPs** to calculate their baseline based on real-time data and **to submit the baseline in real-time under the following conditions**:

[53] We refer to Section 5.3.2 for more details regarding a possible implementation of a declarative baseline for mFRR.

- The BSP provides a sound justification indicating why there is no viable way to submit a baseline one minute in advance while meeting the accuracy requirements and why submission of the baseline in real time would allow achieving a sufficient accuracy.
- The BSP provides a clear description of the method and inputs used to calculate the baseline and can make these inputs available to Elia upon request. The calculated baseline should be traceable to these different inputs.
- The BSP can prove that the calculation of the baseline is fully independent of the aFRR reference set point and independent of the operating conditions (e.g., curtailed or non-curtailed conditions).

In addition, the criteria and monitoring applied for the current declarative baseline (submitted one minute ahead of real time) will also apply for a baseline submitted in real time. More specifically, in moments that Delivery Point is not activated, the baseline submitted will be compared to the measured offtake/injection in order to check whether the baseline submitted is sufficiently accurate.

In order to ensure that a possible implementation would effectively be used, and considering that the conditions proposed above would in any case need to be met before a BSP would be allowed to submit his baseline in real time, **Elia invites stakeholders to provide:**

- **indications on the volumes for which the current baseline is considered to form a barrier for participation and that will/would be participating in case the baseline can be submitted in real-time;**
- indications **of the feasibility of meeting the proposed criteria** applying for a baseline that can be submitted in real time (i.e., baseline test/monitoring, description of the methods and inputs, baseline independent from the operating conditions) or possible alternatives.

Part C: Impact assessment and conclusions

7. Impact assessment

The study involves several recommendations. Specifically, the following elements can be distinguished:

- the possible introduction of a declarative baseline methodology for mFRR;
- the possible introduction of a declarative baseline methodology for the products ToE DA/ID and the CRM;
- enabling the real-time submission of the (calculated) baseline under certain conditions for aFRR.

Below, an assessment of the impact in terms of required IT developments, adaptations of operational processes (at Elia side as well as for FSPs) and changes to the different regulated documents is presented for each of these recommendations.

Introduction of a declarative baseline methodology for mFRR

The possible introduction of a declarative baseline methodology for mFRR (for assets that do not provide a MW schedule and do not have a scheduling obligation) would require modifying the T&C BSP mFRR and the ToE Rules in order to describe the new baseline methodology. More specifically, this would include a description of:

- the Delivery Points for which the declarative baseline could be chosen by the BSP;
- the process for baseline validation/prequalification (i.e., the metrics and criteria used for assessing whether the declarative baseline methodology increases accuracy with respect to the default baseline methodologies);
- the modalities for data exchanges (i.e., when and how the baseline needs to be submitted);
- the monitoring modalities and corresponding penalties for preventing systematic manipulations of the baseline.

Moreover, minor adaptations might be required in the description of the prequalification processes.

In addition, IT developments and adaptations of operational processes are required. Specifically, new data exchanges related to the submission of the baseline would need to be foreseen, and processes need to be developed to perform the baseline prequalification/validation and to perform the monitoring.

Introduction of a declarative baseline methodology for the products ToE DA/ID and the CRM

An introduction of the declarative baseline for ToE DA/ID and the CRM would require modifications to the regulated documents in order to describe the new declarative baseline methodology. Specifically, for ToE DA/ID, this involves modifying the ToE Rules and the FSP Contract DA/ID, whereas for the CRM, modifications to the CRM Functioning Rules would be required.

Similar as for mFRR, a description of the new baseline methodology would require:

- the process for baseline validation/prequalification (i.e., the metrics and criteria used for assessing whether the declarative baseline methodology increases accuracy with respect to the default baseline methodology, and the requirements on providing transparency on the method used by the FSP/Capacity Provider for calculating the baseline);

- the modalities for data exchanges (i.e., when and how the baseline needs to be submitted);
- the monitoring modalities to avoid systematic manipulations of the baseline and corresponding penalties.

In addition, IT developments and adaptations of operational processes are required. Specifically, new data exchanges related to the submission of the baseline would need to be foreseen, and processes need to be developed to perform the baseline prequalification/validation and to perform the monitoring.

Enabling the real-time submission of the baseline under certain conditions for aFRR

Enabling BSPs to submit the baseline for aFRR until real time if certain conditions are met would require changes to the T&C BSP aFRR in order to describe this new option. Specifically, the process and conditions for getting the authorization to submit the baseline in real-time need to be described.

In addition, IT implementations are required at Elia side in order to correctly distinguish between Delivery Points that need to submit the baseline one minute in advance and the Delivery Points that are allowed to submit the baseline in real time.

8. Summary and conclusions

In recent years, Elia has developed baseline methodologies for the participation of different technologies including active demand response, distributed storage and distributed generation to different products. Currently, the following baseline methodologies are available for the different products under consideration:

- Last QH and High X of Y for mFRR;
- High X of Y for strategic reserves;
- High X of Y* for participation to the day-ahead and intraday markets via the Transfer of Energy mechanism (ToE DA/ID);
- Declarative baseline methodology for aFRR, in which the BSP needs to determine and submit the baseline at the latest one minute ahead of real time.

In this study, Elia analyzes the performance of the baseline methodologies currently in place and the possible opportunities for improving the existing baseline methodologies or introducing new baseline methodologies. To this end, an analysis of international practices and a literature review have been performed, the return of experience from market parties active in Belgium is gathered, and an assessment has been done of the performance of the different baseline methodologies. In addition, a detailed analysis of the opportunities for market parties to choose their own baseline methodology and the baselining options for variable renewable generation have been performed. In these different steps, the characteristics of the products that are relevant for the choice and design of baseline methodology have been taken into account (notably, the typical activation duration, the time between the activation request/decision and the start of the activation, and the frequency of activation). The key conclusions and recommendations are summarized below for each product or group of products. Note that the conclusions and recommendations for the products ToE DA/ID, strategic reserves and the CRM are identical because these products are similar in terms of the characteristics relevant for the design of baseline methodologies.

ToE DA/ID / CRM / Strategic reserves

The assessment indicates that the High X of Y(*) baseline methodology current applied is in line with best practices. Combining a good performance in terms of accuracy for a large variety of assets and simplicity, historical baseline methodologies such as High X of Y(*) are assessed to perform well for these products. This is also reflected in the overall satisfaction of market parties with this baseline methodology and the fact that this type of baseline methodology is the standard for similar products internationally.

Many different variants of historical baseline methodologies exist. In this regard, the literature shows that applying a same-day adjustment can significantly improve the accuracy (although coming with an additional risk for baseline manipulation) whereas other parameters of the historical baseline methodology have a minor impact on the accuracy. As such, there are no indications that the current baseline methodologies could be (significantly) improved.

Based on these observations, a first recommendation is to maintain the High X of Y and High X of Y* baseline methodologies but to develop dedicated monitoring to mitigate the possibilities for baseline manipulation.

Despite the fact that the current High X of Y(*) baseline methodology achieves a very good overall performance, it might not be sufficiently accurate for a minority of assets with a (highly) variable and irregular offtake/injection pattern. In this regard, Elia considers that adding a declarative baseline methodology, in which FSPs can determine and submit their own baseline, would be the best option for an alternative choice of baseline methodology to reduce barriers for such type of assets. In addition, the introduction of a declarative baseline methodology more generally enables FSPs to use a suitable method for calculating their baseline.

For these reasons, Elia is willing to consider developing a declarative baseline methodology for the products ToE DA/ID and the CRM. However, although adding a declarative baseline methodology has certain potential benefits, Elia questions the current need for and added value of introducing a declarative baseline methodology for these products. Therefore, Elia invites market parties to provide indications and specific use cases for which a declarative baseline methodology would be needed to reduce entry barriers.

In terms of the design of the declarative baseline methodology that could be implemented in case there is a clear added value, Elia proposes that the baseline needs to be submitted minimally two days prior to the start of the activation in order to minimize opportunities to selectively manipulate the baseline. In addition, Elia proposes to monitor systematic deviations between the submitted baseline and the measured offtake/injection (i.e., the bias) and to request transparency of the process used by the FSP for calculating the baseline, including the inputs used for the calculation. Finally, Elia proposes that validation/prequalification of the baseline is needed before an FSP can make use of the declarative baseline methodology. Such a prequalification of the baseline would imply that, during a test period, the baseline submitted by the FSP needs to be more accurate than the default High X of Y(*) baseline methodology.

mFRR

The assessment indicates that both the High X of Y and the Last QH baseline methodology current applied are in line with best practices. Combining a good performance in terms of accuracy and simplicity, MBMA baseline methodologies such as Last QH as well as historical baseline methodologies such as High X of Y are assessed to perform well for mFRR. This is also reflected in the overall satisfaction of market parties with these baseline methodologies and the fact that these types of baseline methodology are used the most for similar products internationally. Similar as for ToE DA/ID / strategic reserves and the CRM, there are many different variants of historical baseline methodologies, but outside of applying a baseline adjustment, the different parameters tend not to have a significant impact on the accuracy, particularly considering the relatively short activations for mFRR. Based on these observations, a first recommendation is to maintain the Last QH and High X of Y baseline methodologies.

Although market parties were generally satisfied, some market parties did express the concern that both Last QH and High X of Y would not be sufficiently accurate for technologies they might want to add to their pool in the future, such as variable renewables. The literature indeed confirms that MBMA and historical baseline methodologies can be insufficiently accurate for a minority of assets that have a (highly) variable and irregular offtake/injection pattern, such as wind and solar PV generation, but also possibly variable operated batteries/generation units or demand facilities. In this regard, this study analyzed the opportunities for introducing an additional declarative baseline methodology, in which FSPs can determine and submit their own baseline, and a calculated baseline methodology, in which the baseline is calculated in real-time based on external parameters (e.g., based on real-time wind speed measurements).

The study concludes that for mFRR, a declarative baseline methodology, in which the baseline needs to be submitted by the BSP shortly before the activation request, is sufficient to enable the participation of wind and solar PV. This

because it is considered that short-term forecasts for a product with a 15-minute resolution can be sufficiently accurate. At the same time, such a declarative baseline methodology would also be a solution for other Delivery points with a variable and irregular offtake pattern.

Regarding a possible implementation of an additional (i.e., in addition to the last QH and High X o f Y baseline) declarative baseline methodology, it is important to highlight that this would only apply to Delivery Points that do not provide MW schedules in the framework of the T&C SA and that do not have a scheduling obligation. This because for Delivery Points that do provide MW schedules, the MW schedules are by default used as the basis for quantifying the delivered flexibility (i.e., one could consider that there is by default a declarative baseline methodology in place for Delivery Points that provide MW schedules).

For Delivery Points that do not provide a MW schedule and do not have a scheduling obligation, Elia sees some potential benefits of introducing an additional declarative baseline methodology. Therefore, Elia is willing to consider developing a declarative baseline methodology in case there are clear indications of the need and added value. However, given that market parties mainly expressed concerns related to baselining for wind and solar PV, and considering that all Technical Units (including wind and solar PV parks) with a nominal power exceeding 1 MW will have a scheduling obligation as of phase 2 of the ICAROS project, Elia questions the current need of implementing an additional baseline methodology for mFRR. Therefore, Elia invites market parties to provide specific use cases that indicate a current need for a declarative baseline methodology for those assets that do not have (and will not have) a schedule obligation (i.e., demand facilities and generation units or batteries with a nominal power below 1 MW).

In case there is an added value of implementing a declarative baseline methodology, Elia proposes that the submission/final update of the baseline needs to happen prior to the activation request (e.g., 45 minutes before the start of the quarter hour). In addition, Elia proposes to monitor the submitted baselines (i.e., detect systematic deviations) during moments without activations in order to avoid manipulations of the baseline to the benefit the BSP, and to apply penalties in case systematic deviations are detected for which no sound justification can be provided. Finally, the BSP is proposed to receive the choice between the Last QH, High X of Y and the declarative baseline methodology, but a validation/prequalification of the declarative baseline methodology needs to be performed before the BSP can make use of the declarative baseline methodology. Similar to the proposal for ToE DA/ID and the CRM, the baseline prequalification/validation would involve that during a test period, the submitted baseline needs to be more accurate than the default baseline methodologies for mFRR (i.e., Last QH and High X of Y).

aFRR

Although initially out of scope of this study, market parties indicated during the interviews that the current declarative baseline methodology, in which the BSP needs to submit the baseline for each 4-second interval one minute before real time, may be insufficiently accurate for wind and solar PV resulting in barriers for participation. In this regard, the study has analyzed the need and possibilities for introducing a calculated baseline methodology for assets for which the current declarative baseline methodology might not be sufficiently accurate.

The challenge of baselining for wind and solar PV for aFRR differs significantly from the challenge for mFRR. For aFRR, the main challenge results from the high resolution of the aFRR product (i.e., the baseline needs to be submitted for every 4-second interval) in combination with the short-term fluctuations of wind (and possibly solar PV) power output on these time scales that are highly difficult to predict, even close to real time.

To overcome these challenges, Elia is prepared to enable BSPs to calculate their baseline based on real-time data and to submit the baseline in real-time if the need is confirmed, and only under the following conditions (in addition, the criteria and monitoring applied for the current declarative baseline would also apply for a baseline submitted in real time):

- The BSP provides a sound justification indicating why there is no viable way to submit a baseline one minute in advance while meeting the accuracy requirements and why submission of the baseline in real time would allow achieving a sufficient accuracy.
- The BSP provides a clear description of the method and inputs used to calculate the baseline and can make these inputs available to Elia upon request. The calculated baseline should be traceable to these different inputs.
- The BSP can prove that the calculation of the baseline is fully independent of the aFRR reference set point and independent of the operating conditions (e.g., curtailed or non-curtailed conditions).

In order to ensure that a possible implementation would effectively be used, and considering that the conditions proposed above would in any case need to be met before a BSP would be allowed to submit his baseline in real time, **Elia invites stakeholders to provide:**

- **indications on the volumes for which the current baseline is considered to form a barrier for participation and that will/would be participating in case the baseline can be submitted in real-time;**
- indications **of the feasibility of meeting the proposed criteria** applying for a baseline that can be submitted in real time (i.e., baseline test/monitoring, description of the methods and inputs, baseline independent from the operating conditions) or possible alternatives.

Next steps

Elia will consider the feedback provided by stakeholders during the public consultation in the finalization of the study by the end of the year. The final study report will describe the final conclusions with respect to the implementation of the different recommendations. Based on the final conclusions, a detailed implementation plan will be developed and discussed with the stakeholders. In elaborating the implementation plan, Elia will aim to avoid an excessive number of revisions of the different regulated documents (with associated public consultations), and strive to synchronize the required modifications for an eventual implementation as much as possible with other planned modifications of these documents.