

IEA Wind Task 46 Aerodynamic Benchmark: Computational Aerodynamics Approaches for Assessing Blade Airfoil and Wind Turbine Performance Reduction due to Leading Edge Degradation

M. Sergio Campobasso, Alessio Castorrini, David Bretos, Beatriz Mendez, David C. Maniaci, Johannes N. Theron, Alexander Meyer Forsting, Niels N. Sørensen, Kisorthman Vimalakanthan, Marco Caboni, Ruben Gutierrez, Yana Gorbachova, Aya Aihara, Francesco Grasso

We thank very much the anonymous Reviewers and Prof. Leishman for their valuable comments and suggestions aimed at improving the quality of this article, and the Editor for handling our submission and also making constructive suggestions to improve further the quality of our manuscript. We have made substantial revisions to the manuscript and we hope the revised manuscript will address all questions and suggestions of all who have provided the invaluable input to guide our revisions. During the revision process, we also made some additional corrections. Furthermore, the results of AIST, and part of the results of TNO and Nordex have been updated and partly improved over the results in the first submission. In particular, the AIST analyses of the smooth airfoils no longer use wall functions and integrate all governing equations all the way down to the airfoil wall boundary, like all other smooth airfoil analyses reported in the manuscript.

In the following, we address all comments of the Reviewers and the Editor, summarize our response to the community members, and list the changes and additions we have made to the revised article.

Response to Reviewer 1

Major Comments

Q1.R1 Presentation and Comparison of the Turbulence and Transition Models.

The manuscript should provide a clearer and more concise comparison of the turbulence and transition models used by the different participants. In the current version, it is difficult to identify the differences among the participating groups without repeatedly consulting several paragraphs and the appendices. No consolidated summary of the turbulence-modeling approaches is provided in the main body of the manuscript. I recommend moving Tables A1, A2, and A3, or condensed versions of these tables, into the main body of the manuscript. These tables contain essential information needed to interpret the results. Presenting this information earlier would allow readers to understand the differences among the modeling approaches before examining the comparative results.

A1.R1 We thank the reviewer for this important note aimed at improving the accessibility of the information in the manuscript. Many unnecessary repetitions of models regarding turbulence and transition models readily available have been removed. Repetitions regarding the use of turbulence and transition models of different participants have also been removed. Indeed Section 3.1 of the first submission has been substantially

reduced. Tables A1, A2 and A3 have been moved to the main body of the article. All participants used the same specific version of the turbulence and transition models which has now been clearly stated in the section on turbulence and transition modeling.

We have also added in Tab. 1 (basic CFD code specifications) a column indicating the rough-wall treatment, i.e. whether a rough-wall function is used or the turbulent and transition model equations are integrated all the way down to the wall. This information has not been included for the smooth-LE analyses because all CFD codes now integrate the turbulence and transition model equations all the way down to the wall boundary, i.e. they do not use wall functions.

Q2.R1 Precise Identification and Citation of the Turbulence and Transition Models.

The manuscript should pay more attention to the precise identification and citation of the turbulence and transition models. For example, the SST model is cited using both Menter (1994) and Menter et al. (2003). Although the differences between these SST formulations are relatively limited, the specific model version used by each participant should nevertheless be stated clearly.

This issue is more important for the transition model. The manuscript cites Menter et al. (2006), Langtry et al. (2006), and Langtry and Menter (2009) when discussing the transition-model formulation. However, the model formulations and empirical correlations differ considerably among these publications. For example, the earlier formulation included an acceleration parameter, (K) , and did not publish the complete set of empirical correlations, whereas the later formulation removed this parameter and provided the complete correlations.

As demonstrated in studies such as the NASA Turbulence Modeling Resource validation cases and the AIAA Transition Modeling Workshops, relatively small differences in model formulations and implementations can produce substantially different predictions. Therefore, it is not sufficient to refer generally to the “ $(\gamma - Re_\theta)$ ” transition model. The authors should identify the exact model version and implementation used by each participant.

Considering the already considerable length of the manuscript and the diversity of the numerical approaches, the authors could consider removing or reducing the general explanatory subsections on the SST and transition models (Section 3.1) particularly because not all participants used these models. Instead, the manuscript could provide accurate references and concise descriptions of the exact model versions used by the individual participants.

A2.R1 Thank you for this important notes and the helpful suggestions. We have now substantially reduced Section 3.1 of the first submission (turbulence and transition modeling), and clearly indicated the version of turbulence and transition models used by the participants.

Q3.R1 Readability and Presentation of the Figures.

The overall readability of many figures is poor, including Figures 2, 4, 5, 7, 8, 11, 12, 14, 15, 17, and 18. It is difficult to distinguish the results of the different participants because of the large number of curves and the limited differentiation among line styles, symbols, and colors. In addition, some of the y-axis ranges are too broad or otherwise unsuitable for identifying meaningful differences among the predictions. The authors should revise these figures by using more distinguishable line styles and symbols and by increasing the font and marker sizes. The axis ranges should also be

adjusted where necessary to make differences among the predictions more visible. For Figures 14, 15, 17, and 18, please also provide a caption or legend identifying the dashed line.

A3.R1 Thank you for this important notes. We have reformatted all figures of the articles, and we think this has resulted in greatly improved readability of all figures, including those mentioned by the reviewer.

Q4.R1 Differences in Equivalent Sand-Grain Roughness: Tables A2 and A3.

The equivalent sand-grain roughness heights differ among the participants, and several participants appear to have used values that were not derived directly from the experiments. Please explain why different equivalent sand-grain roughness heights were selected. The use of different roughness values raises concerns regarding the fairness and interpretation of the benchmark comparison. If the objective is to compare different approaches for roughness modelling, readers would generally expect all participants to use the same equivalent sand-grain roughness height.

The authors should discuss this issue explicitly and clarify why the experimental equivalent sand-grain roughness heights were not used by all participants.

A4.R1 We thank the reviewer for this important note. The value of K_s to be used by each participant for each rough airfoil experiment was indeed not prescribed. The participants were only given the geometric information of how the particular roughness pattern was generated on the airfoil model of each experiment; the choice of K_s was left to each participant. In the table with the parameters of the Delft experiment simulations only the values of 5000 and 1000 μm do not directly refer to LE roughness specifications of the experiment. The value of 5000 μm was initially proposed by CENER for the CFD simulations of this experiment (Mendez and Pires, 2022) and adopted by some participants of this benchmark. In Mendez and Pires (2022) the value of K_s in the simulations was increased with respect to the sand-grain size of the P40 sandpaper also to account for the additional thickness of the support layers used to apply the sandpaper to the model LE. The value of 1000 μm was the latest choice of CENER for the analysis of the considered experiment.

These notes have been included in Section "CFD simulation set-ups" of the revised article.

One of the aims of the paper was not only to compare different numerical models, but also to assess the variability of the analyses resulting from user practice- and experience-dependent factors. This has been emphasized also in the introduction of the revised manuscript. Unfortunately, this leads to some loss of generality for this particular comparison.

Specific and Minor Comments

Q5.R1 Please verify the citation format throughout the manuscript, including Lines 35 and 40.

A5.R1 We have verified the citation format throughout the manuscript and amended all outstanding inconsistencies in the use of the latex citation commands. We thank the reviewer for this note.

Q6.R1 Line 35:

Current text: rain Keegan et al. (2013)

Suggested revision: rain (Keegan et al., 2013)

A6.R1 Thank you. We have now corrected this error.

Q7.R1 Line 40:

Current text: LE structure at severe damage stages Han et al. (2018); Im and Kim. (2019); Cappugi et al. (2021); Campobasso et al. (2023).

Suggested revision: LE structure at severe damage stages (Han et al., 2018; Im and Kim, 2019; Cappugi et al., 2021; Campobasso et al., 2023).

A7.R1 Thank you. We have now corrected this error.

Q8.R1 Lines 129 and 160: The manuscript states:

“The roughness element height is ($K=200\text{ }\mu\text{m}$), corresponding to a value of ($K=246\text{ }\mu\text{m}$) for a unitary chord.” and “The equivalent sand-grain roughness of the P40 sandpaper is ($K_s \approx 425\text{ }\mu\text{m}$), corresponding to a value of about ($708\text{ }\mu\text{m}$) for a unitary chord.”

Please also provide the equivalent sand-grain roughness for the lower-roughness case. The discussion currently presented in Lines 469–470 could be moved to immediately after Line 129. The equivalent sand-grain roughness could then be briefly restated in the later discussion for the reader’s convenience.

A8.R1 We thank the reviewer for this recommendation, which we have adopted in the revised manuscript. The revised text now reports that the physical roughness-element height in the Texas A&M case is $K = 200\text{ }\mu\text{m}$, corresponding to $K/c = 246 \times 10^{-6}$, and that the empirical correlation of Langel et al. (2017) gives an equivalent sand-grain roughness $K_s/c = 101 \times 10^{-6}$. We also clarified that this K_s value is a model input derived from a roughness conversion and should not be confused with the measured physical roughness height.

Q9.R1 Line 385:

”The roughness-induced transition model of Langel et al. (2017) was originally developed in conjunction with the Langtry–Menter two-equation transition model. Please briefly explain how this roughness model was implemented in RFOIL and how it was coupled with the transition-prediction method used in the present study.”

A9.R1 A brief qualitative description of how the Sandia’s roughness model was used with RFOIL has been added to the revised manuscript. In essence, the roughness model has not been implemented directly in RFOIL, but rather implemented and used externally to determine the variation of the boundary layer transition due to roughness with respect to the smooth airfoil reference. The updated positions of the transition are then prescribed in a second RFOIL analysis to determine the aerodynamic performance (lift, drag, static pressure and skin friction profiles) due to roughness. Therefore, the updated RFOIL analysis, does not account for the higher viscous stress due to roughness, it only accounts for the variation of the position of the transition. We apologize for not making this clear in the original submission. Moreover, a TNO technical report containing the detailed implementation of the method in RFOIL has been added to the bibliography. The report is unfortunately not available online but can be requested.

Q10.R1 Table A1:

”Please define “Ambient turb.” more clearly. Does this term indicate the use of the sustaining terms by Spalart and Rumsey (2007)?

The table currently reports the physical first-cell height. Please provide the corresponding nondimensional wall spacing, (y^+) , because this quantity is more informative. It is unclear whether the mesh-topology column provides useful differentiation because all participants used structured (all-quadrilateral) meshes. Other mesh parameters may be more informative, including:

- Wall-normal growth ratio
- Leading edge / trailing edge spacings”.

A10.R1 We thank the reviewer for these important recommendations. In the revised manuscript we have added in Tables 1-3 all the data requested. We have also greatly improved the discussions and completeness of Section 4, providing a more clear overview of the set-ups of all CFD codes. With regard to ambient turbulence, that term indicates the level of TI in the test section of each experiment. Most CFD codes had the target of ensuring that the TI past the airfoil in the simulations had a TI level comparable to that in the experiments. This can be achieved with different approaches, one of which is the use of the turbulence-sustaining terms proposed by Spalart and Rumsey in 2007. However, only the DTU code implemented these terms. Most other codes used alternative methods to achieve this target, and the methods are now more clearly defined in the subsections summarizing the set-up of each code. The column now labeled Ambient-Turbulence Preserving (ATP) in Tab. 1 now indicates whether the codes actively used any of several methods to ensure the target TI past the airfoil.

Response to Reviewer 2

Main Comments

Q1.R2 The presentation and critical discussion of the experimental validation datasets should be improved.

A1.R2 We thank the reviewer for this important recommendation. We agree that the validation datasets needed to be described and assessed more thoroughly, particularly because the numerical results are evaluated against two different wind-tunnel campaigns. We have substantially expanded Section 2 of the revised manuscript. In particular, we added a new Section 2.3, in the cross-tunnel comparison and assessment of the validation datasets, which directly compares the Texas A&M and Delft experiments and discusses their respective strengths and limitations for validation. The revised manuscript now includes a summary table comparing the two datasets in terms of airfoil chord, model span and mounting, test-section dimensions, wall-clearance/sealing treatment, turbulence intensity, roughness type and extent, wake-rake configuration, static-pressure treatment, transition measurements, and availability of formal uncertainty information. We also added a direct clean-airfoil comparison between the two datasets, including clean polars and clean pressure-coefficient distributions. This comparison is used to establish the level of cross-facility agreement that can reasonably be expected before interpreting the roughness-validation results. The revised text also explains that the two datasets are complementary rather than interchangeable. The Texas A&M dataset is particularly useful for assessing moderate, distributed leading-edge roughness because the roughness pattern is repeatable and uncertainties are documented for the integrated coefficients. The Delft dataset

is particularly useful for assessing severe leading-edge roughness because the P40 sandpaper produces a large aerodynamic penalty and the model-wall gaps were carefully sealed. The revised manuscript also discusses the limitations of both datasets, including wall-clearance effects, availability of uncertainty information, exact model-geometry availability, sandpaper/tape step effects, and wake-rake limitations in post-stall flow.

Q2.R2 Similarly the discussion on the differences between predictions and experiments can be improved.

A2.R2 Thank you for this note, on which we totally agree. We have substantially improved the discussion of numerical results in two ways: 1) highlighting all most likely cause-effect relations between numerical and physical settings of each computational set-up and the answers it delivered, which allowed us to also better explain the differences among predictions, 2) emphasizing the sources of uncertainty of the numerical predictions and, qualitatively, also the variations of the level of this uncertainty in relation to physics (e.g. stall and post-stall regimes) and also engineering aspects (e.g. implementation of the roughness in the two experiments, controllability of which was different (e.g. known roughness element height and area density of roughness elements applied directly on the model surface in one case, and unknown thickness of the base of the P40 sand-paper, applied on top of model surface, in the other). These revisions are distributed throughout the Results and Conclusion sections. Furthermore, we have also improved the description of all codes' inputs and compacted all this information in Section 4, providing all input data relevant to the presented analyses and also cross-comparing the set-ups.

Q3.R2 A statement on LE roughness being beneficial at $Re=3M$ needs to be justified, as it contradicts present experience.

A3.R2 We thank the reviewer for this important note. Indeed we have no evidence that roughness is beneficial on the post-stall behavior at high Reynolds numbers, and the noted result is unexpected. The result may be affected by significant uncertainty due to the phenomenon occurring in an operating condition range characterized by significant three-dimensionality and unsteadiness of the flow field. Nevertheless, similar unexpected trends have been observed in a few other studies at not-too-low Reynolds numbers. Therefore, we have added the following text in the revised manuscript: "Above $\alpha = 11^\circ$, the slope of the measured Δc_l curve becomes positive, and from $\alpha = 12^\circ$ the measured Δc_l values are positive. This trend points to an apparent reduction of the post-stall lift penalty for the low-level roughness case. This result is unexpected and should not be interpreted as conclusive evidence of a robust aerodynamic benefit of LE roughness. In this regime, the measured sectional lift is influenced by separated-flow three-dimensionality and unsteadiness, and the comparison with steady two-dimensional simulations is inherently less definitive than in attached flow. A similar trend was also observed in the experimental study of Zhang 2018, which reported an improved c_l of an airfoil with low-level LE roughness over that of the smooth airfoil in the post-stall regime. However, the measurements were performed for Reynolds numbers up to only 0.26M and considered a different airfoil. Unexpected improvements of the maximum c_l of the NACA 63₃-418 featuring low-level LE roughness were reported by Aagaard Madsen 2010 in relation to wind tunnel measurements at Reynolds number of 1.6M in the Delft wind tunnel. One possible physical explanation in some low- or moderate-Reynolds-number cases is

that roughness or tripping modifies transition and suppresses laminar-separation-bubble effects, thereby altering the stall process. However, the present data do not establish this mechanism for the Texas A&M case. Confirming whether the observed effect of roughness on post-stall behavior is physical, and to what degree it depends on the facility, airfoil, Reynolds number, and post-stall measurement and data-reduction methods, would require targeted experiments and simulations.”

Q4.R2 The conclusions are purely descriptive. Based on all the nice work presented here, the paper should include some form of best practice guidelines, otherwise its usefulness remains limited.

A4.R2 We thank the Reviewer for their appreciation of the presented work, and we fully agree with their note on the Conclusions. Therefore, we have rewritten Section ”Conclusions”, highlighting both the features of present practice of CFD analyses of wind turbine airfoil roughness that we found work adequately and should therefore be considered as Best practice, and also aspects where present simulation-based technology needs further improvements, such as improvement of turbulence modeling for relatively thick wind turbine airfoils and adoption of multi-scale approaches to the CFD analysis of severe LE degradation, featuring both uniformly distributed roughness and jagged features, such as steps at the ends of applied sand-paper in wind tunnel experiments and highly irregular surface profiles of real LE erosion encountered in wind turbine operation.

Detailed Comments

We thank the reviewer for their thorough comments in the annotated PDF file attached to the review. We have comprehensively answered all 46 comments. Below we report only those where we propose some variants with respect to what kindly suggested, and those where we believe it is important to highlight some essential clarifications we made in the revised manuscript.

Q5.R2 Line 103 of first submission (Texas A&M experiment): ”could you please state the distance between the model tips and the wind tunnel walls?”

A5.R2 We have added the requested information to Section 2.1 and to the new experimental-summary table in Section 2.3. The Texas A&M model had chord $c = 0.813$ m and span 2.1 m, and it was mounted vertically in a 2.1 m-high test section. Therefore, the nominal clearance was approximately 12 mm at both the floor and ceiling. We also state that no dedicated sealing of these gaps is documented in Ehrmann et al. (2017).

Q6.R2 Line 131 of first submission: ”Is this based on field data from inspecting actual blades? Is there a reference you could cite?”

A6.R2 We have clarified the basis of the Texas A&M roughness pattern in Section 2.1. The distributed vinyl roughness pattern was informed by measurements of roughness on operational wind-turbine blades and by insect-impingement calculations, as documented by Ehrmann et al. (2017). We now state explicitly that the benchmark case should be interpreted as a simplified, repeatable representation intended for aerodynamic-model validation, rather than as a direct reproduction of one specific blade condition. We also clarified that the chordwise roughness extent, from the leading edge to $0.13c$ on the pressure side and $0.02c$ on the suction side, was guided by the

30 % normalized mass-accumulation cutoff from the insect-impingement calculations reported in Ehrmann et al. (2017).

Q7.R2 Line 150 of first submission (Delft experiment): "could you please state the distance between the model tips and the wind tunnel walls?"

A7.R2 We have corrected and expanded the Delft model description in Section 2.2 and in the new experimental-summary table in Section 2.3. The revised manuscript now states that the Delft model had chord $c = 0.6$ m and total span of approximately 1.9 m. Since the TU Delft test-section height was 1.25 m, the model protruded through the upper wind-tunnel wall and spanned the full test-section height. We also added that all model-wall gaps were carefully sealed to avoid leakage, following the IRPWIND/Timmer documentation.

Q8.R2 Line 150 first submission: "For both campaigns, how was the static pressure at the wake rake location treated (e.g. averaged?)"
"Also, how was drag measured post cl_{max} ?"

A8.R2 We have expanded the wake-rake descriptions for both experiments. For the Texas A&M measurements, the revised manuscript now states that the wake rake was located $0.9c$ downstream of the model trailing edge and contained 25 Pitot probes and three static-pressure probes. The static-pressure variation across the wake-rake location was represented by a linear least-squares fit through the three static-pressure measurements, and this fitted variation was subtracted from the measured wake-pressure deficit before evaluating the wake dynamic pressure and applying the wake momentum integral.

For the Delft measurements, the revised manuscript now states that drag was evaluated using the Jones wake-rake momentum method, in which the measured total- and static-pressure distributions across the wake are included directly in the integral. We also corrected the Delft wake-rake distance to $0.82c$, or 490 mm, downstream of the model trailing edge. The revised text reports that the Delft static wake rake contained 16 static-pressure probes and the total-pressure wake rake contained 67 tubes. We have also clarified that the wake-rake pressure signals were temporally averaged; for fixed-rake measurements, the Delft documentation reports a 20 s averaging time. Thus, the static pressure was temporally averaged at each probe location, while the measured spatial static-pressure distribution was retained in the Jones wake integral rather than replaced by a single spatially averaged value.

We also added a discussion of post-stall drag limitations in the new validation-dataset assessment section. The revised manuscript explains that wake-rake drag is most reliable in attached and mildly separated flow. In post-stall conditions, separated wakes can contain large total-pressure losses, non-equilibrium wake development, three-dimensional stall cells, and strong spanwise nonuniformity. The Delft report explicitly notes that wake-rake drag after maximum lift can show strong spanwise irregularity due to stall cells and can become invalid at high AoA or large total-pressure losses. Pressure drag derived from the model pressure taps may be more consistent in this regime, although it does not include the viscous contribution to total drag. For the sandpaper cases, the Delft documentation notes that a clean crossover from wake-rake drag to pressure drag was not straightforward. Therefore, the revised manuscript treats attached-flow and near-stall comparisons as the primary quantitative validation target and post-stall drag comparisons as qualitative.

Q9.R2 I think it would help a lot if you could add a section here (or later in the text if you so wish) discussing the cross-tunnel clean airfoil results comparison. This can only be implicitly assessed at this point. The authors also have two sentences on this in section 5.2. I understand the paper is about computational modeling, but the numerical approaches are validated against wind tunnel data. It is imperative the validation dataset is presented in detail and understood correctly.

This section should contain at least the following four plots, comparison table, and authors assessment.

(a) force coefficient polar comparison between the two data sets for the smooth cases. (two plots)

(b) C_p vs X comparisons at an angle of attack in the linear region and an AoA in post stall (two more plots).

(c) A table where the relevant parameters for the two wind tunnel datasets are summarized. This will facilitate understanding of the commonalities/differences between the two sets.

(d) your assessment on the use of the specific datasets for validation purposes. See also comment in section 5.1 on the three-dimensionality of the flow. Finally, all experimental data should include clearly defined error bars.

A9.R2 We agree that a direct comparison and assessment of the two experimental datasets is important for interpreting the benchmark. We have added a new Section 2.3 on cross-tunnel comparison and assessment of the validation datasets. This section includes a summary table comparing the Texas A&M and Delft experiments, including model dimensions, test-section dimensions, mounting and sealing, turbulence intensity, roughness type and extent, wake-rake configuration, static-pressure treatment, transition measurements, and uncertainty availability.

We also added the requested clean-airfoil cross-tunnel comparison. The revised manuscript now includes a comparison of the clean polars from the Texas A&M and Delft experiments and a comparison of clean pressure-coefficient distributions at selected angles of attack. The $\alpha = 0^\circ$ pressure comparison is used as an attached-flow comparison, while the $\alpha = 10^\circ$ case is described as a high-angle-of-attack, near-stall comparison rather than as a universally post-stall condition, because stall onset differs between the two datasets.

The new text explains that differences between the Texas A&M and Delft clean-airfoil measurements should not be interpreted solely as statistical measurement uncertainty. They may also reflect facility effects, model geometry and surface finish, model mounting and wall sealing, turbulence intensity, wall corrections, pressure referencing, wake-rake implementation, and data-reduction procedures. We also cite previous work showing that measured sectional coefficients, particularly near stall and in separated-flow regimes, can depend on wind tunnel facility effects, wall/corner-flow treatment, and measurement method. This supports interpreting the Texas A&M–Delft clean-airfoil comparison as a cross-facility consistency check rather than as a strict pointwise uncertainty estimate.

Finally, Section 2.3 now includes an explicit validation-dataset assessment. We identify the Texas A&M dataset as particularly useful for moderate, distributed leading-edge roughness and the Delft dataset as particularly useful for severe leading-edge roughness. We also discuss the main limitations of each dataset and state that the most meaningful quantitative comparisons are in attached and near-stall flow, while fully post-stall comparisons should be interpreted qualitatively.

We have added a dedicated discussion of experimental uncertainty and data quality in the new Section 2.3. For the Texas A&M measurements, the revised manuscript clarifies that lift uncertainty was obtained from first-order propagation of pressure-measurement uncertainties, not from a constant percentage uncertainty assigned to c_l over the full angle-of-attack range. The underlying pressure uncertainties are associated with pressure-transducer accuracy and temperature-related drift and are expressed as pressure uncertainties, rather than as a constant percentage of the measured pressure. Consequently, the resulting relative uncertainty in an integrated coefficient varies with operating condition and angle of attack.

The revised manuscript reports representative uncertainty values from Ehrmann et al. (2017). The median relative lift uncertainty was approximately 0.8 % at $Re_c = 1.6 \times 10^6$, with the corresponding uncertainty at $Re_c = 3.2 \times 10^6$, the Reynolds number used in the present benchmark, expected to be no larger. Texas A&M drag uncertainty was estimated separately using a bootstrap procedure applied to the wake-rake drag calculation. The reported mean relative drag uncertainty was approximately 12 % at $Re_c = 1.6 \times 10^6$ and 5.1 % at $Re_c = 2.4 \times 10^6$, with drag uncertainty at higher Reynolds numbers generally near 5 %. For the $Re_c = 3.2 \times 10^6$ clean case, a representative absolute wake-rake uncertainty was 4.4 drag counts, equivalent to 4.4×10^{-4} in c_d . We also state that pointwise c_p uncertainty values are not tabulated in the available report, although the uncertainty-propagation procedure is documented.

For the Delft data, formal uncertainty values for the force and pressure data were not available in the documentation provided to the benchmark participants. The revised manuscript states this explicitly and summarizes the documented Delft data-quality procedures, including pressure sampling at 330 Hz, 20 s averaging for fixed-rake measurements, and wake-rake traverses over a 40 cm spanwise range to assess two-dimensionality and identify a representative drag-measurement region. We also clarified that, in the Delft wake-rake drag calculation, the pressure signals were temporally averaged, while the measured spatial distribution of static pressure across the wake was retained in the Jones wake integral rather than replaced by a single spatially averaged value.

Because uncertainty bars are not plotted for all experimental curves, the revised figure caption notes that experimental uncertainty bars are omitted for readability, that representative Texas A&M uncertainties are summarized in Section 2.3, and that formal Delft uncertainty values were not available in the documentation provided to the benchmark participants.

We also clarified the pressure-reference definitions used for the c_p comparison. For the Texas A&M data, Ehrmann et al. (2017) define the pressure coefficient using the tunnel total pressure, $c_p = (p - p_0)/q_\infty + 1$, because this formulation yields a stagnation pressure coefficient closer to unity. For the Delft data, the Timmer report defines $c_p = (p - p_0)/q_0$, with q_0 obtained from the wind-tunnel dynamic-pressure calibration curve and $p_0 = p_{t,0} - q_0$, where $p_{t,0}$ is generally obtained from the average of wake-rake total pressures outside the wake. Because the pressure-reference definitions and wall-correction procedures differ between facilities, the revised manuscript states that the cross-tunnel c_p comparison is used as a qualitative consistency check rather than as a strict pointwise validation metric.

Q10.R2 With reference to the Fluent set-up (Section 2.3, line 227 of fist submission), in which we wrote:

"The constant a_1 has been set to 0.29 in all the analyses ...", the reviewer asks:
"how did the users end up with this value? Is this a value they normally use? Did they play around just for this study? Please provide more info. Similarly in all cases where non-standard parameter values are used."

A10.R2 All users who changed the default value of a_1 for the analyses reported in the manuscript performed some sensitivity analysis, looking at the sensitivity of the predicted airfoil performance as a function of this parameter. The selected value was that which led the closest match of measured and computed stall AoA. This manuscript does not report these sensitivity analysis for brevity, but a fairly comprehensive sensitivity analysis of this type is available in Gutierrez et al. (2024). These notes have been reported in the general section on turbulence and transition modeling of the revised manuscript.

Q11.R2 With reference to the notes on the use of rough-wall functions in the Fluent set-up (line 232, Section 3.2 of first submission), the reviewer asks:
"What is the requirement for the first cell distance for the employed wall functions? Is it the same as for a fully resolved BL simulation and y^+_1 ? If not, is it advisable to use the same grid? Is this the reason at least one participant used two different grids for smooth and rough-surface simulations?"

A11.R2 We thank the reviewer for this excellent question. With all rough-wall functions the requirement is that the minimum wall distance ensures that y^+ be in the logarithmic part of the boundary layer. Hence the requirement is not that y^+ be 1 or smaller. The requirement that y^+ be in the logarithmic part of the boundary layer was fulfilled by all three codes using rough-wall functions. The Fluent simulations can use the same grid for smooth and rough airfoil simulations thanks to the so-called virtual wall model implemented in Fluent, for which we have provided a reference in the revised manuscript. This approach is unfortunately not available in OpenFOAM, which requires the wall-adjacent cell height to be comparable with the sand-grain roughness. This is the reason why the two OpenFOAM set-ups using rough-wall functions had to use different grids for the smooth and rough airfoil analyses. All this has been explained in detail in the section on CFD simulation set-ups of the revised manuscript.

Q12.R2 The reviewer asks for clarifications regarding the use of the method by Hellsten and Laine to improve the convergence of simulations modeling the effects of high levels of roughness in the DTU set-up (line 268 of first submission).

A12.R2 Thank you for driving our attention to this item. The point has been clarified in the revised manuscript.

Q13.R2 The reviewer proposes a more clear rewording of the description of how the two variables of the SST model were determined in the Nordex OpenFOAM simulations (line 320 of the first submission).

A13.R2 We thank the reviewer for their helpful suggestion. The interpretation of the original sentence is fully correct, and we have adopted the wording suggested by the reviewer in the revised manuscript.

Q14.R2 With reference to first line of section 3.4.4 of first manuscript submission, the reviewer asks: "Is change of font necessary?"

A14.R2 Thank you for the note. Indeed, it was not necessary and we have changed this in the revised manuscript. In general, we have simplified the description of the specific purely numerical methods because we have realized that this is only marginally relevant to the focus of our study.

Q15.R2 With reference to following statement in TNO OpenFOAM set-up (Section 3.4.4, line 358 of first submission):

"This model is an alternative to rough-wall functions for incorporating surface roughness effects on BL transition, a modeling feature essential for predicting accurately the aerodynamic performance of airfoils in the presence of moderate LEE.", the reviewer asks:

"does this mean that other approaches presented herein, where this alternative model was not adopted, were bound to fail?"

A15.R2 We thank the reviewer for this observation and we apologize for the inaccurate statement. The implementation of the transport equation for the roughness amplification variable in OpenFOAM by Vimalakanthan et al. (2023), based on the model proposed by Dassler et al. (2012), potentially enables predicting the location of BL transition in the presence of LE roughness integrating all turbulence and transition transport equations all the way down to the wall. The alternative approach is using the 4-equation $\gamma - Re_\theta$ SST model suitably coupled to a rough-wall function. One of the potential benefits of the 5-equation model is the avoidance of possible inaccuracies resulting from the use of wall functions. However, a relatively small amount of validation studies of this approach presently exists, a circumstance which makes this promising approach very interesting and worth further testing, one of the key reasons for its use in this study. These comments have been summarized in the revised manuscript.

Q16.R2 "Please state whether data from the two wind tunnels mentioned in the present study have been used to inform the DART tool."

A16.R2 No. None of the experimental data presented in this article have been used to inform the DART tool. We have stated this also in the revised manuscript.

Q17.R2 With reference to the statement on line 407 of the original submission "For the analyses using smooth- or rough-wall functions, it was verified that δ_w^+ always lied in the logarithmic range of the turbulent BL.", RC2 asks:

"Was this not ensured for all participants using rough wall functions?"

A17.R2 Yes, it was ensured for all three participants using rough-wall functions. We have stated it more clearly in the revised article. Additionally, all smooth-LE simulations now resolve the BL all the way down to the wall; hence smooth-wall functions are no longer used by any code.

Q18.R2 With reference to line 432-433 (section 5.1 of original manuscript on Texas A&M experiment simulations), RC2 notes:

"The agreement is indeed very good and the partners are commended on this.

For higher AoA, however, the flow separates and in fact higher variability is observed between relevant wind tunnel data than the numerical results shown here [1].

For completeness, the authors should discuss the three-dimensionality of the flow in the experiments and the fact that their approaches are 2D.

More specifically, they should comment on whether agreement should be expected

at and beyond stall between 2D CFD data and sectional pressure lift from a 3D separated wind tunnel flow [2]?

If this is an engineering approach, or an industry-led practice, please explicitly state so.

-[1] DOI: 10.1002/we.2630

-[2] doi: 10.1088/1742-6596/1618/3/032040

A18.R2 We are grateful to the reviewer for their valuable insight. We agree with all observations made and we have updated the indicated part of the manuscript accordingly, also citing the indicated sources to support the new statements.

With regard to the question on the suitability of 2D RANS analyses for post-stall analyses, we are aware that these simulations are unlikely to capture the complete physics of the flow three-dimensionality and unsteady effects in that post-stall regime. This is now explicitly stated also in the revised manuscript.

Q19.R2 Please use consistent colours/symbols for each partner between plots.

Symbols could be larger to facilitate distinguishing between the different lines.

A19.R2 Thank you for this comment. We have updated all figures, checked them for consistency and corrected any issue encountered.

Q20.R2 With reference to sentence of the first submission (line 475): "... pointing to a possible performance benefit due to small roughness in this region.", RC2 comments: "This could also be the fact that the measurements concern the sectional lift of a 3D separated flow. Have the authors confirmed this is not the case?"

Roughness can have a beneficial effect at much lower Re numbers, but if the authors wish to state that at Re=3M adding LE roughness improves the airfoil performance, they will have to convincingly justify this."

A20.R2 We thank the reviewer for this insightful note. This question has been addressed in our answer A3.R2 above.

Q21.R2 With reference to line 510 of first submission, RC2 writes:

"please include this definition in the description for both experimental campaigns". This refers to the definition of the reference static pressure in the two experiments.

A21.R2 We thank the reviewer for requesting clarification of the pressure-reference definitions. We have added this information to the revised manuscript. For the Delft experiment, the revised manuscript now states that pressure coefficients were computed as $c_p = (p - p_0)/q_0$, where q_0 is the undisturbed dynamic pressure in the test section and p_0 is the undisturbed static pressure. The value of q_0 was obtained from the wind-tunnel dynamic-pressure calibration curve, which relates the dynamic pressure in the center of the empty test section to the tunnel control pressure Δp_b . The tunnel control pressure is formed from the difference between the settling-chamber total pressure and a wall static pressure measured approximately halfway through the contraction, with this wall static pressure obtained as the average of four pressures around the contraction circumference. The undisturbed static pressure was then inferred from $p_0 = p_{t,0} - q_0$, where $p_{t,0}$ is the undisturbed total pressure. For most cases, $p_{t,0}$ was taken as the average of wake-rake total pressures outside the wake. For very high-drag cases, where insufficient wake-rake tubes were outside the wake, $p_{t,0}$ was taken from a Pitot-static tube at the test-section entrance.

For the Texas A&M experiment, the revised manuscript states that the pressure distribution was expressed in nondimensional form as c_p and then integrated to obtain lift, and provides details of the equations that can be compared to the Delft method. In particular, the Texas A&M reduction uses $c_p = (p - p_t)/q_\infty + 1$, where p_t is the wind tunnel total pressure, p_0 is the reference static pressure and q_∞ is the freestream dynamic pressure. The detailed treatment of the reference pressure p_0 and uncertainty propagation are documented in Ehrmann et al. (2017), which is now cited as the accessible public Sandia report.

The approach used in the Texas A&M wind tunnel for obtaining c_p does not require a direct measurement of a reference static pressure. It uses instead the wind tunnel total pressure and the reference dynamic pressure. This has been thoroughly explained in section 2.1 of the revised manuscript.

Q22.R2 Fig. 6b of first submission: "26. It appears there is a step in the TNO data, here, could the authors comment on this?"

A22.R2 We thank the reviewer for drawing our attention to this feature. The result of TNO was updated and that discontinuity has now nearly disappeared.

Q23.R2 Line 522 of first submission: "again, this is sectional lift of a 3D separated flow"

A23.R2 Thank you for this note, we agree. We added a comment on this in the revised manuscript.

Q24.R2 With reference to line 529 of first submission: "The larger differences among the computed c_l curves are due to the fact that the geometry of the tested airfoil model is not available, and participants used slightly different airfoil geometries for the numerical analyses", RC2 suggests:

"are potentially due to.

A positive statement here would require confirming that this is the reason by simulating the tested profile, too."

A24.R2 Thank you for this note. We fully agree with the evaluation and have added "potentially" in the revised manuscript.

Q525R2 RC2's comment on line 530 regarding the slightly different airfoil geometry used by the participants for the analysis of the Delft experiment, which may be the reason for the larger differences among computed results: "do you mean all participants used the same nominal geometry, which was different to the (unknown) measured one? Or that each participant used a different profile. Please clarify".

A25.R2 We thank the reviewer for this important question. The nominal airfoil geometry used by all participants for the Delft experiment was indeed the same, but input file of the geometry had a participant-dependent format, i.e. was based on participant-specific CAD systems. Moreover, some partners use grid generator that do not even require a CAD file as input, and import instead a text file with (x,y) coordinates. The variability of the same nominal geometry resulting from different format of the input geometry for each grid generation system, may also account for the larger result spread in the Delft case. For the Texas A&M experiment, conversely, the control points of a generalized a Bezier parametrization of the tested airfoil model, outputting 1000 points for the airfoil geometry, was used by all participants. A summary of these notes has been included in the revised manuscript.

Q26.R2 Comment on line 550, regarding the scatter of computed c_l curves of the rough airfoil of the Delft experiment with respect to the Texas A&M experiment. In first submission we write: "Inspection of the left plot of Fig. 9 reveals a notably larger data scatter than observed for the rough airfoil of the Texas A&M analysis considered in Fig. 4. These differences are partly due to the aforementioned small differences in the airfoil geometry used by the participants."

The reviewer commented: "The scatter for the rough case is larger than that for the clean case. For the geometry differences to justify a larger scatter, they should also be greater than in the clean case. Is this what the authors mean? If yes, how was this quantified? If they mean something else, please clarify."

A26.R2 We thank the reviewer for this comment. We fully agree that the scatter of the computed c_l curves of the Delft rough airfoil is larger than that of Texas A&M rough experiment. We also agree that the scatter of the Delft rough computed curves is larger than that of the computed curves of the same wind tunnel for the clean case. Indeed this is why we wrote "partly to the aforementioned small geometry differences". The rest of the differences is due to the use of different models in the CFD simulations. We have added this note in the revised manuscript, apologies for not reporting it in the first submission.

Q27.R2 With regard to sentence at line 595 of first submission: "Quite interestingly, at high AoA, the predicted drag reduction of all simulations is lower than that based on the measurements.", RC2 writes:

"I am probably missing something here, but all Δc_d values are positive, is this a reduction or increase?"

A27.R2 We apologize and thank the reviewer for their very careful reading. That was indeed a mistake, we meant "increase". This has now been corrected.

Q28.R2 Comment referring to Fig. 10c: "Is this $(Cl/Cd)_{cfd} - (Cl/Cd)_{exp}$? How can the difference be at the order 100? is this normalised somehow? Please consider adding the equation for all deltas, to avoid any misunderstandings. The same comment holds for Fig. 5

A28.R2 We apologize for the insufficient clarity on this matter. The definition of the reported differences is:

$$\Delta\phi_{\text{method}} = (\phi_{\text{rough}} - \phi_{\text{smooth}})_{\text{method}}$$

where ϕ stands for c_l , c_d or c_l/c_d , and subscript "method" denotes the approach adopted for the evaluation, i.e. experimental or based on one of the simulation set-ups.

This text has been added in the revised manuscript at the start of the text commenting the plot with the aforementioned measured and computed Deltas for the Texas A&M experiment.

Q29.R2 With reference to lines 638-639 of the first submission: "... the required time-series of the inflow wind maps have been generated with the TurbSim code (Jonkman and Kilcher)", RC2 writes:

"I think it would increase the uptake and impact of this paper if inflow dataset was publicly available on an online repository. Have the authors considered this?"

- A29.R2 We thank the reviewer for this good suggestion. Since TurbSim is a well documented and supported widely used opensource code we have opted for a slightly more practical approach. Specifically, we have added an appendix to the paper in which we provide the values of all relevant input variables that were used to generate the wind time-series for the reported analyses.
- Q30.R2 With reference to the first part of section 5.3 on analyses of power and AEP losses, RC2 writes (note on line 644 of annotated file of first submission): "In this section, up to this point, the employed methodology is described. This belongs to a methodology section, not to the results. Please move this part to an independent section, before the results."
- A30.R2 We thank the reviewer for this thoughtful note. Indeed the first part of the indicated section explains the procedure used for the subsequent power and energy loss analyses. In the revised manuscript, however, we preferred not to move the initial part of this section towards the beginning of the article, in the methodology section. The description of the analysis of the power output of the turbine is very small with respect to the methodology of the experiments and the simulations. Moving this description in the general methodology section, before the fairly long experimental and aerodynamic analyses would make more difficult to follow the results of the power and energy loss analyses, in that readers would have to flip back and forth between the initial and final part of the paper to check details of the methodology at the start of the manuscript required to follow the results of the end.
We have, however, introduced two sub-sections, making more clear that the first part refers to methodology and reminder of the section regards results.
- Q31.R2 With reference to their comment on line 646 of the first submission, RC2 asks: "Have the measured polars been extended to 0-360deg? If yes how? Please state in the relevant methodology section."
- A31.R2 The measured lift and drag curves have been extrapolated over the range -180° to $+180^\circ$ by means of a Python implementation of Viterna's method made available by NREL. This has been reported in the updated manuscript.
- Q32.R2 Line 650 of first submission: "I think this has already been stated in this paragraph."
- A32.R2 Apologies, this is correct. We removed the repetition in the revised manuscript.
- Q33.R2 Line 651 of first submission: "I think this is also duplicated"
- A33.R2 This is correct, thank you. We have removed also this duplication.
- Q34.R2 With reference to caption of Fig. 15 of first submission, RC2 asks: "should this read between clean and rough? It would be very helpful if the authors used clear definitions (incl. equations) of all quantities".
- A34.R2 Thank you for this helpful suggestion. We have replaced preposition "of" with "between" in the indicated figures and all other with a similar content.
We have also added a new equation defining the considered percentage differences. We have also clarified all other definitions in the reminder of the revised manuscript.
- Q35.R2 RC2 writes: "The presented benchmark can be an extremely useful exercise. Unfortunately, the conclusions section remains purely descriptive. Based on the large number of simulations and researchers involved, can the authors present a section

dealing with best-practice guidelines? If these cannot be generalized, then they could be tool or method specific. This would significantly increase the significance and impact of the paper for anyone dealing with similar challenges in the future.”

A35.R2 We thank the reviewer for this important note. As reported in our answer A4.R2, the reviewer’s advice has been adopted in the conclusions of the revised manuscript. It was possible to provide some recommendations for Best practice in simulations (e.g. verifying the applicability of the standard $k - \omega$ SST model as it is or consider changing some parameters (e.g. a_1), try to control in best possible manner level of TI around airfoil in relation to target conditions, and avoid use of wall-functions). We also provided recommendations on how the type of benchmark presented in this article could be improved in future exercises of this type, e.g. thorough investigations of the impact of airfoil profile discontinuities arising from the application of sand paper, aimed at defining less uncertain values of ”aggregate” K_s values accounting for both grain-induced roughness and steps).

Q36.R2 RC2 asks about the availability of the numerical data (results) of this article:
”What about the numerical data? Are they uploaded on an online repository too? This would also be very helpful.”

A36.R2 Thank you for this excellent idea. We have created a public repository on Zenodo containing all results presented and discussed in the article, i.e. all data used to generate all figures of the paper. The repository, available on Zenodo at <https://doi.org/10.5281/zenodo.22810001>, will be activated after the article is accepted. The data can be made available to the reviewer at any time during the revision process if needed. The link has already been added in the revised manuscript at the end of Section ”Conclusions” and in the section on Data Availability.

Q37.R2 RC2 asks why the AIST values of δ_w/c in Tables A2 and A3 are larger than those of all other CFD users; they also ask to provide the y^+ values of the CFD simulations:
”This value is two to three orders of magnitude greater than other approaches in the same table.
Please state the relative y^+ value for each simulation.
The same for Table A3.

A37.R2 The AIST simulations of the clean airfoil were initially using wall functions, which allow a larger minimum wall-normal spacing, since the innermost part of the turbulent BL is not resolved. In the revised article, AIST resolved the BLs of the clean airfoil all the way down to the wall, and wall functions were not used. Thus the considered variable has become comparable to that of all other participants.
A suitably selected estimate of y^+ has been added for all smooth airfoil simulations in Tables 2 and 3.

Q38.R2 With reference to PhD Thesis of Ehrmann 2017, RC2 reports: ”This thesis was not publicly available, when I checked. Could this reference be replaced by a public document?”

A38.R2 Thank you for noting this. We have removed this bibliographic entry and replaced it with a Sandia report by the same author available online, which contains all relevant information for which the previous entry was cited.

Response to comments and suggestions of Prof. G. Leishman

We thank very much Prof. Leishman for his insightful comments and suggestions. They were very helpful in guiding our revision from the start of the revision process. We believe that all comments and recommendations of Prof. Leishman have been addressed and resolved in the revised version of our manuscript.

In particular, we have removed the majority of the equations of CFD methods that can be found in literature and/or were not strictly relevant to the main focus of our study. The equations we have left in the revised manuscript are those of models which have a direct impact on the discussion of all presented results and are not used by all CFD set-ups.

The last subsection of Section 3 now reports a comprehensive overview of all input parameters of the CFD set-ups that directly influence the smooth- and rough-LE analyses presented, compared and discussed in this study.

Section Conclusions has been rewritten, highlighting the methods which appeared to produce results more consistent with the experimental observations (thus highlighting aspects of Best Practice in the simulation of the considered problems), as well as areas where methodology improvements requiring further research are deemed necessary.

Editor

Q1.ED Dear authors, please ensure that you provide a detailed response to the comments of both anonymous reviewers and the public comment, and revise the manuscript accordingly.

Furthermore, one of the reviewers commented on the range of the y-axis in the plots. I understand the the range of the axes has been selected to accommodate the maximum values of the aerodynamic parameters across all conditions (clean, eroded, and severely eroded). It is clear that this choice is intentional and aims to illustrate the quantitative impact of the various surface conditions. However, still in some cases the comparison between different models is difficult. Please try to somehow address this point.

A1.ED We thank the Editor for his comments and recommendations. We confirm we have carried out a thorough revision and update of the manuscript addressing in the best possible way we could all the comments and recommendations of the anonymous reviewers and the community. We have reformatted all figures of the article and believe we have substantially improved the quality and accessibility of these too.