Physics of Plasmas

Validation metrics for turbulent plasma transport

Cite as: Phys. Plasmas **23**, 060901 (2016); https://doi.org/10.1063/1.4954151
Submitted: 07 March 2016 • Accepted: 05 June 2016 • Published Online: 22 June 2016

C. Holland

Validation metrics for turbulent plasma transport

C. Holland[a]
Center for Energy Research, University of California, San Diego, La Jolla, California 92093-0417, USA

(Received 7 March 2016; accepted 5 June 2016; published online 22 June 2016)

Developing accurate models of plasma dynamics is essential for confident predictive modeling of current and future fusion devices. In modern computer science and engineering, formal verification and validation processes are used to assess model accuracy and establish confidence in the predictive capabilities of a given model. This paper provides an overview of the key guiding principles and best practices for the development of validation metrics, illustrated using examples from investigations of turbulent transport in magnetically confined plasmas. Particular emphasis is given to the importance of uncertainty quantification and its inclusion within the metrics, and the need for utilizing synthetic diagnostics to enable quantitatively meaningful comparisons between simulation and experiment. As a starting point, the structure of commonly used global transport model metrics and their limitations is reviewed. An alternate approach is then presented, which focuses upon comparisons of predicted local fluxes, fluctuations, and equilibrium gradients against observation. The utility of metrics based upon these comparisons is demonstrated by applying them to gyrokinetic predictions of turbulent transport in a variety of discharges performed on the DIII-D tokamak [J. L. Luxon, Nucl. Fusion **42**, 614 (2002)], as part of a multi-year transport model validation activity.
[http://dx.doi.org/10.1063/1.4954151]

I. INTRODUCTION

All models are wrong, but some are useful.
—*George E. P. Box*

The concept of model validation, defined as[1]

The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model

has become an increasingly important part of research and engineering efforts across many disciplines, as numerical simulation results play an ever larger role in diverse areas, such as the design of buildings, aircraft, and automobiles, weather forecasting and climate modeling, and epidemiology. One such discipline is fusion energy research, where development of a validated predictive modeling capability is now commonly identified as a top priority for the US fusion energy program.[2,3] In the nearest term, such a capability is desired to maximize the scientific returns from future burning plasma experiments such as ITER,[4] while minimizing the risks of operating in regimes or configurations likely to damage the extremely expensive device infrastructure. On a longer timescale, the hope is that a sufficiently accurate predictive modeling capability can be applied to confidently design future experimental devices and reactors, reducing the cost and time to deploy fusion energy as a viable commercial energy source. More broadly, the conceptual models and computational codes that constitute a predictive modeling capability represent the most tangible realization of our understanding of the fundamental nonlinear plasma dynamics that govern fusion plasmas. As such, validation studies which assess the accuracy and fidelity of these models in predicting (or postdicting in many cases) experimental observations represent one of the clearest and most direct means of quantifying our progress in developing a true understanding of fusion-relevant systems. More importantly, through these assessments, validation allows us to rigorously identify those parameter regimes where current models do not provide sufficiently accurate descriptions of reality, and where improvements to the models are needed.

Given the importance of rigorous validation studies in many fields, it is no surprise that there is now a broad and well-established literature in the field, ranging from guidelines and best practices prescribed by professional societies[1,5] to journal reviews[6–11] and textbooks,[12,13] in addition to the numerous articles and reports detailing the outcomes of individual studies. In the area of plasma physics and fusion energy research, key review articles include those by Terry *et al.*[14] and Greenwald.[15] These articles lay out many of the key fundamental ideas and concepts for validation research in the context of magnetic confinement based fusion energy (MFE) research. A common feature of both these reviews and the broader literature is the identification of validation metrics as key components of any serious, robust validation study. While both the Terry *et al.* and Greenwald reviews discuss these metrics in some detail, including potential mathematical formulations, both also emphasize the need for further work. This paper aims to build upon those discussions, by focusing on the practical issues often encountered in formulating validation metrics for plasma physics studies. In order to provide a "hands-on" illustration

[a]Electronic mail: chholland@ucsd.edu

of these issues, this paper will use the formulation of validation metrics relevant for turbulent transport in MFE plasmas as a type of worked example. However, the underlying concepts and challenges are broadly relevant to a wide variety of different phenomena. The most important of these concepts are the need to incorporate both experimental and model uncertainties into the validation metrics, as a fundamental part of any assessment of model performance, and using synthetic diagnostics for meaningful model–experiment comparisons.

The remainder of this paper is structured as follows. In Sec. II, an overview of key validation concepts is presented, including both formal and "plain English" definitions. In Sec. III, a brief review of turbulence and transport modeling in MFE plasmas is presented, including a review of the widely used global transport validation metrics defined in the ITER physics basis[16] in Sec. III A. Sec. IV then presents an alternate approach based on local transport metrics, which builds upon the pioneering work by Ross and co-workers.[17,18] Extensions of this approach to local turbulence characteristic validation metrics, including the role of the synthetic diagnostics, are detailed in Sec. V. In Sec. VI, construction of composite metrics is discussed, along with the related concept of the primacy hierarchy, while conclusions and future research directions are presented in Sec. VII.

II. OVERVIEW OF KEY VALIDATION CONCEPTS

A useful entry point to the study of model validation is to begin by reviewing and defining the terminology commonly used in the literature. In this paper, the definitions given by Terry *et al.*[14] will be used. As noted above, alternative definitions and terminology discussions can be found in professional association guidelines,[1,5] journal reviews,[6–11] and textbooks.[12,13] A summary of simple "plain English" descriptions (not definitions) of the key concepts is provided in Table I.

As a starting point, one must specify what the term model means in model verification and validation. The AIAA guidelines define a model as "*A representation of a physical system or process intended to enhance our ability to understand, predict, or control its behavior*." Because this definition is so broad, a more practical starting point is to follow Terry *et al.* and distinguish between a conceptual model and computational model. A conceptual model is defined as "*The observations, mathematical modeling data, and mathematical (partial differential) equations that describe the physical system, including initial and boundary conditions*," while a computational model is "*The program or code that implements the conceptual model*." The key here is the separation of the system of equations under consideration, along with specifications of domain geometry and initial and boundary conditions, from their specific computational implementation. Thus, what is often referred as a model in plasma physics literature corresponds to the conceptual model, regardless of whether the equations under consideration are obtained directly from fundamental physics relationships such as the Boltzmann equation or Maxwell's equations (sometimes termed "first-principles modeling" or even just "theory"), or incorporate some amount of free parameters whose values are determined via analytic (e.g., moment closures used in gyrofluid theory[19]) or empirical (e.g., confinement scaling laws derived from database regressions[16,20]) methods. In this paper, this shorthand of using the term model to refer to the conceptual model will be used frequently, while the term code will be used to refer to a particular computational model. Also implicit throughout this discussion is that the conceptual model(s) being tested are sufficiently complex as to require numerical computation to obtain solutions sufficiently accurate for meaningful quantitative comparison against experiment. In some cases the numerical implementation may be fairly trivial, but nonetheless some level of computation is always needed in validation studies.

TABLE I. "Plain English" descriptions of common V&V terms.

Term	Description
Model	The set of equations being tested against experiment
Code	The numerical implementation of the model
Verification	"Does the code solve the model equations correctly?"
Validation	"Does the model have the right equations?"
Validation metric	A measure used to quantify the experimental fidelity of the model

As most every set of physics-based model equations is based upon some set of simplifying assumptions and mathematical orderings, it is important to understand the limits these assumptions place on when and where the model is expected to be valid, or at least useful. Identifying the domain in parameter or operational space where a model is expected to perform acceptably is termed model qualification, formally defined as the "*theoretical specification of the expected domain of a conceptual model and/or approximations made in its derivation*."[14] Experimentally assessing whether a given model performs acceptably (as quantified by validation metrics) over this theoretically specified domain is one of the most important functions careful model validation studies provide.

Given the distinction between conceptual and computational models, the ideas of verification and validation naturally arise. Verification is formally defined as "*The process of determining that a model implementation accurately represents the developer's conceptual description of the model and the solution to the model*." It is distinct from validation, defined as "*The process of determining the degree to which a model is an accurate representation of the real world from the perspective of the intended uses of the model*." However, the two are clearly related and often discussed together, leading to the commonly used acronym "V&V" to refer to verification and validation. Thus, one might describe (but should not define) verification as asking "Does a code solve the model equations correctly?" and validation as "Does a model have the right equations?"[12] Although a detailed discussion of verification is outside the scope of this paper, it is important to emphasize that verification encompasses far more than elimination of coding errors and other software quality assurance activities (both of which are themselves intensive efforts). Verification also includes quantification of inherent numerical errors due to sources such as discrete spatiotemporal grids, accuracy of time-integration algorithms, and noise accumulation in Monte Carlo simulations, all of which must be known to assess whether a given code yields sufficiently accurate solutions of the model equations for their intended use. It is also clear that verification must precede validation. One cannot assess the experimental fidelity of the conceptual model without first knowing the accuracy of the computational model it is implemented in, for the specific parameters or regime being considered. The process of ensuring that the model equations and computational algorithms are correctly implemented is known as *code verification*, while testing that a particular code result has been correctly calculated to the desired accuracy is known as *solution verification*. Note that one must verify all computational tools used in a validation study, including synthetic diagnostics (discussed further in Sec. V A).

Finally comes the concept of validation metrics, defined by Terry *et al.* as "*A formula for objectively quantifying a comparison between a simulation result and experimental data*." Similarly, Oberkampf and Roy define a validation metric as "*A mathematical operator that measures the difference between a system response quantity (SRQ) obtained from a simulation result and one obtained from experimental measurements*."[13] Thus, validation metrics are the mathematical relationships used to quantify how closely the model predictions reproduce the experimental observations of interest. As such, they provide the basis for assessing the fidelity of a given model, and establishing confidence in using that model predictively. This formal quantification of model accuracy, as opposed to more subjective assessments of model fidelity often found in the literature such as "good agreement," "qualitatively consistent," or "clearly differs," is what distinguishes validation as an activity distinct from many earlier model–experiment comparison studies. Indeed, Oberkampf *et al.*[9] state that "*[t]he specification and use of validation metrics comprises the most important practice in validation studies*."

Examination of the broader validation literature indicates a number of characteristics well-designed validation metrics should incorporate,[10,11] which will be discussed in detail in Sec. IV. In the author's opinion, the most important of these characteristics is the incorporation of experimental, modeling, and computational uncertainties and errors into the metric's assessment of model accuracy. Uncertainties and errors are a fundamental reality of both experiment and simulation, and without accounting for them no meaningful assessment of model validity can be made. Without such an assessment, one cannot build confidence in the predictive capabilities of a model, undermining the entire goal of a validation study. As a practical illustration, consider the common case where one has a set of model predictions with associated uncertainties $y_M(x) \pm \sigma_M(x)$ (e.g., predictions of energy confinement time in a tokamak as a function of plasma current) to be compared against a corresponding set of experimental data $y_E(x) \pm \sigma_E(x)$. One can only assess the level of model accuracy (i.e., the difference between $y_M(x)$ and $y_E(x)$) to a level of precision set by the combination of uncertainties σ_E and σ_M. Put another way, without knowledge of σ_M and σ_E, one cannot assess whether differences (or agreement) between model predictions and experiment are truly meaningful or relevant for building confidence in the predictive capabilities of the model. The processes used to quantify experimental and model uncertainties are often referred to in the literature as uncertainty quantification (UQ). Research into sophisticated UQ methods, particularly into efficient and robust approaches for propagating uncertainties in model inputs into uncertainties in model outputs, is currently a rapidly growing field in its own right with entire journals dedicated to its study,[21] and some aspects of this work will be touched upon in Sec. IV and V of this paper.

To illustrate the importance of incorporating uncertainties into validation metrics, consider the progression of metric visualizations discussed in Oberkampf *et al.*,[9] reproduced here in Fig. 1. At one end of the spectrum lies the purely qualitative "viewgraph norm" comparison (Fig. 1(a)), which consists of two figures separately visualizing (presumably comparable) experimental data and simulation results. While such an approach may provide useful insights into the gross performance of the model, most often for experts in the problem under consideration who have already developed an intuition for where the model is most likely to perform well or poorly, it does not provide a quantitative assessment of model fidelity. Judgments based on such visualizations are always subjective, and easily influenced by the presentation (e.g., due to choice of colors used in visualizing the data[22]). Moreover, in many cases such as the comparisons of contour plots as shown in Fig. 1(a), no information is (or can be) included as to the experimental and model uncertainties, which prevents one from determining whether any suggested (dis)agreement is (un)fortuitous. A significantly improved comparison is shown in Fig. 1(b), which directly visualizes magnitudes and trends in both the measured and simulated quantity of interest (the system response) as a function of a given parameter (the system input). However, without inclusion of uncertainties, the significance of differences between model and experiment still cannot be assessed. Figs. 1(c)–1(e) address this issue through increasingly comprehensive inclusion of uncertainties in both the input variable and the measured and predicted system responses. This progression culminates in Fig. 1(f), which represents a robust quantitative comparison between the experiment and simulation. Note that it transitions from separate plotting of the simulated and measured responses as a function of the input variable to plotting their difference, and that it includes visualization of the total uncertainty in both the difference of the predicted and measured SRQ and system input parameter (here one and two standard deviation contours are visualized). Although Fig. 1(f) contains the same information as Fig. 1(e), it provides a clearer visualization of both the discrepancy between model and experiment and the statistical significance of this difference as a function of input parameter, and as such forms the most robust basis for a validation metric. Developing metrics (and their corresponding visualizations) of this form suitable for plasma turbulence modeling is the focus of Secs. IV–VI of this paper.

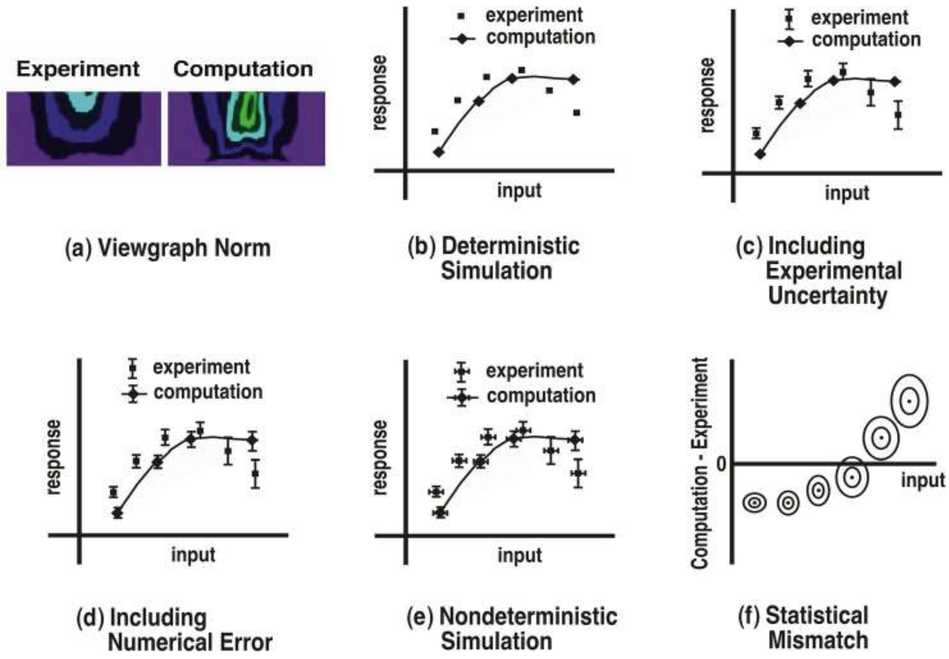

FIG. 1. Illustration of validation metrics of increasing quality and utility. Reprinted with permission from Oberkampf *et al*., App. Mech. Rev. **57**, 345 (2004). Copyright 2004 ASME Publishing.[9]

As the above discussion makes clear, the process of validation rests upon comparisons of model predictions to experimental data. The astute reader might therefore note a contradiction between the necessity of experimental data for validation, and the goal of much fusion-related validation work, which is to build confidence in models used to predict future experiments that lie outside the parameter space spanned by current-day experiments. In the strictest and most rigorous sense, validating a model for one experimental condition or regime does not necessarily imply anything about its predictive capability for regimes where it has not been validated. Thus, demonstrating a model that works well in describing current-day experiments does not necessarily mean that it will perform as well in future burning plasma experiments such as ITER. This tension is a specific instance of a general challenge for validation research, namely, understanding to what degree validation of a model for a given area of parameter space enables meaningful *extrapolation* of the model to other parameter regimes.

Although the tension between validation and extrapolation has not been explicitly discussed much within the fusion community to date, the implicit approach taken to resolving it has been to focus on understanding the relevant underlying physical processes in great detail, rather than investing heavily in empirical models calibrated to current experiments. This approach is based upon the idea that while the absolute values of many plasma quantities (such as plasma volume, density, temperature, etc.) will be different in future experiments from current ones, many of the fundamental physical processes governing the dynamics of those future plasmas will be the same. Thus the focus of much current work is on developing models of these processes built upon physical understanding, and eliminating free parameters and dependencies determined by calibration of the models to current experiments. This approach is mostly clearly seen in the widespread use of various dimensionless parameters (such as the ion-electron mass ratio m_i/m_e and ratio of plasma thermal to magnetic pressure $\beta = 2\mu_0 nT/B^2$) to characterize plasma regimes, and formulation of model behavior in terms of scalings with these dimensionless parameters. When combined with application of dimensional analysis and scale invariance techniques to the model equations, understanding gained from the study of current-day experiments can be extrapolated to future devices with increased confidence, in a manner analogous to how wind tunnel experiments are used in the design of aircraft and automobiles. This approach has been employed with great success to advance our understanding of transport physics by combining data from multiple experimental devices with different physical parameters (size, density, temperature, etc.) but matched dimensionless parameters.[23–25] In this context, basic plasma physics and smaller-scale experiments focused on specific plasma dynamics play a key role in MFE plasma model validation, by enabling access to parameter regimes significantly different than those obtainable in large high-power fusion experiments. Such experiments also often allow use of diagnostic approaches and measurement techniques not available on larger devices, enabling testing of predictions of phenomena which cannot otherwise be measured easily, if at all. While the physics-based approach to model extrapolation offers many advantages, it is important to keep in mind its key limitation, which is that it presumes that no significant new physics or dynamics arise in future devices that are not captured in the model equations. Thus, while confidence in extrapolation to future regimes can be increased through better physical understanding, there is no substitute for actual experimental data for truly validating a model in the parameter regime of interest.

III. BASICS OF TURBULENCE AND TRANSPORT MODELING IN MAGNETICALLY CONFINED FUSION PLASMAS

One of the greatest challenges for plasma theory is developing an accurate understanding of plasma confinement: the physics which relates the equilibrium density n and temperature T of magnetically confined plasmas to externally applied fueling and heating sources. Since plasma confinement in most MFE devices is believed to be determined by plasma turbulence, building, validating, and qualifying models of this turbulence is essential for establishing an accurate predictive capability for designing and optimizing future MFE devices. Given the importance of this topic to predicting future performance, as well as the extensive active research efforts in MFE plasma turbulence measurements, simulation, and validation, we have chosen to use it as the example problem for this tutorial. In this section, we briefly review the fundamental physics of plasma confinement and approaches taken to modeling it, to inform the development of turbulent transport validation metrics in Secs. IV and V. This discussion is intended to provide a complete and concise introduction to turbulent transport in magnetically confined plasmas for those who are not already expert in the area. Those who are experts can likely skip ahead to Sec. III A for a discussion of global transport metrics or Sec. IV for discussion of local turbulence metrics.

There are a number of different approaches taken to quantifying and understanding confinement in MFE-relevant plasmas discussed in the literature, including the transport and confinement chapters of the original[16] and updated[20] ITER physics basis reports, which form the basis of the discussion here. The most basic measure of plasma confinement is the energy confinement time τ_E, which is equal to the ratio of the volume-integrated stored energy of the plasma $W_p = \int dV\, 3nT$ to the injected power P_{inj}. Since the total fusion power produced in a given experiment is proportional to the square of the stored energy, models which describe the dependence of τ_E on various global system parameters such as toroidal magnetic field and total plasma current provide a compact way of roughly predicting the net power production (and thus performance) expected for a given device configuration and heating scenario. A variety of empirical and analytic scaling laws for τ_E have been derived, which are often used as measures of performance in current experiments. For instance, the H_{89} and $H_{98(y,2)}$ measures are the ratios of the energy confinement time in a given plasma to the empirical ITER89P[16] or IPB98(y,2)[20] scaling laws for τ_E which have been extensively used in the design of ITER target scenarios.[26] Each of these measures provides a basis for assessing the level of confinement achieved in a given experiment relative to what is required for a given ITER operational scenario, and a significant fraction of current experimental work is focused on demonstrating plasmas which can maintain these normalized confinement levels in different operational scenarios (such as non-inductive steady-state operation,[27] edge localized mode (ELM)[28] suppression via application of resonant magnetic perturbations,[29] or operation in the presence of "ITER-like" metal walls[30]).

While the energy confinement time provides a useful global measure of plasma confinement, it is not sufficient for accurate predictions of many important aspects of reactor performance such as plasma stability, because it does not contain information about the spatial variations or dynamics of the equilibrium density and temperature profiles. In a toroidally axisymmetric plasma (to which we will restrict our attention from here onwards), and absent large-scale (i.e., device-size) instabilities such as sawteeth or tearing modes, these profiles are approximately constant on closed magnetic flux surfaces, essentially as a consequence of the plasma equilibrium satisfying the force balance equation $\vec{J} \times \vec{B} = \vec{\nabla} P$ which implies that $\vec{B} \cdot \vec{\nabla} P = 0$. However, as will be seen below, small deviations and fluctuations about this equilibrium result in particle, energy, and momentum fluxes which transport the plasma across flux surfaces, limiting the level of confinement achieved. Describing the relationship between the dynamics of these kinetic profiles (i.e., the equilibrium density, temperature, and rotation profiles) and their associated sources and sinks is the goal of transport modeling, and the equations used to describe this relationship are called the transport equations. These equations are generally formulated as a set of one-dimensional (1D) fluid balance equations of the form[31]

$$\frac{d\langle n_j\rangle}{dt} = \frac{1}{V'}\frac{d\left(V'\,\Gamma_j\right)}{dx} + S_{n,j}, \qquad (1)$$

$$\frac{d\langle W_j\rangle}{dt} = \frac{1}{V'}\frac{d\left(V'\,Q_j\right)}{dx} + \Pi_j\frac{d\Omega_{tor}}{dx} + S_{int,j} + S_{aux,j}, \qquad (2)$$

$$\frac{d\left(\langle R^2\rangle\Omega_{tor}\sum_j n_j M_j\right)}{dt} = \frac{1}{V'}\frac{d\left(V'\,\sum_j \Pi_j\right)}{dx} + \sum_j \mathcal{T}_j, \qquad (3)$$

for each species (ions or electrons) j. Here, the brackets denote an average over magnetic flux surface, $W_j = (3/2)\, n_j T_j$ is the energy of species j, $\Omega_{tor}(x)$ is a common toroidal rotation frequency of all species, $S_{n,j}$ corresponds to the particle source and sink terms, S_{aux} to auxiliary (external) heating and fueling systems such as neutral beams or various wave-based systems such as ion or electron cyclotron resonance heating, S_{int} to internal heating and cooling processes such as ionization, recombination, radiation, and collisional energy exchange, and $\mathcal{T}_j$ to auxiliary momentum sources. The quantities Γ_j, Q_j, and Π_j correspond to the magnetic flux-surface-averaged radial fluxes of particle, energy, and toroidal angular momentum, respectively. There are a variety of possible choices for radial coordinate x, each of which provides a unique label for a given closed magnetic flux surface. Two of the most common choices are the square root of normalized toroidal magnetic flux $\rho_{tor} = \sqrt{\chi_{tor}(x)/\chi_{tor}(a)}$ or the normalized poloidal magnetic flux $\psi_N = \{\psi - \psi(0)\}/\{\psi(a) - \psi(0)\}$, where χ_{tor} and ψ are the unnormalized toroidal and poloidal magnetic fluxes $\int \vec{B} \cdot d\vec{A}$ passing through the a surface labeled by x, respectively, and a is the label of the last closed flux surface (LCFS).

Given these equations, and specifications of the vacuum toroidal magnetic field, assorted initial and boundary conditions, and auxiliary sources, one can then predict the evolution (or steady-state profiles) of the plasma equilibrium density, temperature, rotation, and current, for a given model of the plasma fluxes Γ_j, Q_j, and Π_j. The level of plasma confinement obviously depends directly upon these fluxes—the larger the fluxes (in particular, the energy flux Q), the faster the energy leaves the system, and thus the poorer the confinement. The irreducible minimum for these fluxes is determined by collisional diffusion, the details of which can be derived via use of generalized Chapman–Enskog theory in analogy to neutral fluids; this approach is detailed in the classic review paper by Braginskii.[32] The key difference from neutral fluid collisional transport is that in a well-magnetized plasma, the typical diffusive step-size for cross-field (i.e., radial) transport is given the gyroradius $\rho = v_T/\Omega_c$ of the species in question (where $v_T = \sqrt{k_B T/m}$ and $\Omega_c = qB/mc$ are the thermal velocity and cyclotron frequency, respectively) rather than the mean free path $\lambda_{mfp} = v_T/\nu_{coll}$ (where ν_{coll} is a collisional scattering rate), leading to radial energy fluxes of the form $Q = -n\chi dT/dx$, with χ scaling as $\rho^2\nu_{coll}$. In toroidal plasmas, additional complications arise due to the $1/R_{maj}$ dependence of the toroidal magnetic field (where R_{maj} is the major radius) that leads to trapped particles and larger values of χ. Collisional transport in this case is described by neoclassical theory.[33,34] The practical implication for toroidal MFE devices is that while this neoclassical transport is often an order of magnitude larger than collisional processes in cylindrical geometry, it is still sufficiently small that if it were the only process acting, the needed confinement for net energy gain could be obtained in relatively small devices.

Unfortunately, many tokamaks (as well as other MFE devices) observe thermal confinement levels approximately 10–100 times worse than what is expected from neoclassical transport theory. In most of these devices, this difference is now commonly ascribed to the presence of "microturbulence"—small scale (i.e., correlation lengths much less than the plasma minor radius), small amplitude ($\tilde{n}/n_0 \ll 1$ where $\tilde{n}$ is the density fluctuation and n_0 the equilibrium density) fluctuations driven by the inherent cross-field gradients of the equilibrium plasma density, temperature, and rotation.[35,36] These fluctuations nonlinearly couple and exhibit collective behavior which manifests as cross-field fluxes of the form $Q_{turb} = (3/2)\langle\tilde{p}\,\tilde{v}_r\rangle$ (where $p = nT$, $\tilde{p}$ is the pressure fluctuation, and $\tilde{v}_r$ the radial velocity fluctuation), which act to reduce the driving equilibrium gradient(s) and thereby limit confinement achieved. Given the number of potential free energy sources—the cross-field gradients of density, temperature, and toroidal rotation of multiple ions as well as electrons—it is perhaps not surprising that a correspondingly wide array of instabilities driven by these gradients has been identified. These instabilities are often classified by their dominant driving mechanism. In current tokamak plasmas, the dominant instabilities are generally observed to be the ion temperature gradient (ITG) mode which operates on ion gyroradius ρ_i scales and the corresponding electron temperature gradient (ETG) mode which operates on electron gyroradius ρ_e scales, as well as the trapped electron mode (TEM) driven by both the electron density and temperature gradients, and which spans the ion and electron gyroradius scales (where it smoothly transitions to the ETG mode). Beyond these modes, both resistive and kinetic ballooning modes may be unstable (most often in the near-edge and pedestal regions close to the LCFS of the plasma), while microtearing modes may be unstable at sufficiently high normalized plasma pressure $\beta = 2\mu_0 nT/B^2$ and collisionality. In many (if not all experiments), multiple instabilities are simultaneously present, operative, and nonlinearly coupling with each other throughout the plasma. Thus, in order to carry out accurate predictive transport modeling, we must develop microturbulence models which can correctly describe the dependence of the cross-field turbulent particle, energy, and momentum fluxes on various driving gradients and parameters as the mix of instabilities present changes. Additional information on the physics of these instabilities can be found in a variety of review articles and textbooks on the subject, such as those by Horton[35] and Weiland.[36]

Because these instabilities generally have small amplitudes, characteristic cross-field correlation lengths on the order of 1–10 ρ_i or smaller, much smaller than the plasma minor radius a, and characteristic frequencies small relative to the ion cyclotron frequency, they are most accurately described via the coupled gyrokinetic-Maxwell equations.[31,37] In their most common form, these equations describe the self-consistent gyromotion-averaged dynamics of small fluctuations $\tilde{f}(\vec{X},\vec{v},t)$ in the ion and electron distribution functions and their associated electromagnetic fields, for a specified set of equilibrium electric and magnetic fields and equilibrium kinetic distribution functions $f_0(\vec{X},\vec{v},t)$ which include the radially varying equilibrium density, temperature, and rotation profiles that drive the turbulence. This model is often referred to as the "δf" formulation since it is based upon the assumption that $\delta f = \tilde{f}/f_0$ is a small parameter on the same order as ρ/L, where ρ is the gyroradius of the species in question, and L a characteristic equilibrium scale length such as the major radius R_{maj}, minor radius a, or an equilibrium gradient scale length. A kinetic rather than fluid approach is generally required for accurate description of microturbulence in order to correctly capture important velocity-space dynamics such as particle trapping and Landau damping which play significant roles in determining mode growth rates. However, sophisticated generalized fluid equations have been developed which can approximately capture the dominant gyroaveraging and velocity-space effects through combinations of higher-order moments and closure formulations.[38–45] The most sophisticated of these gyrofluid models can accurately reproduce many characteristics of the full gyrokinetic model (including electromagnetic, geometric shaping, and collisional effects) at significantly reduced computational cost.[46]

Another important aspect of these equations is that in their conventional "δf" formulation (and corresponding gyrofluid reductions), they can be derived as part of a formal expansion theory in $\rho^* = \rho_i/a$, starting from the Fokker–Planck equation. This hierarchy can be summarized as requiring (in the limit of small rotation) that:

1. to order unity, each equilibrium guiding-center distribution is a time-stationary Maxwellian with constant density and temperature on a given magnetic flux surface,
2. to 1st-order in ρ^* fast, small-scale fluctuations in the distribution function are described via the gyrokinetic equations, while slowly varying, large-scale distribution function fluctuations and corrections are described by the drift-kinetic equation,

3. to 2nd-order in ρ^* the equilibrium profiles slowly evolve due to the fluxes arising from the turbulent and neoclassical processes, as well as internal and external sources, as described by the transport equations Eqs. (1)–(3).

From the perspective of validation, the greatest advantage of this approach is that it provides a rigorous definition of the conceptual model to be validated against experiment, which provides clarity in formulating the desired comparisons. The primary drawback is that the idealized physical situation this ordering describes—the slow evolution of equilibrium profiles on perfectly nested axisymmetric flux surfaces due only to neoclassical and small-amplitude fluctuations—is in practice almost never realized experimentally. For example, large-amplitude fluctuations near the LCFS, non-axisymmetric equilibria arising from both internal (e.g., sawteeth[47] and tearing modes[48]) and external (e.g., error fields[49] and resonant magnetic perturbations[29]) processes,[50] and rapidly varying external heating sources operated in feedback to maintain plasma performance[51] all lead to violations of the formal ordering outlined above in different regions of the plasma. However, in many cases, these violations are weak, or localized to certain regions of the plasma, and the underlying physical picture embodied in this model is believed to represent a useful practical paradigm for understanding and predicting plasma confinement.

As might be expected, there are very few (if any) analytic solutions to the gyrokinetic-Maxwell equations useful for quantitative predictions of turbulent transport, and so numerical solution (i.e., a code implementing the model) is required. The most accurate solutions are obtained through nonlinear, initial value simulations of either the gyrokinetic-Maxwell equations or their fluid variants, analogous to direct numerical simulation (DNS) of neutral fluid turbulence. A variety of such codes have been developed, which can generally be divided between those taking a continuum approach[52–56] and those that use particle-in-cell methods.[57–62] As in neutral fluid DNS, both approaches initialize the simulation with some very small amplitude fluctuations, which first grow exponentially at the linear growth rate(s) of the instabilities being considered (the "linear" phase), and then saturate at a finite amplitude set by the balance of these linear drives and nonlinear couplings between different wavenumbers (the "saturated" phase). The statistics of various quantities (such as mean energy flux or fluctuation power) from this saturated phase are then used for transport modeling predictions,[63,64] as well as comparisons with other models and experiments in V&V studies. Implicit in this approach is the assumption that the saturated phase represents a "statistical steady-state" from which well-converged estimates of the quantities of interest can be made, and that the results are independent of the initial conditions. Appropriate calculation of these statistics and their related uncertainties is discussed further in Secs. IV and V. Here, it is important only to emphasize that for virtually any V&V effort, it is only comparisons of these converged, initial-condition independent quantities that are meaningful. Claims based upon small averaging windows (or sample sizes) and early simulation times have no physical value or relevance, regardless of the computational cost required to obtain those results. Specifically for validation of plasma turbulence models, the appropriate comparison is between predictions of assorted statistical quantities derived from simulation and experimental data, and not specific time traces or "snapshot" visualizations (i.e., viewgraph norm comparisons).

Since these nonlinear gyrokinetic simulations can require 10^3 processor-hours or more (even 10^7 and beyond in some cases[65–68]) to calculate the statistics at a single location in the plasma, reduced models of the turbulence have been developed which can make predictions on the processor-minute or less timescale, to facilitate practical transport modeling with feasible resource requirements. The general approach of most such models is to decompose the turbulent fluxes into two components along the lines of (using the ion energy flux Q_i as an example) $Q_i = (3/2)\mathrm{Re}\langle\sum_k \tilde{p}^*_{i,k}\tilde{v}_{r,k}\rangle = (3/2)\mathrm{Re}\sum_k R_k^{p_i}\langle|\tilde{v}_{r,k}|^2\rangle$ where the angular brackets denote averaging over both magnetic flux surfaces and fast turbulent timescales, and the sum is a sum over wavenumbers k. In these models, the wavenumber-dependent ion pressure response function $R_k^{p_i} = \tilde{p}^*_{i,k}/\tilde{v}_{r,k}$ is generally calculated via direct solution of linearized (gyro)fluid equations. It is then convolved with a second model for the fluctuation spectral intensity $\langle|\tilde{v}_{r,k}|^2\rangle$ which may come from calibration of the model against experimental data,[69] analytic theory arguments,[36,70,71] or fits to databases of nonlinear simulations.[46,72,73] The underlying physical motivation for this approach is that unlike many neutral fluid turbulence systems, in the core region of MFE-relevant core plasmas, the fluctuations saturate at small amplitude and retain many of the linear wave characteristics of the underlying instabilities. As discussed above, this assumption of small amplitude levels often breaks down toward the LCFS, limiting the region of plasma to which these models can be appropriately applied. Obviously such "quasilinear" approaches contain many more approximations than the DNS approach, and as such effectively constitute separate conceptual models than the DNS approach. Nonetheless, they also often provide sufficient experimental fidelity at such greatly reduced computational cost as to currently be the only practical models for predictive and interpretive transport modeling. Finally, one should note that both approaches implicitly assume that there is a single, unique nonlinear solution for the specified input parameters. While to the author's knowledge there are no known counterexamples that disprove this assumption for numerical converged simulations of gyrofluid or gyrokinetic simulations, the possibility of multiple physical solutions cannot be formally ruled out.

More extensive discussions of the underlying physics of the various microturbulence-relevant instabilities are available in a wide variety of review articles and textbooks such as Refs. 35 and 36, and further details on their dynamics can be found therein. For the purposes of this paper, we need to consider only one particular defining feature of their dynamics—a hypersensitivity to the primary driving gradient. In many cases (most notably the ITG and ETG modes), there is a critical value of the driving gradient that must be exceeded for the mode to become unstable. Moreover, once this gradient is exceeded, the turbulent fluxes are observed to increase superlinearly, whereas the neoclassical collisional fluxes scale linearly with driving gradients (Fig. 2). This phenomenon is often referred to as transport stiffness in the literature (see, e.g., Refs. 74–78), and has two important implications. Experimentally, stiff transport means that once the critical gradient has been exceeded, it becomes increasingly hard to increase the core pressure by increased heating (hence the term stiff). From a modeling standpoint, first note that most microturbulence models predict the turbulent fluxes and fluctuation levels for a given (i.e., input) set of local parameters and gradients. The inherent stiffness of the turbulence magnifies any uncertainty in the driving gradient(s) into larger uncertainties in the predicted quantities, which must be included in any comparison against measured values of the predicted quantities. Confronting this property of the turbulence is therefore essential for any useful turbulent transport validation metric, and is discussed in detail in Sec. IV B. Note that since experimental determinations of these gradients are obtained from derivatives of profile measurements, the gradients can have non-negligible uncertainties even if the profile measurements themselves have small uncertainties.

TABLE II. ITER physics basis figures of merit for evaluating transport models. Adapted with permission from ITER Physics Basis Expert Groups on Confinement and Transport and Confinement Modelling and Database and ITER Physics Basis Editors, Nucl. Fusion **39**, 2175 (1999). Copyright 1999 IAEA.[16]

1: Ratio of incremental total stored energy $W_{inc}^{sim}/W_{inc}^{exp}$

where $W_{inc} = \frac{3}{2}\int dV\,\left(n_e\hat{T}_e + n_i\hat{T}_i\right)$ and $\hat{T} = T(\rho) - T(\rho_{BC})$

2: $(W_{inc}^{sim}/W_{inc}^{exp})_e$ and $(W_{inc}^{sim}/W_{inc}^{exp})_i$

(separate e and i)

3: $(n_{i,\rho=0.3}T_{i,\rho=0.3}W)_{sim}/(n_{i,\rho=0.3}T_{i,\rho=0.3}W)_{exp}$

4: $\chi^2 = [\sum(T_{sim} - T_{exp})^2]/N\sigma^2$, where σ is the expt. error and N the number of observations

5: $\beta_{sim}^{*2}/\beta_{exp}^{*2}$ where $\beta^{*2} = \int dV\, n_i^2 T_i^2$

6a: $\mathrm{STD} = \sqrt{\int_0^{\rho_{BC}} dx\,(T_{sim} - T_{exp})^2} \Big/ \sqrt{\int_0^{\rho_{BC}} dx\,T_{exp}^2}$

6b: $\mathrm{OFF} = \left[\int_0^{\rho_{BC}} dx\,(T_{sim} - T_{exp})\right] \Big/ \sqrt{\int_0^{\rho_{BC}} dx\,T_{exp}^2}$

A. Historical transport modeling metrics

Historically, the first widely utilized validation metrics in MFE transport and turbulence modeling were the six “figures of merit” detailed in the ITER physics basis,[16] listed in Table II. Most of these metrics are focused only upon core confinement, often defined as the region inside of some boundary radius ρ_{BC} (=0.9 in the ITER physics basis analysis). Of these six metrics, the most commonly used are metric 1, the ratio of predicted to measured incremental stored energy W_{inc} (Fig. 3), and metric 6, the normalized mean offset and root-mean-square (RMS) error between individual predicted and measured profiles.

These metrics have several appealing features. Foremost, they all have fairly simple mathematical forms corresponding to basic statistical measures (e.g., means and standard deviations). They are therefore easily accessible to a broad audience, particularly non-experts. The value of such simplicity and accessibility should not be underestimated, particularly for communicating the results of validation studies to management and decision-makers that may be removed from details of day-to-day research. These metrics also focus on quantifying the ability of transport models to accurately predict the key global measures of reactor performance, such as stored energy and $\langle\beta^2\rangle \propto Q_{fusion}$, which are the bottom-line quantities such modeling is intended to predict. However, the metrics also have several drawbacks. First, only figure of merit 4, which to the author’s knowledge has not been widely used in published MFE transport modeling validation studies, incorporates any estimate of experimental or model uncertainties. Previous studies using other figures of merit implicitly address this issue by assessing model performance using databases of experimental discharges (including those assembled by the ITER Topical Physics Activity Transport and Confinement working group[79]), and looking at ensemble statistics of the metrics (illustrated in Fig. 4), but this approach is not a completely satisfying substitute for explicitly confronting the

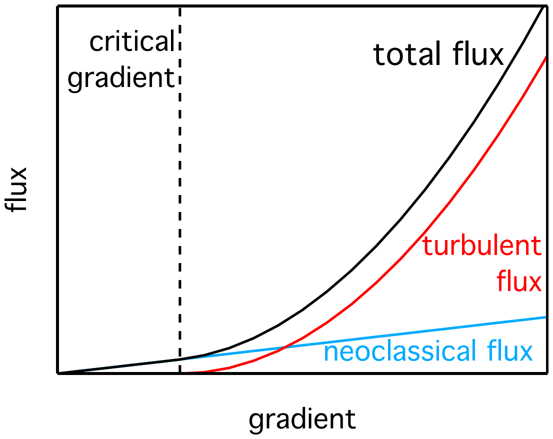

FIG. 2. Schematic illustration of transport stiffness, showing the scaling of the total energy flux (—) as a function of driving gradient. Once the critical gradient (– – –) for a given microturbulence instability is exceeded, the turbulent flux (red line) rapidly increases and quickly exceeds the neoclassical flux (blue line) to become the dominant component of the total flux.

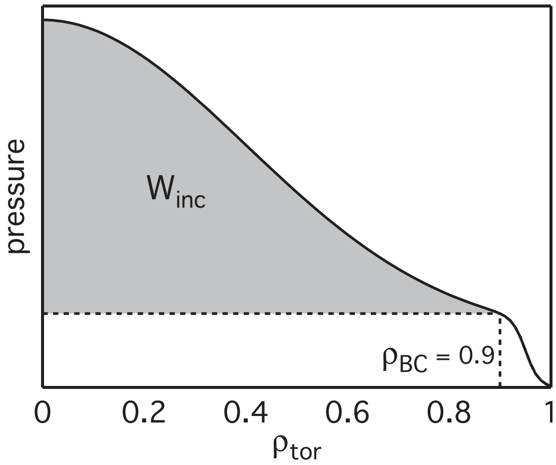

FIG. 3. Schematic illustration of the fraction of plasma pressure which contributes to the incremental stored energy W_{inc}, using the $\rho_{BC} = 0.9$ boundary condition of Ref. 16.

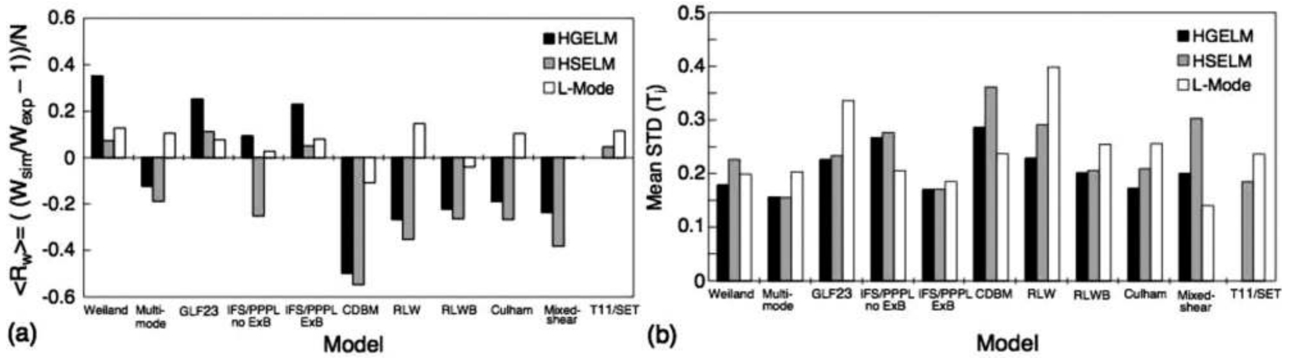

FIG. 4. Calculations of ensemble-averaged (a) incremental stored energy ratio (Figure of Merit 1) and (b) STD error (Figure of Merit 6a) in predicted T_i profiles for a variety of transport models and confinement scenarios. Reprinted with permission from ITER Physics Basis Expert Groups on Confinement and Transport and Confinement Modelling and Database and ITER Physics Basis Editors, Nucl. Fusion **39**, 2175 (1999). Copyright 1999 IAEA.[16]

uncertainties in the measurements of each discharge modeled. Second, because all of these metrics are global (in that they average over the entire simulation domain), one cannot easily determine where (or why) the model disagrees with the experiment. Namely, from these metrics alone one could not discriminate between a case where the model had only a significant error in a small fraction of the domain, or was off by a uniform amount everywhere. Finally, and most importantly for MFE transport modeling where the implicit assumption is that validation against current experiments will enable more confident extrapolation to the prediction of future regimes, there is no clear connection to the underlying turbulence physics which is presumably determining the confinement. Thus, it is hard to determine whether model performance (particularly good performance) is due to a sufficient underlying understanding and encapsulation of the relevant physics, or is simply (un)fortuitous. Addressing these issues requires local turbulence-focused (rather than global transport and confinement) metrics.

IV. BUILDING ROBUST LOCAL TURBULENT TRANSPORT VALIDATION METRICS

In order to go beyond the global transport metrics discussed in Sec. III A, we must construct metrics that will allow us to systematically quantify and compare the experimental fidelity of different turbulence models at multiple radial locations in multiple tokamak discharges. Building upon the discussions of Secs. II and III, and drawing from experience gained by the community in performing gyrokinetic validation exercises over roughly the last decade, we can identify a set of criteria that these metrics should meet, beyond those outlined in Sec. II;

1. The metrics should utilize simple mathematical formulas to make the results as transparent as possible to non-experts.
2. The initial set of metrics should be easily extensible to incorporate comparisons of additional quantities.
3. Calculation of the metrics should be practical for use with nonlinear gyrokinetic simulations, which can individually require 1000 or more core-hours to obtain converged results.
4. The fundamental challenge for validating plasma microturbulence models—the interplay between stiff model responses and experimental uncertainties in equilibrium profiles and gradients—is explicitly incorporated into the metric design.

In order to address these issues, the plasma microturbulence community has begun to converge on a common approach of using local sensitivity plots for presenting verification[80,81] and validation[17,18] results. In this approach, one identifies a single input parameter (often the driving gradient of the dominant microinstability), and performs a discrete set of simulations in which this parameter is systematically varied about the experimental value, holding all other model inputs fixed. A typical example is shown in Fig. 5(a), where results from (effectively) three different models of ITG dominant transport are compared against an independent power balance flux calculation as a function of the normalized ion temperature gradient $R/L_{Ti} = -(R/T_i)dT_i/dx$, also known as the normalized inverse ion temperature gradient scale length. Here we have used the major radius R to normalize the ion temperature scale length since we are most often concerned with the curvature-driven ITG instability in tokamak plasmas. As R serves to parameterize the strength of the toroidal curvature and corresponding drift velocity, R/L_{Ti} is the dimensionless control parameter that appears in analytic calculations of the mode growth rate, rather than simply dT_i/dx or $1/L_{Ti}$. Other common normalization choices include the plasma minor radius a and density scale length L_n (particularly for the slab ITG instability). It is important to note that this discussion implicitly assumes a local model of the turbulence, in which the local turbulence properties (including cross-field fluxes) depend only upon the local gradients and other plasma parameters. A full discussion of local versus nonlocal turbulence models is beyond the scope of this paper; we focus on the local approach here since it is widely used in both turbulence and transport modeling, and note that the conceptual approach to validation described in this paper can be readily adapted to nonlocal turbulence and transport modeling.

This sensitivity plot approach was first published as a means of plasma microturbulence validation studies in a pair of seminal papers by Ross and co-workers,[17,18] and then utilized in a number of subsequent studies.[66,82–94] Many other validation studies have assessed the sensitivity of

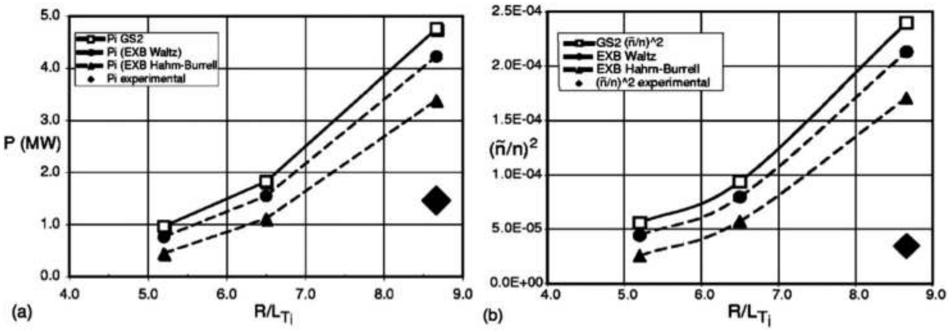

FIG. 5. Local sensitivity plots comparing predictions of (a) total energy flow $P_i = Q_i V'$ and (b) $|\tilde{n}_e/n_{e0}|^2$ from GS2 gyrokinetic simulations using different $E \times B$ shear suppression "quench rules" to experiment, as a function of normalized inverse gradient scale length R/L_{Ti}. Reprinted with permission from Phys. Plasmas **9**, 5031 (2002). Copyright 2002 AIP Publishing LLC.[18]

inputs but not explicitly presented the results in the form shown in Fig. 5. For the purposes of this paper, we note that this approach meets all of the desired criteria for a validation metric except for explicit inclusion of experimental and modeling uncertainties. In fact, there are three specific sets of uncertainties that must be quantified and incorporated into the metric:

1. The uncertainty in the power balance flux calculation (or more generally the measured system response quantity in the language of Oberkampf *et al.*[8–10,13]).
2. The uncertainty in the measurement of the local normalized driving gradient (the system input quantity).
3. And the uncertainty in the simulation predictions of the flux (the predicted system response quantity).

Quantifying the uncertainty in each of the quantities is a challenging process that requires careful consideration, and is discussed further below in Secs. IV A–IV C.

As will be discussed further in Sec. IV A, direct measurements of core tokamak cross-field fluxes are virtually nonexistent, and what are commonly termed the "experimental fluxes" are in fact the results of power balance calculations performed independently of the turbulence modeling. An important implication of this lack of direct flux measurements is that the comparison between power balance and turbulence model fluxes presented in Fig. 5(a) cannot rigorously be interpreted as a validation of the turbulence model, since it involves the comparison of two computational model outputs rather than a computational model to experiment. However, one can interpret Fig. 5(a) as a validation of a joint power balance and turbulence model calculation. In this view, the joint model calculation yields the prediction of a local inverse gradient scale length R/L_{Ti}^{sim} (or whichever other input quantity was varied) at which the separate power balance and turbulence models fluxes are equal to each other, which can be compared with the corresponding measured gradient scale length R/L_{Ti}^{expt}. To the extent one has confidence in the power balance analysis, such a comparison becomes primarily (but never entirely) an assessment of turbulence model fidelity. This interpretation of the local sensitivity plot approach is used to formulate a quantitative metric based upon the difference of R/L_{Ti}^{sim} and R/L_{Ti}^{expt} in Sec. IV E. In order to remove the dependence upon the power balance analysis, sensitivity plots comparing predictions for directly measured turbulence characteristics can also be made, such as is shown in Fig. 5(b). These calculations and their associated challenges are discussed in Sec. V. Finally, we note that a key advantage of the local approach and local simulations is that they allow one to decouple variations in the local value of dT_i/dx from variations in T_i itself, and thus one can isolate the impact of variations in the relevant control parameter R/L_{Ti} while holding T_i and its associated dimensionless quantities that appear in the gyrokinetic equations such as T_i/T_e fixed. Therefore, in the discussions that follow which focus upon the local approach, variations in dT_i/dx are implicitly assumed to be performed at fixed T_i, and we will treat discussions of variations or uncertainties in the local value of dT_i/dx as equivalent to those in R/L_{Ti} unless otherwise noted.

A. Quantifying uncertainties in power balance fluxes

The most common assessments in plasma microturbulence validation studies are comparisons of the predicted magnetic flux-surface averaged particle, energy, and momentum fluxes to what are often referred to as the experimental values. However, this label is incorrect, as in the core of MFE-relevant plasmas, such fluxes are never directly measured—there is no diagnostic capable of measuring these quantities that could survive at or access the multi-keV core plasma temperatures. Instead, the turbulence model predictions are compared with independent power balance calculations, which first calculate internal $S_{int}(x)$ and auxiliary source $S_{aux}(x)$ terms, and then calculate fluxes from the transport equations, e.g.,

$$Q_{PB}(x) = \frac{1}{V'}\int_0^x d^3x'\, V'\left(S_{int} + S_{aux} - \frac{dW}{dt}\right), \tag{4}$$

where $V(x)$ is the plasma volume enclosed within the flux surface labeled by x and $V'(x) = dV/dx$ is the surface area of that flux surface. The dW/dt terms on the right-hand side related to the temporal evolution of the kinetic quantities are often small relative to the internal and auxiliary source terms and so are frequently neglected in the calculation of Q_{PB}. Verification and validation of the various codes and models used in these power balance calculations are challenging exercises in their own right, and it is essential to always bear in mind that the lack of direct measurements of local fluxes is a significant constraint on the validation of these tools. However, through combinations of extensive verification exercises,[95] global cross-checks (e.g., comparisons of predicted and measured neutron production rates to constrain fast ion densities injected by neutral beams[96–98] or hard X-ray emissions associated with interactions between fast electrons and lower hybrid, electron cyclotron, and ion cyclotron waves[99–101]) and comparisons with both measured changes in equilibrium profiles[102–106] and core fluctuations,[107,108] quantitative confidence in these models to a fairly high level has been established for many operating conditions of interest. Nonetheless, power balance analyses are still subject to significant aleatory (i.e., statistical) and systematic uncertainties.

One can identify two main sources of uncertainties in power balance calculations. The first is the inherent statistical uncertainty due to uncertainties in the magnetic and kinetic equilibrium profiles input to the power balance analysis. For instance, both the ion and electron energy transport equations (as defined for each species by Eq. (2)) have an internal source term related to collisional inter-species energy transfer $S^{exch}_{e/i} = n_e(T_{i/e} - T_{e/i})/\tau_{ei}$, where τ_{ei} is the electron-ion collision time. Obviously, any uncertainties in the plasma density or temperatures will translate directly into an uncertainty in this exchange term, and its corresponding contributions to the energy fluxes. In dense, highly coupled plasmas (such as is expected for ITER or a future fusion reactor), this exchange term can be a dominant component in the total power balance analysis of the individual ion and electron energy channels. Therefore, if there is a large uncertainty in this term, it can be impossible to determine the relative amounts of power transported through the ion and electron channels with confidence, even though the total energy flow is well known since the ion and electron exchange terms exactly cancel when Eq. (2) is summed over all species. If this ratio of ion to electron energy fluxes cannot be determined with confidence, it is unlikely that comparisons with turbulence models (which make specific predictions of this ratio based upon the mixture of underlying instabilities) will be of significant validation utility.

In order to formally determine these power balance uncertainties, one must first construct probability distribution functions (PDFs) for the various magnetic and kinetic profiles, and then propagate these PDFs through the power balance model to yield PDFs of source and flux profiles. In practice, what is commonly done is to create an ensemble of equilibrium profiles, from which an ensemble of power balance calculations can be made, the statistics of which are used to calculate the (mean) source terms and their uncertainties (generally quantified as the standard deviation of the ensemble). Obviously the key to obtaining meaningful results through this method lies in appropriate construction of the profile ensembles. For most experimental MFE situations, there are two (non-exclusive) approaches to generating these ensembles. First, recognizing that the equilibrium profiles to be used are almost always parameterized fits (usually via various splines or polynomials) to experimental point measurements, one can transform the uncertainties in the measured data points into uncertainties in the fits. This propagation can be achieved in principle through the computational fitting routines themselves, or more commonly by a Monte Carlo approach in which ensembles of data points are generated by randomly varying each point in proportion to the quoted uncertainty, and then generating corresponding ensembles of fits. Measurement of different kinetic profiles requires use of different diagnostic techniques,[109,110] such as Thomson scattering, reflectometry, electron cyclotron emission (ECE), charge exchange recombination (CER), X-ray crystal spectroscopy (XCIS), and motional Stark effect polarimetry (MSE). Each such technique has its own challenges and uncertainties that must be understood and incorporated for this uncertainty quantification approach to be meaningful. An example of this approach can be seen in White *et al.*[111] in which a 100-element ensemble of profiles "sets" was propagated through the ONETWO power balance code[112] to generate a corresponding ensemble of power balance fluxes and thermal diffusivities.

In the second approach, the temporal variation of the measurements is used to generate the ensemble. The full experimental time-averaging window is broken up into a series of subwindows (whose length is often set by the sampling rate of the slowest relevant profile diagnostic), and profile fits to the measurements in each subwindow are generated to create an ensemble of time slices. An example of this approach is shown in Fig. 6, where a 11-element profile fit ensemble is generated by decomposing temperature profile point data from a typical DIII-D[181] L-mode discharge[113,114] collected over 220 ms into eleven 20 ms subwindows, and fitting the profile point data in each subwindow (Fig. 6(a)) via splines. The ONETWO code is then used to calculate a power balance analysis for each subwindow, and the statistics of this temporal ensemble is utilized to calculate the uncertainty in the fluxes (Fig. 6(b)). Here, a 95% confidence interval is calculated via

$$\sigma_{95} = t^*(0.025, N)\sqrt{\frac{\sigma^2_{R/L_{Ti}}}{N}}, \tag{5}$$

where t^* is Student's t-statistic, $\sigma_{R/L_{Ti}}$ is the standard deviation of the ensemble of R/L_{Ti} profiles derived from the fits to the experimental data, and N the number of elements in the ensemble ($N = 11$ here). For the case shown in Fig. 6, note that although there is quite small statistical uncertainty in the mean temperature profiles, the variations are still large enough that the power balance 95% confidence intervals correspond to an approximately 10% uncertainty in the energy fluxes at larger radii ($\rho_{tor} > 0.5$). One could also apply this approach to comparable time windows from multiple "repeat" discharges which hold equilibrium parameters and profiles constant across each discharge.

In addition to these inherent statistical uncertainties, one must also consider potential systematic uncertainties and errors in the power balance analysis. To illustrate the challenges in quantifying these uncertainties and errors more concretely, consider the various individual source and sink terms of the electron energy transport equation in the same DIII-D discharge, illustrated in Fig. 7. The net electron heating source (and thus energy flux) is composed of four terms: direct heating of the electrons by collisions with injected fast neutral deuterium beams, resistive Ohmic heating, collision energy exchange with the plasma thermal ion populations, and radiation. The calculation of these individual terms will only be as good as the individual models and assumptions used. In these calculations, propagation of the ensemble of profiles shown in Fig. 6 enables calculation of the

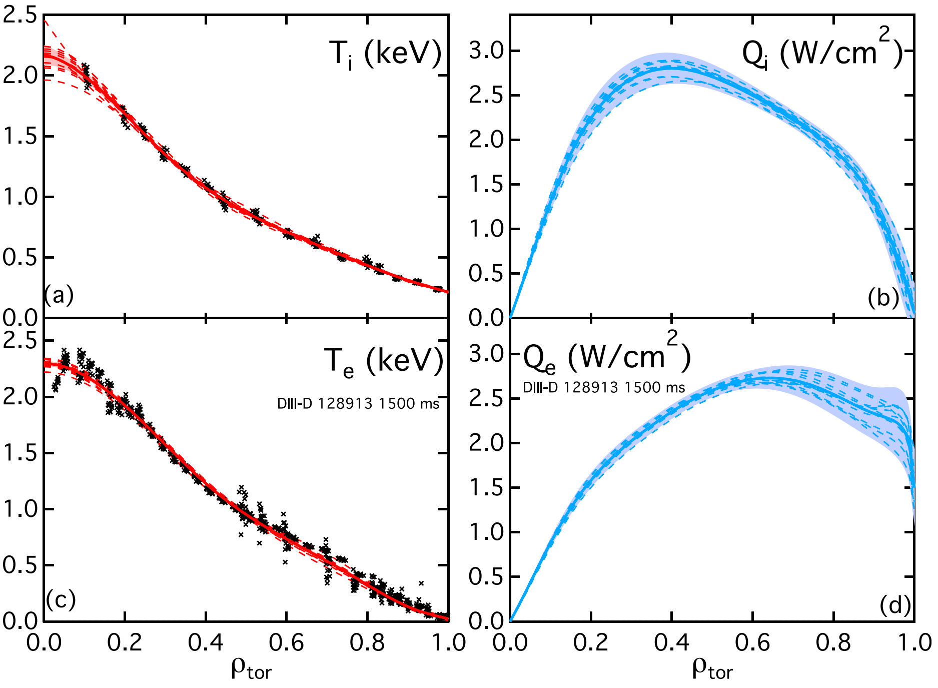

FIG. 6. Ensemble of fits to measured (×) (a) T_i and (c) T_e profiles in a DIII-D L-mode discharge, derived from decomposing a 220 ms collection window into eleven 20 ms subwindows. The fits are then propagated through the ONETWO power balance code to generate ensembles of (b) Q_i and (d) Q_e. Individual ensemble members are plotted as dashed lines (- - -), the ensemble mean as a solid line (—), and the 95% confidence interval for each ensemble mean is denoted by the shaded region.

corresponding statistical uncertainties of each term, denoted by the shaded bands. However, potential systematic errors in the calculations could arise from a variety of sources, such as:

1. Choice of fast beam ion slowing down and equilibration model.
2. Anomalous fast ion transport, most often due to Alfvén eigenmodes, which would broaden the beam heating profile.[115]
3. The availability and quality of measured, calibrated radiation profiles to be used as inputs.
4. Correct reconstruction of the current profile in the magnetic equilibrium calculation.
5. Choice of resistivity model and code used (i.e., Chang–Hinton analytic theory,[116] NCLASS,[117] NEO,[118] etc.).
6. Uncertainties or errors in the effective charge $Z_{eff} = \sum_i q_i^2 n_i / \sum_i q_i n_i$ (and resulting uncertainties or errors in τ_{ei}) which impact both the Ohmic heating calculation and collisional exchange term.
7. Systematic errors or biases in the profiles fits due to choice of fitting form or algorithm (discussed further in Sec. IV B).

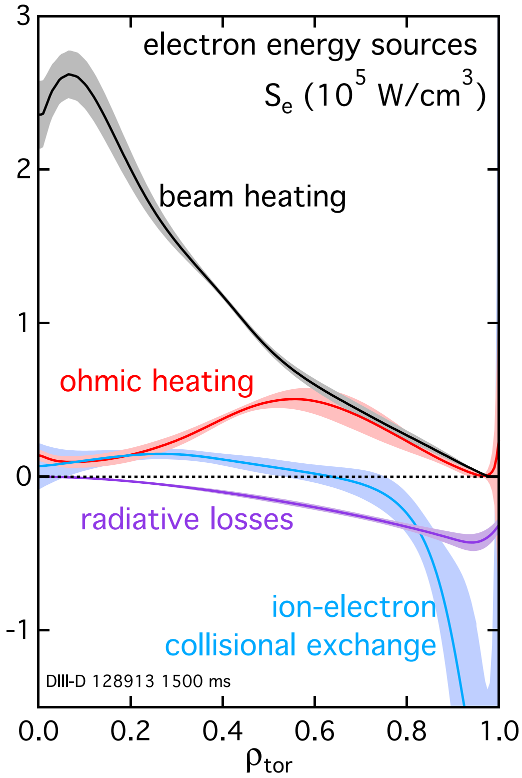

FIG. 7. Illustration of various components of total electron energy source term S_e and their uncertainties, for the ensemble of fits and corresponding power balance analysis shown in Fig. 6.

Similar issues arise in the calculation of other source terms of other profiles; particularly important are charge-exchange and prompt fast ion losses in the calculation of ion thermal and momentum sources, and edge neutral penetration and ionization in determining the total particle flux. As with most systematic errors, one can only try to minimize these errors through judicious choice of appropriate models and experimental design, and maintain awareness of them in interpreting results.

B. Quantifying the uncertainty in the local driving gradient

Equally important to quantifying the uncertainty in the power balance calculations is quantifying the uncertainty in the driving gradient (or other model input parameter under consideration) a particular local sensitivity analysis is focusing upon. Knowledge of this uncertainty is essential regardless of whether one is approaching the problem in terms of predicting fluxes and fluctuations, in which case the gradient uncertainty translates into the dominant turbulence model input uncertainty, or as a joint power balance-turbulence model prediction of the local gradients, whose sensitivity is obviously limited by the uncertainty in the measured local gradient. For most MFE experiments and conditions, the key challenge for quantifying this uncertainty arises from the fact that the equilibrium kinetic profiles used in turbulence and transport modeling are generally smoothly varying fits to assorted point measurements, as discussed in Sec. IV A. These fits are made not just for convenience of analysis, but also reflect an underlying assumption (embodied in the gyrokinetic formulation described above in Sec. III) that the equilibrium profiles vary slowly (relative to turbulent fluctuations) spatially as well as temporally.

Fig. 8 shows R/L_{Ti} calculated from both the ensemble of spline fits and direct finite difference of the CER point measurements shown in Fig. 6(a). The horizontal error bars on the point values of R/L_{Ti} equal the radial separation of the channels used in the finite differencing calculation. One can immediately see that while spline fit R/L_{Ti} profile captures the bulk trend of the point measurements, it by design does not capture the rapid radial variation and scatter of the point measurements. At the same time, one can see that the 95% confidence interval of the spline fits is both much smaller than the variations of the point measurements, and itself has non-negligible radial variations arising from the locations of the spline knots. Thus, neither approach is fully satisfactory for quantifying the uncertainty in either the local gradients or inverse gradient scale lengths. In practice, to the extent these uncertainties are considered in transport and modeling analysis, the fit ensemble approach is more commonly used, most frequently simply using the standard deviation of the ensemble to estimate the uncertainty.

Obviously, this issue is one that could benefit from further study and work toward a common, community-accepted approach to quantifying these crucial uncertainties. If the current ensemble statistics approach is not deemed sufficient, one might consider utilizing a uniform fractional uncertainty in the gradients based on the radially averaged ratio of the 95% confidence interval to the mean spline R/L_{Ti}

$$\delta_{LT_i}^{avg95} = \frac{1}{a}\int_0^a dr\, \frac{\sigma_{95}}{R/L_{Ti}^{fit}}, \qquad (6)$$

or a RMS fractional difference between the point measurements and spline R/L_{Ti} profile fits

$$\delta_{LT_i}^{RMS} = \sqrt{\frac{1}{a}\int_0^a dr \left(\frac{R/L_{Ti}^{data} - R/L_{Ti}^{fit}}{R/L_{Ti}^{fit}}\right)^2}. \qquad (7)$$

Both of these suggested measures are based upon the idea of using radial averaging to smooth out the fast variations of R/L_{Ti} and its associated confidence interval due to either the spline knot locations or inherent scatter of the point measurements, at the cost of sacrificing information about physically meaningful radial variations in the uncertainty of the local R/L_{Ti} value. The results of applying Eqs. (6) and (7) to the data shown in Fig. 8 yield values of $\delta_{LT_i}^{avg95} = 19\%$ and $\delta_{LT_i}^{RMS} = 43\%$; these results are also illustrated by additional shaded regions in Fig. 8. Alternatively, a more modest radial smoothing could be utilized, although this would require identification of some objective method for selecting the smoothing function and width. More generally, it would be desirable to increase utilization of more sophisticated fitting and uncertainty quantification techniques that have been developed in other communities. One promising technique in this vein is Gaussian process regression (GPR),[119] which takes a Bayesian approach to determining the probability distribution function of the profile and its gradient. Chilenski *et al.*[120] have recently applied this technique to modeling of impurity transport in the Alcator C-mod tokamak, and adaptation of the approach to other studies appears tractable. Certainly further study of GPR, and more detailed assessments of its potential benefits and costs relative to traditional MFE profile-fitting algorithms, is warranted. It should also be noted that for some profiles, multiple independent diagnostics are often combined (such as Thomson scattering and electron cyclotron emission for measurements of T_e profiles) to increase confidence in the final fit. In such cases, the use of integrated data assessment (IDA) techniques[121–124] can enable significantly improved estimates of profile and gradient uncertainties. Finally, in some experiments, it is possible to use small "jogs" or scans of the plasma through the diagnostic viewing locations to obtain more smoothly varying profile (and thus gradient) measurements;[125] similar results for diagnostics such as ECE can be obtained through small variations in toroidal field strength.[126]

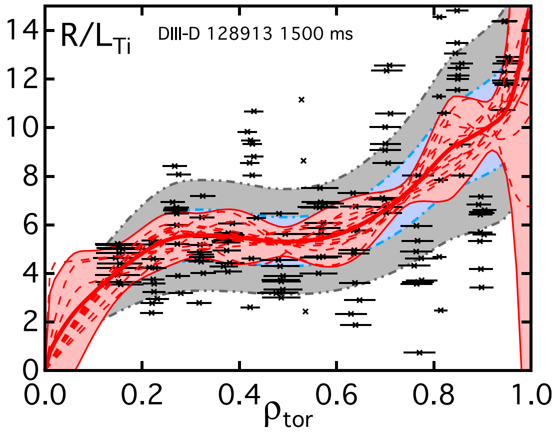

FIG. 8. Calculation of R/L_{Ti} from data and fits shown in Fig. 6(a). The calculation of R/L_{Ti} from direct finite differencing of the measurements is plotted as (×), with associated horizontal bars indicating the separation of CER channels differenced to obtain that point. Calculations of R/L_{Ti} from individual member profile fits are plotted as (red dashed line), and the ensemble mean as (red line). The shaded region bounded by solid lines denotes the 95% confidence interval on the ensemble mean R/L_{Ti}. Additional bounded shaded regions indicate the uncertainty intervals associated with $\delta_{LT_i}^{avg95} = 19\%$ (blue dashed line) and $\delta_{LT_i}^{RMS} = 43\%$ (gray dashed line).

C. Quantifying the uncertainty in the simulation predictions

The third group of uncertainties that we must quantify is the simulation output uncertainties. For nonlinear initial-value simulations, one source of such uncertainty arises from the time-averaging of the simulation results, which as described above is necessary for any meaningful V&V study. To see how one should address and quantify these particular uncertainties, consider the time trace of the magnetic flux-surface averaged ion energy flux Q_i from a nonlinear gyrokinetic simulation of a DIII-D discharge shown in Fig. 9(a). Q_i is output every $\Delta t = 1\,a/c_s$ (where $c_s = \sqrt{k_B T_e/m_i}$ is the ion sound speed), and the thick bar indicates the mean value of $Q_i = 1.26\,\mathrm{W/cm^2}$ averaged over the time window of $200\,a/c_s \le t \le 600\,a/c_s$. Note that the averaging does not begin until after the initial transients have damped away and the early linear physics (exponential growth) phase has ended. Although there is no easy formal rule for defining when this linear phase ends and the nonlinear phase of physical interest begins, the practical rule of thumb to use is that any quoted results should be insensitive to the choice of averaging window endpoints. More broadly, any nonlinear initial value simulations used in V&V studies should be run for sufficiently long times that the uncertainty associated with the time-averaging is subdominant relative to all of the other sources of input and output uncertainty, since it is (relatively speaking) one of the most straightforward to minimize. Toward this end, average values of Q_i for all possible choices of averaging window, parameterized in terms of starting time t_{start} and averaging window size t_{avg}, are shown in Fig. 9(b). So long as the initial transient linear phase ($t \lesssim 150\,a/c_s$) is not included, and a sufficiently long averaging window ($t_{avg} > 100\,a/c_s$) is utilized, the exact values of the time-averaged ion energy flux are insensitive to the choices made.

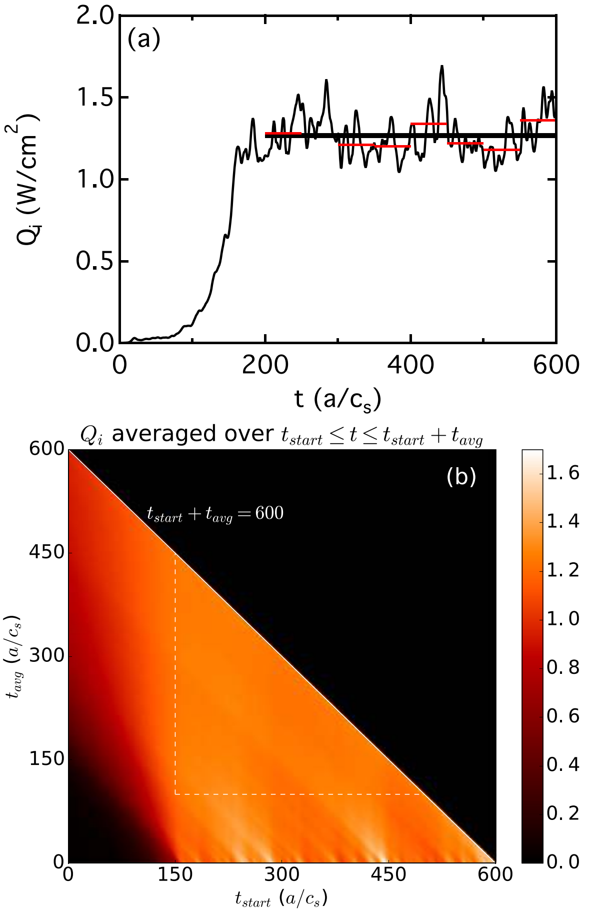

FIG. 9. (a) Time trace of Q_i from gyrokinetic simulation of a DIII-D discharge. The thick line (—) indicates the average value over the window $200\,a/c_s \le t \le 600\,a/c_s$, and the thin lines (red line) indicate mean values of sequential $50\,a/c_s$ subaveraging windows. Adapted from Phys. Plasmas **18**, 056113 (2011). Copyright 2011 AIP Publishing LLC. (b) Contour plot of mean Q_i values averaged over the window $t_{start} \le t \le t_{start} + t_{avg}$, for arbitrary values of t_{start} and t_{avg}. The average value is seen to be insensitive to the choice of specific averaging window when $t_{start} > 150\,a/c_s$ and $t_{avg} > 100\,a/c_s$, indicated by (– – –).

Although Fig. 9(b) suggests that the time-averaging uncertainty in the mean value of Q_i is "small" for appropriately chosen averaging windows, a more rigorous and impartial means of quantifying this uncertainty is needed. To date, there has not been a widely adopted common approach to this question within the MFE turbulence community. One possible approach is illustrated by the series of thin horizontal lines plotted in Fig. 9(a), which mirrors standard experimental signal processing techniques. In this approach, the full averaging window is first decomposed into a series of sequential subwindows, each of which is chosen to be long enough to average over fast variations, such that the mean values of Q_i calculated for each subwindow can be taken to form an ensemble of uncorrelated, independent estimates of the "true" mean value of Q_i. The uncertainty in this "true" mean value is then taken to be the standard deviation of the mean $\sigma_M = \sqrt{\sum_{j=1}^{N} (Q_{i,j} - \bar{Q}_i)^2 / N^2}$ (where $Q_{i,j}$ is the mean value of Q_i in the j-th subwindow, and $\bar{Q}_i$ ensemble mean value), calculated from the ensemble of subwindow means. For the particular Q_i timeseries shown in Fig. 9(a), applying this approach to a full averaging window of $200\,a/c_s \le t \le 600\,a/c_s$ with $50\,a/c_s$ subwindows yields a mean Q_i of $1.26\,\mathrm{W/cm^2}$ with $\sigma_M = 0.03\,\mathrm{W/cm^2}$. While there have been some efforts to formalize this approach,[127] there remains a significant opportunity and need to develop more rigorous algorithms for identifying appropriate time-averaging windows and their corresponding uncertainties. Future studies in this direction should look to draw upon the experiences and expertise of other communities which routinely utilize initial value nonlinear fluid turbulence simulations.

In applying this subwindowing technique, one should note that it assumes that the simulation saturates about a constant mean value that is large relative to amplitude of the fluctuations about that mean level. Such a result is often obtained for cases when the plasma is robustly unstable. However, for cases near marginal stability ($R/L_{Ti} \sim R/L_{Ti,crit}$), gyrokinetic simulation outputs are often observed to exhibit slow secular dynamics and significantly skewed fluctuations about mean values. For such cases, there is no commonly accepted methodology in the MFE community known to the author of calculating a well-justified mean value (in terms of choosing an averaging window) or associated uncertainty. Given the interest in improving predictions of ITER and reactor plasmas which are expected to lie in such a near-marginal regime over much of the plasma volume, it is clear that more work is needed to define appropriate analysis and uncertainty quantification methods for such cases.

Beyond these finite time-averaging uncertainties, there will be additional uncertainties in the model outputs due to uncertainties in any model inputs other than the control variable under consideration. In the context of the discussion thus far, these would be uncertainties in any model inputs other than R/L_{Ti}. In particular, uncertainties in other key instability-drive gradients such as the electron temperature (R/L_{Te}) and density (R/L_n), as well as the equilibrium $\vec{E} \times \vec{B}$ shearing rate γ_{ExB} (which in general suppresses the turbulence[128,129]) and key dimensionless parameters such as magnetic safety factor q, magnetic shear $s = (x/q)dq/dx$, and impurity fraction (sometimes expressed in terms of Z_{eff}) can yield significant uncertainties in model outputs. These uncertainties exist for both nonlinear initial value turbulence simulations as well as purely deterministic reduced turbulent transport models. Figure 10 shows values of Q_i predicted by GYRO simulations of an ITG-dominant DIII-D L-mode discharge for the nominal "base case" parameters, along with ±25% variations in R/L_{Ti}, R/L_{Te}, R/L_n, and γ_{ExB}, all of which are within experimental uncertainties. One can see that the model exhibits a nonlinear response to changes in these parameters as well as in R/L_{Ti} (i.e., a ±25% variation in any input does not necessarily yield a uniform $\pm\ X\%$ change in Q_i), which significantly complicates formulation of a simple statistical uncertainty estimate. Moreover, each such variation requires its own execution of the model, which quickly becomes prohibitively expensive on currently available computing platforms for nonlinear gyrokinetic simulations. In addition, "cross-terms" and couplings between different parameters are possible (due to, e.g., a mix of strong ITG and TEM instabilities being simultaneously present in the simulation) that may not be well captured by varying each input individually. Given these challenges, there is currently no widely used practical and robust model for estimating such uncertainties for gyrokinetic simulations, and to the author's knowledge no significant exploration of potential methods for use with computationally cheap reduced models has been undertaken. As such, it represents one of the ripest areas for more research and collaboration between the MFE community and broader computer science, applied math, and UQ communities. Whether next-generation exascale computing platforms can be profitably engaged to productively address this challenge remains to be seen.

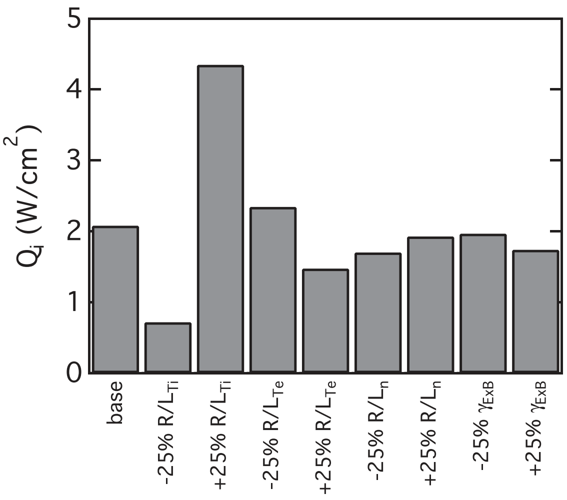

FIG. 10. Illustration of sensitivity of GYRO predictions of Q_i to ±25% changes in R/L_{Ti}, R_{LT_e}, R_{Ln}, and γ_{ExB}.

Finally, we note that one could also potentially classify numerical errors due to effects such as finite grid resolution, boundary conditions, domain size, source and sink terms, or the accuracy of the time-integration scheme as a third source of contributions to model output uncertainties. However, since these are more properly viewed as (hopefully known and minimized when feasible) systematic errors rather than uncertainties, we do not consider them further. In cases where sufficient computational resources are not available to minimize these errors, it is important to include them in the formulation of the validation metrics.

D. Example: Quantifying local gyrokinetic performance in DIII-D L-mode discharges

In order to illustrate the practical application of the local sensitivity plot analysis combined with uncertainty quantification to a "real-world" validation problem, we consider in this section an assessment of gyrokinetic predictions of ion and electron energy fluxes at different radii in a set of seven DIII-D neutral beam heated L-mode discharges. Experimental details of these discharges can be found in Refs. 111, 114, 130, and 131, and key global experimental parameters are summarized in Table III. Each of these discharges was performed as part of a coordinated transport model validation effort by the DIII-D experimental team, with particular attention paid to obtaining comprehensive, well-converged profile and fluctuation measurements in repeatable conditions.

For each case, the experimental data are averaged over at least 200 ms during which the plasma is slowly evolving, and uncertainties in experimental profiles, gradients, and power balance calculations estimated by ensemble calculations generated from subdivision of the full averaging window into 20 ms subwindows. Uncertainties in R/L_{Ti} are calculated using $\sigma^{RMS}_{LT_i}$, defined above in Eq. (7). The power balance analysis was performed with the ONETWO code,[112] using the Callen analytic model[132] to calculate neutral beam sources and TORAY-GA code[133] to calculate electron cyclotron heating sources. The gyrokinetic simulations are performed using the GYRO code,[54] and all simulations were averaged for over $250\,a/c_s$. However, we utilize a constant 10% fractional uncertainty for all GYRO flux predictions as a conservative estimate of the finite time-averaging statistical uncertainty. Since we have no tractable way of fully quantifying uncertainties in the GYRO outputs due to uncertainties in inputs other than R/L_{Ti}, we leave these uncertainties as unspecified "known unknowns."

All simulations were performed with resolutions and algorithms similar to those reported in Refs. 111, 114, 130, and 131, but using a common version of the GYRO source code.[134] Time integration was performed with a 4th-order Runge–Kutta scheme that treats fast parallel electron dynamics implicitly and other terms explicitly. The integration timestep h was less than or equal to $0.01\,a/c_s$ in all cases, such that estimates of the numerical integration error are less than 0.1%. Each simulation used a standard 128-point velocity space discretization (eight energy points, eight pitch angles, and two signs of parallel velocity), and physical simulation domains of approximately 100 ρ_s across in both the radial and binormal dimensions, where $\rho_s = c_s/\Omega_{ci}$, with

TABLE III. Global parameters for DIII-D transport model validation discharges. The plotting symbol column indicates the plotting symbol to be used for that discharge in Figs. 11–14 and 19–21.

Discharge number	Avg. window (ms)	B_T (T)	I_p (MA)	$\bar{n}_e$ (10^{19} m^{-3})	P_{NBI} (MW)	P_{ECH} (MW)	Plotting symbol	Reference
128913	1400–1600	2.05	1.05	2.3	2.6	0	● (filled circle)	114
136674	1300–1500	2.05	1.15	3.2	2.6	0	▲ (brown triangle)	130
136693	1300–1500	2.05	0.7	4.1	5.2	0	▼ (red inverted triangle)	130
138038	1400–1650	2.05	1.0	2.3	2.6	2.2	◀ (blue left pointing triangle)	111
138040	1400–1650	2.05	1.0	2.3	2.6	0	▶ (blue right pointing triangle)	111
142351	1400–1600	2.1	0.98	2.3	2.6	0	◆ (green rhombus)	131
142371	1800–2000	2.1	0.98	2.3	2.6	3.2	■ (green square)	131

additional 10 ρ_s wide buffer regions on at either end of the radial domain. The radial grid resolution was approximately 0.5 ρ_s for simulations at $\rho_{tor} \simeq 0.25$ and 0.5, and 0.3 ρ_s for those simulations at $\rho_{tor} \simeq 0.75$, and all cases use 16 toroidal modes with separation Δn chosen such that binormal wavenumbers span the range $0 \leq k_y\rho_s \lesssim 1$, where $k_y = nq/r_{min}$. Since these simulations consider only long-wavelength ρ_s-scale dynamics for which $k_y\rho_e = 60\,k_y\rho_s \ll 1$, the electrons are treated with a simpler drift-kinetic model (which assumes $k_\perp\rho_e = 0$) rather than a full gyrokinetic description. The simulations include finite perpendicular (but not parallel) magnetic fluctuations and two dynamic ion species (deuterium and carbon), and use a generalized Miller representation[135] to describe shaped geometry.

In Fig. 11(a), a comparison of the GYRO prediction of the ion energy flux Q_i with the ONETWO power balance calculation at $\rho_{tor} = 0.75$ in the most well-studied of the discharges considered (as seen in Refs. 113, 114, 131, and 136–139) is plotted as a function of R/L_{Ti} including all of the uncertainties described above. One can clearly see that the difference between the GYRO prediction of Q_i and corresponding ONETWO calculation at the nominal experimental gradient, or alternatively the difference of the predicted flux-matching gradient (i.e., the value of R/L_{Ti} for which $Q_i^{GYRO} = Q_i^{ONETWO}$) and the experimental gradient clearly lie outside the net model and experimental uncertainties. The question then arises as to whether this result is unique to this particular location and discharge, or robust across multiple radii and plasmas. To answer this question, we begin by recasting the local sensitivity plot shown in Fig. 11(a) into a comparison of fractional differences and uncertainties, shown in Fig. 11(b). Thus, instead of plotting the power balance and gyrokinetic flux predictions in W/cm^2 as a function of R/L_{Ti}, we plot the fraction difference of the ion energy fluxes $\Delta_{Q_i} = (Q_i^{GYRO} - Q_i^{ONETWO})/Q_i^{ONETWO}$ as a function of the fractional change in input value of $dT_i/dx = \nabla T_i$, $\Delta_{LT_i} = (\nabla T_i^{GYRO} - \nabla T_i^{expt})/\nabla T_i^{expt}$. The uncertainty in Δ_{LT_i} is simply the fractional uncertainty $\delta_{LT_i}^{RMS}$ in the experimental measurement of R/L_{Ti}, and the uncertainty in Δ_{Q_i} is calculated as $\sqrt{(\sigma_i^{ONETWO}/Q_i^{ONETWO})^2 + (\sigma_i^{GYRO}/Q_i^{GYRO})^2}$ since the propagation of profile uncertainties through ONETWO represented by σ_i^{ONETWO} is completely independent of the time-averaging uncertainty σ_i^{GYRO}. Recasting Fig. 11(a) in this form is analogous to the final step in the validation metric progression of Oberkampf *et al.* shown in Fig. 1(f), and also yields a simple physical interpretation. If the curve of Δ_{Q_i} vs. Δ_{LT_i} passes within its uncertainties through the origin (highlighted by the bold star symbol in Fig. 11(b)), then one can legitimately claim that the model prediction is consistent with the experimental observations, given the known uncertainties. Conversely, if the distance between the Δ_{Q_i} curve and the origin is always larger than the uncertainties, one has demonstrated a statistically significant difference between the model prediction and experimental measurements.

Recasting the local sensitivity plot in terms of fractional differences has a second practical benefit, which is that it

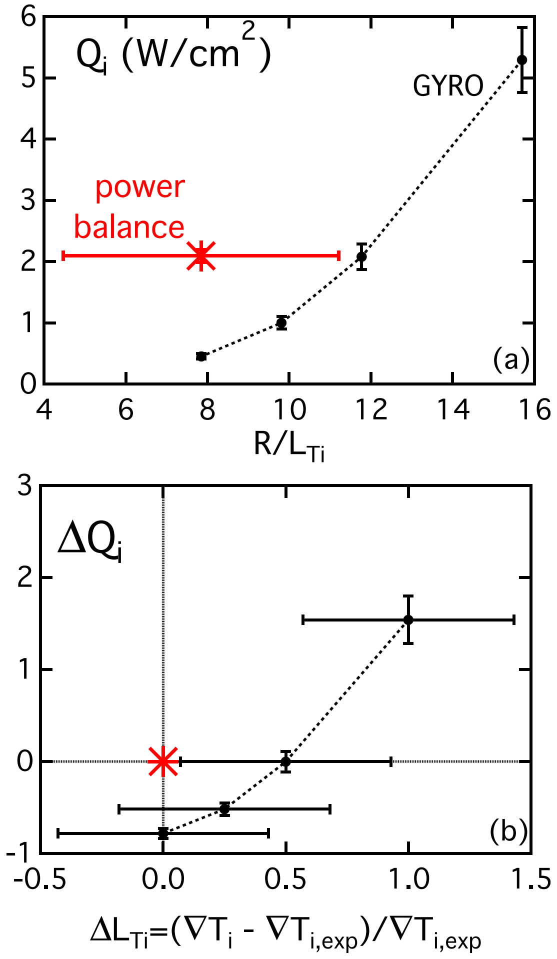

FIG. 11. Scaling of GYRO predictions (●) of Q_i vs R/L_{Ti} at $\rho_{tor} = 0.75$ in DIII-D discharge 128913, in both (a) physical units and (b) fractional differences. The broken lines indicate where ΔQ_i and $\Delta L_{Ti} = 0$ and the joint experimental-power balance results are plotted as (red asterisk).

facilitates direct comparisons of multiple conditions (e.g., different radii and/or discharges) that have different experimental and simulation values. This utility is shown in Fig. 12, which plots the fractional ion and electron energy flux differences for all seven discharges listed in Table III at $\rho_{tor} = 0.25$, 0.5, and 0.75. From these plots, we can immediately draw two conclusions. First, we see that for all seven discharges there is a systematic underprediction of Q_i and Q_e at $\rho_{tor} = 0.75$ (or equivalently an overprediction of the flux-matching gradient) which is larger than the experimental and model uncertainties. Second, for all seven discharges, the model predictions are significantly more consistent with the power balance ion and electron energy fluxes at both $\rho_{tor} = 0.25$ and 0.5. We can therefore conclude that this particular model (ion-scale microturbulence predicted with the GYRO code) cannot simultaneously match the experimental gradients and power balance fluxes (within uncertainties) in NBI-heated DIII-D L-mode discharges at $\rho_{tor} = 0.75$, but can (at least in some cases) at $\rho_{tor} = 0.25$ and 0.5. To make these conclusions more quantitative requires formulation of an explicit validation metric, which is the subject of Sec. IV E.

There are other possible choices for normalization of the simulation inputs and outputs beyond the experimental or power balance calculations, and the most useful choice will be case-dependent. For example, while the choice of normalizing quantities to the power balance and experimental levels works well here for Q_i and R/L_{Ti}, such a choice is not appropriate or possible when the experimental level is very small relative to the expected range of simulation inputs or outputs. A typical example here would be predictions of momentum transport in a case for which there was no meaningful auxiliary torque source $\mathcal{T}_{inj} = \sum_j \mathcal{T}_j$. In this case, a power balance analysis would predict $\Pi_{PB} \propto \int dV \mathcal{T}_{inj} \simeq 0$, and experimentally one often observes small rotation gradients in these plasmas, particularly relative to their uncertainty. In such cases, one might choose to normalize quantities relative to theoretical scalings (e.g., gyroBohm scaling $\Pi_{gB} = n_i T_i v_{ti} \rho^{*2}$) or some combination of experimental and model uncertainties.

E. Using flux-matching gradients to construct validation metrics

While the normalized local sensitivity plots shown in Fig. 12 provide a useful means of illustrating a model's fidelity for a single experimental condition, they rapidly become cluttered when multiple conditions are plotted. In particular, while one can clearly see in Fig. 12(c) that none of the simulations at $\rho_{tor} = 0.75$ simultaneously match the fluxes and R/L_{Ti}, there is enough scatter in the results at $\rho_{tor} = 0.25$ (Fig. 12(a)) and 0.5 (Fig. 12(b)) that it is not easy to determine how robust the model (dis)agreements are at those radii. Moreover, in order to better quantify how a model's performance varies with radius, or as a function of global parameters such as the auxiliary heating mix, plasma current, density, *etc*. it is desirable to condense the local sensitivity plot to a more compact and quantitative measure of model fidelity. Since the goal of the turbulent transport models discussed here is the prediction of the equilibrium kinetic profiles and gradients, the natural reduction of the local sensitivity plot is the difference between the measured experimental gradient and predicted flux-matching gradient, i.e., the gradient for which the model prediction of the associated flux matches the power balance result. In the spirit of utilizing normalizations that allow comparisons across multiple experimental conditions, we define a flux-matching fractional gradient error metric

$$E_{LT_i} = \Delta_{LT_i}|_{\Delta_{Q_i}=0} \tag{8}$$

with an associated uncertainty

$$\sigma_{E_{LT_i}} = \sqrt{\sigma^2_{\Delta_{LT_i}} + \sigma^2_{\Delta_{Q_i}}}|_{\Delta_{Q_i}=0}. \tag{9}$$

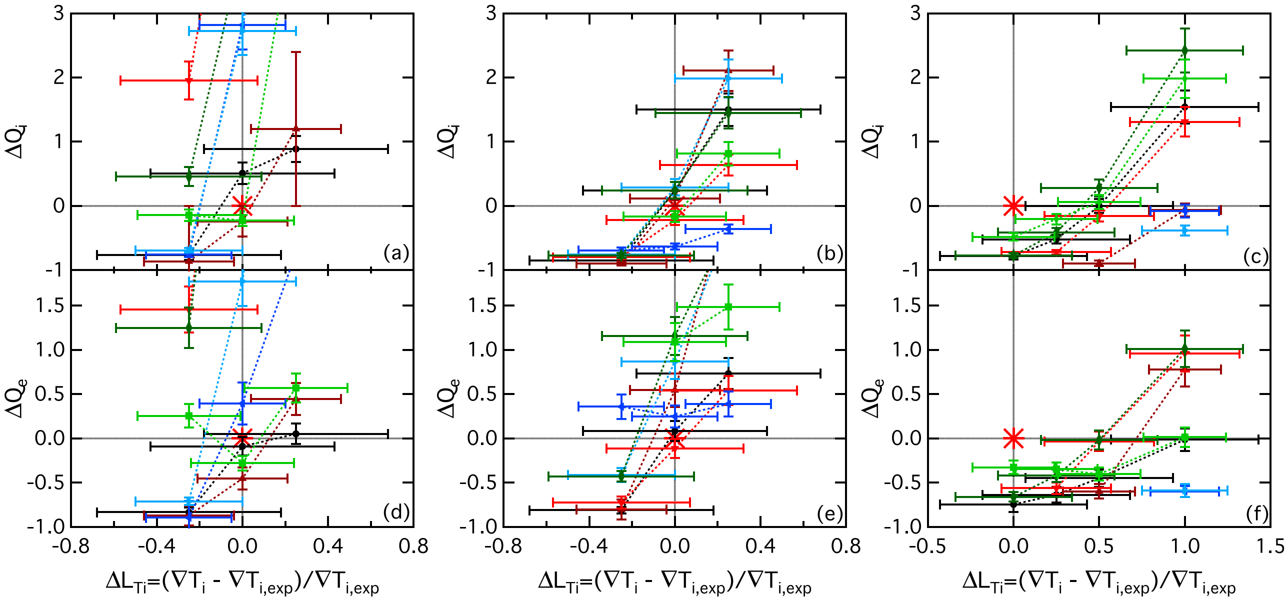

FIG. 12. Plots of (a)–(c) ΔQ_i and (d)–(f) ΔQ_e vs. ΔL_{Ti} at (a) and (d) $\rho_{tor} = 0.25$, (b) and (e) 0.5, and (c) and (f) 0.75 for all seven discharges listed in Table III. The broken lines indicate where ΔQ_i and $\Delta L_{Ti} = 0$ and the origin is indicated by (red asterisk).

Although the choice of requiring a match between the turbulent and power balance ion energy fluxes is natural (since Q_i is the relevant flux in the ion temperature transport equation (Eq. (2)), and we are considering ion-scale simulations of ITG turbulence), it is only one several possible choices. Equally viable would be the electron energy flux, total energy flux, particle or momentum flux, or even fluctuation amplitudes or other characteristics (discussed further in Sec. V), depending upon the specific goals of the validation exercise.

Once the fractional gradient error is defined, one can define error metrics for any other comparison quantity such as Q_e based upon the difference between the power balance or experimental measurements and the model predictions of that quantity at the flux matching gradient. Thus we would define a fractional error metric for Q_e as

$$E_{Q_e} = \Delta_{Q_e}|_{\Delta_{Q_i}=0}. \tag{10}$$

For this metric, we define the error only as

$$\sigma_{E_{Q_e}} = \sigma_{\Delta_{Q_e}|_{\Delta_{Q_i}=0}}, \tag{11}$$

i.e., the uncertainty in Δ_{LT_i} is not included.

The relationship between E_{LT_i} and E_{Q_e} is illustrated in Fig. 13, and the results of calculating these metrics for the simulation data shown in Fig. 12 are plotted in Fig. 14. In addition to the individual error metrics, the uncertainty-weighted average over all seven discharges (i.e., using weights $W_i = 1/\sigma_i^2$) is also plotted vs. radius for both E_{LT_i} and E_{Q_e} in Fig. 14. Confirming the visual impressions of model fidelity from Fig. 12, we see that at both $\rho_{tor} = 0.25$ and 0.5 both the mean E_{LT_i} and almost all individual cases have values of E_{LT_i} smaller than the associated uncertainty, while every case at $\rho_{tor} = 0.75$ predicts a positive value of E_{LT_i} (i.e., overpredicts the local value of R/L_{Ti}) that is larger than the associated uncertainty.

Even more interesting are the results for E_{Q_e}, which exhibit a fair amount of scatter. If we consider only the ensemble-averaged value at each radius, we see that at $\rho_{tor} = 0.25$ and 0.75, the mean value of E_{Q_e} is slightly negative but with an absolute value larger than the associated uncertainty, corresponding to a statically robust residual underprediction of Q_e even when Q_i has been matched. On the other hand, Q_e is on average overpredicted at $\rho_{tor} = 0.5$. However, in all cases we see clear individual outliers with significantly different values than the ensemble mean. The question here naturally arises as to whether these discrepancies lie within the full range of variability and undetermined uncertainty in model predictions related to uncertainties in inputs other than R/L_{Ti}. The natural quantity to focus on for Q_e would be R/L_{Te}, in order to determine how much uncertainty there is in the level of Q_e driven by the ITG modes themselves, as well as any other ion-scale modes such as TEMs which might be present but subdominant. To address this question, one would extend the validation methodology discussed so far to calculation of a local sensitivity map in which both R/L_{Ti} and R/L_{Te} are varied, from which a flux-matching fractional gradient error vector $\vec{E}_z = (E_{LT_i}, E_{LT_e})$ could be calculated by determining the *simultaneous* values

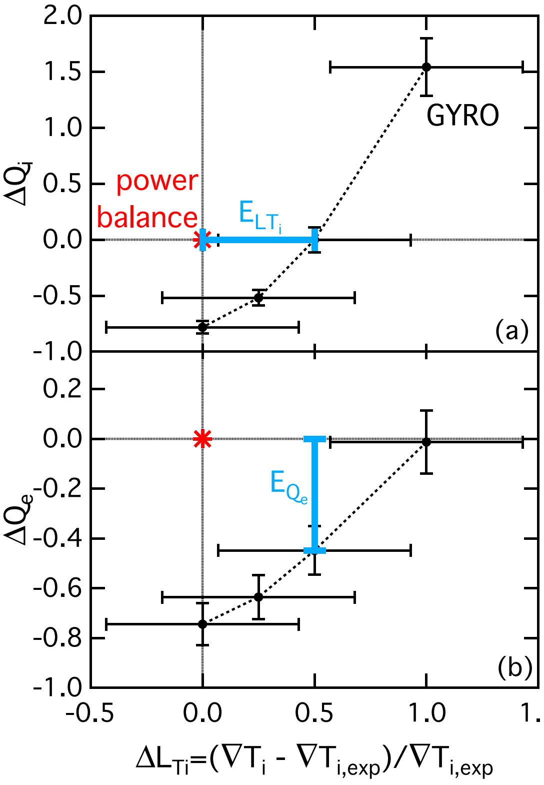

FIG. 13. (a) ΔQ_i and (b) ΔQ_e vs. ΔL_{Ti} for $\rho_{tor} = 0.75$ in discharge 128913, illustrating the connection between E_{LT_i}, E_{Q_e}, ΔQ_i, ΔQ_e, and ΔL_{Ti}. The broken lines indicate where $\Delta Q_{i,e}$ and $\Delta L_{Ti} = 0$ and the origin is indicated by (red asterisk).

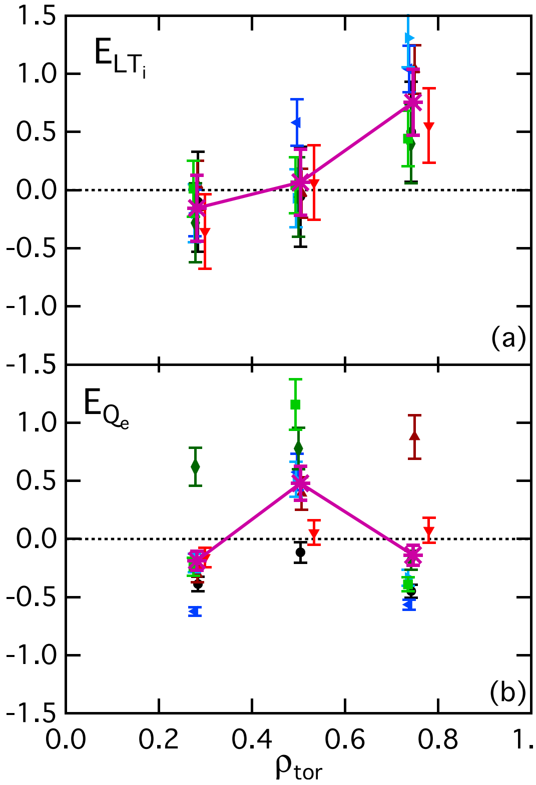

FIG. 14. (a) E_{LT_i} and (b) E_{Q_e} vs ρ_{tor} for all seven discharges listed in Table III. The uncertainty-weighted ensemble mean values at each radius are plotted as pink asterisk.

of R/L_{Ti} and R/L_{Te} which when input into the microturbulence model yield predictions of Q_i and Q_e that *simultaneously* match the power balance Q_i and Q_e results. While the computational resources needed to perform such an analysis using long-wavelength gyrokinetic simulations (to say, nothing of multiscale simulations which incorporate electron-scale ETG dynamics that likely contribute to Q_e in many cases[65–68,140–146]) over many conditions or discharges remain prohibitive for current-day computing platforms, such approaches will likely be feasible on next-generation exascale platforms. Moreover, such an approach, or even further generalizations to include matching of particle and momentum fluxes via additional variations of density and rotation gradients, is readily feasible now for most reduced turbulent transport models with fairly modest computing resources, and should be pursued further.

F. Alternative metric formulations

While the structure of the E_{LT_i} and E_{Q_e} metrics proposed above have a clear physical interpretation and are consistent with some recommendations in the literature,[10,15] other choices are possible and have been pursued. In particular, normalizing the model–experiment differences in terms of uncertainties rather than mean experimental values is an equally viable choice, e.g.,

$$E^{alt}_{LT_i} = \frac{\left(\nabla T_i^{model} - \nabla T_i^{expt}\right)}{\sigma_{\nabla T_i^{expt}}}\bigg|_{\Delta Q_i=0}. \qquad (12)$$

The primary advantages of this approach are a further compactification of the metric, from $(value) \pm (uncertainty)$ to simply $(value)$, which facilitates an clear means of assessing whether the model and experiment are consistent within uncertainties (i.e., is the metric greater or smaller than some order-unity threshold value[147,148]). On the other hand, by using such a formulation one cannot discriminate between cases that achieve small metric values ("good agreement") due to small differences between the model and experimental predictions or large uncertainties in the experimental and/or simulation data.

More complex metrics have also been proposed in and utilized in a variety of studies. For instance, drawing upon a widely used climate validation metric,[149] Terry *et al.*[14] assess the fidelity of ITG correlation lengths predicted by different models using the data published in Rhodes *et al.*[150] These approaches quantify agreement between model and experiment using the correlation coefficient between the predicted (x) and measured (y) values of a particular observable obtained at N distinct points

$$R = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_x \sigma_y} \qquad (13)$$

and RMS deviation

$$E = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\frac{x_i - \bar{x}}{\bar{x}} - \frac{y_i - \bar{y}}{\bar{y}}\right]^2}. \qquad (14)$$

Here $\bar{x}$ and $\bar{y}$ are the mean values of x and y, and σ_x and σ_y their standard deviations. As another example, Ricci *et al.*[147,148] define an uncertainty-normalized distance metric

$$d = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{(x_i - y_i)^2}{\Delta x_i^2 + \Delta y_i^2}} \qquad (15)$$

for validations studies of the GBS Braginskii fluid code[151] using data from the TORPEX experiment.[152] A bounded error metric of the form

$$R = \frac{1 + \tanh[(d - d_0)/\lambda]}{2} \qquad (16)$$

for each observable is then defined, such the $R = 0$ denotes perfect agreement and 1 complete disagreement. The quantities d_0 and λ are free parameters chosen to quantify the threshold level for agreement (d_0) and sharpness of transition from agreement to disagreement (λ), with $d_0 \simeq 1.4$ corresponding to the case of the distance between simulation and experiment being comparable with their uncertainties. For their studies, they found that the conclusions drawn were fairly insensitive to the specific choice of d_0 and λ, so long as they were in the ranges of $1 \le d_0 \le 2$ and $0.1 \le \lambda \le 1$. Whether this property would hold for other studies remains to be seen. One particular advantage of this bounded metric formulation is that it lends itself well to incorporation into composite metrics,[14] which are discussed in Sec. VI.

V. VALIDATION METRICS FOR PREDICTIONS OF TURBULENT FLUCTUATIONS

In order to fully validate a turbulent transport model, one must assess the fidelity of the predicted turbulent fluctuations themselves as well as the cross-field fluxes. Comparisons of predicted and measured fluctuations serve two specific, complementary purposes. First and foremost, such comparisons provide a more stringent means of testing our understanding of the fundamental underlying physics of plasma transport, which is needed for confident extrapolation to future regimes. More specifically, by assessing the ability of a given model to accurately predict a wide variety of measured fluctuation characteristics (such as amplitudes, spectra, correlation lengths, etc.) and their scalings with plasma parameters in conjunction with the cross-field fluxes, we can determine whether or not the models are predicting (in)correct flux-gradient relationships because they have (in)correct models of the underlying turbulence dynamics. In doing so, these fluctuation comparisons also provide a means of addressing the myriad potential systematic uncertainties of the power balance analyses discussed in Sec. IV A.

A. Using synthetic diagnostics to enable quantitative code-experiment comparisons

In order to carry out these comparisons of predicted and measured plasma fluctuation characteristics, one must invariably use synthetic diagnostics as part of the comparison. Synthetic diagnostics are computational algorithms used to transform the quantities output by a simulation into experimentally measured quantities to enable meaningful quantitative comparisons.[153] While there is a wide array of diagnostic techniques used to measure plasma fluctuations, in virtually every case the quantity measured by the diagnostic differs in some way from the "native" variables of the various turbulence models. For instance, a synthetic Langmuir probe[109] would translate the electron density, temperature, and plasma potential fluctuations predicted by a simulation into the measured ion saturation current and floating potential fluctuations, while a synthetic gas puff imaging diagnostic[154] would calculate predicted light emission fluctuations based upon the underlying density and temperature fluctuations and plasma ion and neutral species. In many cases, the synthetic diagnostic algorithm itself can be a sophisticated computational model requiring its own verification and validation, possibly even including the use dedicated experimental devices.

For the family of discharges considered in Sec. IV D, both beam emission spectroscopy (BES)[155] and correlation electron cyclotron emission (CECE) radiometry[113] measurements at multiple locations were obtained as part of the experiments. Obtaining measurements from these diagnostics was prioritized in the design of the experiments because they provide spatially localized measurements of the long-wavelength ρ_i-scale fluctuations described by the GYRO simulations utilized in Sec. IV D. Both diagnostics work by integrating plasma radiation emitted from small but finite spatial volumes, the intensity of which is then related back to instantaneous local density or electron temperature values. To model each diagnostic, a point-spread function (PSF) is convolved with the relevant simulation outputs to account for the finite integration volume of the diagnostic, as described in Ref. 114. Denoting the diagnostic-specific structure of the PSF as $\psi_{PSF}(R,Z)$, synthetic electron density or temperature time traces are generated as

$$\delta X_{syn} = \frac{\int\int dR\, dZ\, \psi_{PSF}\, \delta X_{GYRO}}{\int\int dR\, dZ\, \psi_{PSF}}, \qquad (17)$$

where δX can refer to either the normalized electron density fluctuation $\delta n_e = \tilde{n}_e/n_{e0}$ or temperature fluctuation $\delta T_e = \tilde{T}_e/T_{e0}$. Typical PSFs for both BES and CECE are visualized in Fig. 15. For the BES system, calculation of the PSF is made by a separate code developed by the BES diagnostic group;[156] for the CECE system the PSF is a Gaussian in both R and Z with widths provided by the CECE diagnostic group. In both cases the toroidal extent of the experimental integration volume is significantly smaller than the typical toroidal correlation lengths of the turbulence, which is approximated in the synthetic diagnostic as a perfect toroidal localization (i.e., ψ_{PSF} does not depend upon toroidal angle). In Fig. 16, typical lab-frame time traces and frequency spectra of the synthetic fluctuations are plotted against those from corresponding unfiltered signals which are simply recorded at the nominal diagnostic channel location (i.e., $\delta X_{unfiltered}(R,Z,t) = \delta X_{GYRO}(R,Z,t)$). One can immediately see that as would be expected for PSFs with spatial dimensions comparable to the turbulent eddy size, there is significant attenuation of fluctuation power at all frequencies in the synthetic spectra, relative to the unfiltered case. However, the exact frequency dependence of this attenuation depends upon the specific shape of the PSF. For instance, the synthetic CECE spectra shown in Fig. 16(d) is more heavily attenuated at higher frequencies. This particular dependence arises from the poloidally extended but radially narrow PSF of the CECE diagnostic (Fig. 15(b)), which preferentially attenuates higher poloidal wavenumbers and thus higher frequencies due to the strong Doppler shift driven by the finite rotation of the plasma. In contrast, the BES PSF is more widely elongated radially, which leads to non-negligible attenuation of higher radial wavenumbers for all poloidal wavenumbers and frequencies (Fig. 16(b)).

B. Frequency-spectra based fluctuation analysis and comparisons

Once the synthetic time series δX_{syn} has been generated, they can be analyzed in the same way as the experimental time series, with any differences primarily arising due to elimination of experimental noise sources (such as background photon shot noise and electronic noise in the case of BES and CECE) that are not present in the synthetic data. Note that both diagnostics are comprised of multiple channels measuring fluctuations at distinct spatial locations, such that the synthetic diagnostic algorithm generates a set of time series corresponding to the different spatial channels of the diagnostic being modeled. For turbulence modeling, the most common analysis and quantities of comparison are

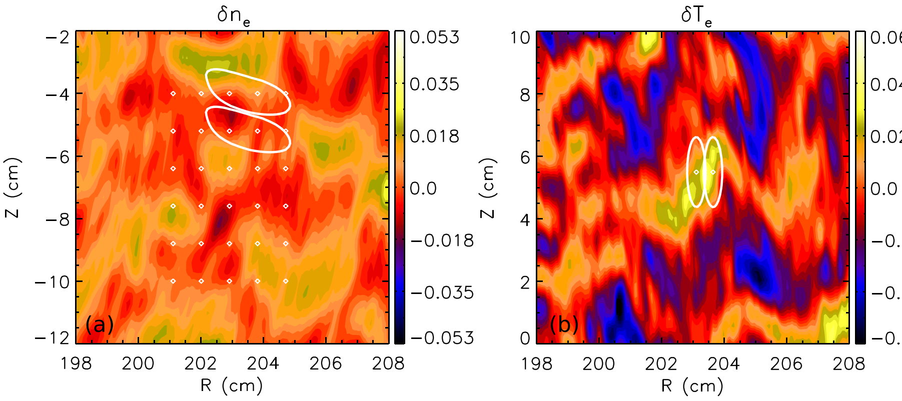

FIG. 15. Visualization of 50% contours of (a) BES and (b) CECE PSF functions overlaid on corresponding δn_e and δT_e fluctuations from a GYRO simulation. Nominal viewing locations of individual BES and CECE channels are shown as (◇). Reprinted with permission from J. Phys: Conf. Ser. **125**, 012043 (2008). Copyright 2008 IOP Publishing.[157]

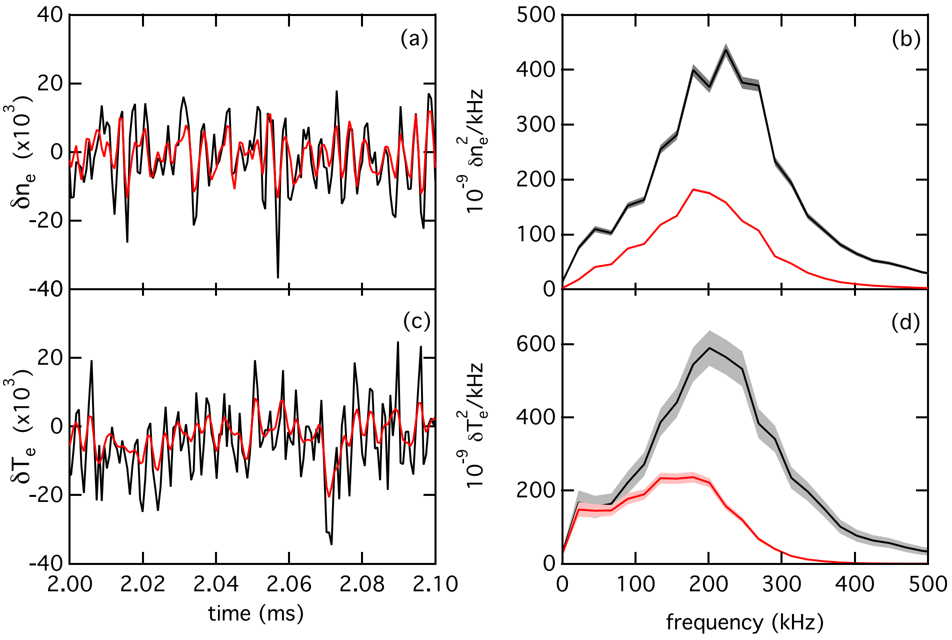

FIG. 16. Time-traces of unfiltered (—) and synthetic (red line) (a) BES and (c) CECE signals. Corresponding (b) BES and (d) CECE frequency spectra illustrate the attenuation of each synthetic signal relative to the unfiltered case, with a frequency dependence arising from the interplay of PSF shape and Doppler shifts. (b) and (d) adapted with permission from Phys. Plasmas **16**, 052301 (2009). Copyright 2009 AIP Publishing LLC.[114]

correlation functions and power spectra, which are often integrated over some frequency or wavenumber range to yield a fluctuation intensity or RMS amplitude. In many cases cross-spectra from neighboring diagnostic channels are utilized rather than single-channel auto-spectra when analyzing the experimental data to suppress uncorrelated noise sources; the synthetic analysis should follow the experimental analyses in these cases. The calculation of auto- and cross-spectra, along with related quantities such as coherence and cross-phase, as well as more complex measures such as bispectra, are standard signal processing techniques which are well-described in a variety of textbooks (e.g., Refs. 158 and 159), to which the interested reader is referred to for further details. For the purposes of this discussion, we note only that there are standard procedures for estimating the uncertainties in these spectral quantities based upon the length and number of averaging windows used, which should be calculated and included in any validation analysis. However, as with the time-averaging uncertainties of the simulation flux predictions described in Sec. IV C, the dominant uncertainty in predictions of fluctuation quantities is likely to arise from uncertainties in the model inputs, rather than the time-averaging uncertainty.

Typical synthetic and experimental BES spectra are shown in Fig. 17, from which quantitative comparisons can be formulated in terms of different moments of the power spectrum density $S(f)$. The most commonly used comparison is the total fluctuation power contained within some frequency band, often expressed in terms of RMS fluctuation amplitudes δX_{RMS}, defined as

$$\delta X_{RMS}^2 = \int_{f_{min}}^{f_{max}} df\, S(f). \tag{18}$$

Integration over a finite frequency band, rather than the entire spectrum, is often used to remove components of the measured or calculated spectrum that are not related to the turbulence. In the example shown in Fig. 17, the experimental BES measurements below 40 kHz are dominated by fluctuations in the neutral beam voltage rather than the microturbulence of interest, and so would be excluded from a comparison with the simulated fluctuations. In cases where rotation or the presence of large coherent modes leads to a clear relationship between frequency and mean wavenumber, comparisons of different frequency bands can be used as an effective proxy for comparisons of different wavenumbers.

Although comparisons of fluctuation amplitudes are the clearest zeroth-order test of consistency between simulated and measured turbulence properties, determining that the fundamental characteristics of the turbulence are being captured accurately by a given model also requires comparisons

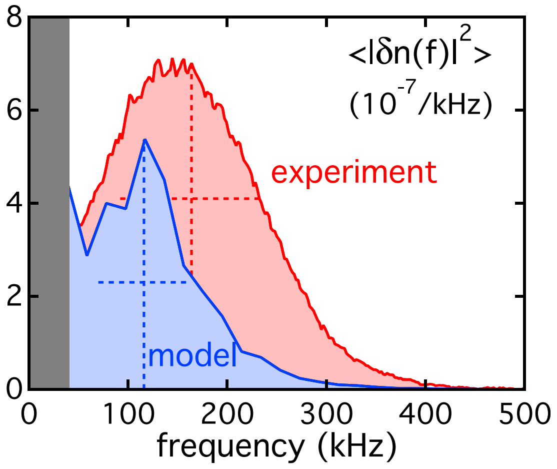

FIG. 17. Comparison of synthetic and measured BES spectra at $\rho_{tor} = 0.75$ in DIII-D discharge 128913, using data from a GYRO simulation with $\Delta Q_i = 0$. Dashed lines (- - -) indicate mean frequency $\bar{f}$ (Eq. (20)) and spectral width W_f (Eq. (21)) calculated for each spectrum over the range 40 kHz $\leq f \leq$ 400 kHz. The dark gray shaded region spanning 0 to 40 kHz indicates the portion of the measured spectrum dominated by non-turbulence fluctuations, and thus should not be included in comparisons with the turbulence model. Adapted with permission from Phys. Plasmas **16**, 052301 (2009). Copyright 2009 AIP Publishing LLC.[114]

of the spatiotemporal structures of the turbulence, as well as of the couplings and cross-phases between different fields. The most obvious means of comparison would be the frequency-integrated difference between predicted and measured spectra

$$R_{\delta X} = \int_{f_{min}}^{f_{max}} df\, [S_{sim}(f) - S_{expt}(f)] \tag{19}$$

perhaps normalized to the experimental value of δX^2_{RMS}. However, this formulation is sensitive to uncertainties in rotation and resulting Doppler shifts (described further below), and moreover loses much of the information about the shape of the spectrum that is of interest. Therefore, additional comparisons of quantities such as higher moments of the frequency spectra than simply the integrated power (or differences in $S(f)$) are more desirable. For instance, one could compare not just the RMS fluctuation levels but also the mean frequency

$$\bar{f} = \frac{1}{\delta X^2_{RMS}} \int_{f_{min}}^{f_{max}} df\, f S(f) \tag{20}$$

and spectral width

$$W_f = \frac{1}{\delta X^2_{RMS}} \int_{f_{min}}^{f_{max}} df\, \left(f - \bar{f}\right)^2 S(f) \tag{21}$$

of the measured and simulated spectra, as illustrated in Fig. 17. Alternatively, one could transform the frequency spectra into correlation functions via

$$C(\tau) = \frac{1}{\delta X^2_{RMS}} \mathrm{Re}\left\{ \int_{f_{min}}^{f_{max}} df\, S(f) e^{-2\pi i f \tau} \right\} \tag{22}$$

and formulate comparisons in terms of different quantities derived from $C(\tau)$. The most common of these is the decorrelation time, generally obtained by fitting an exponential or Gaussian function to the envelope of $C(\tau)$, which can be calculated using the Hilbert transform. Other measures such as spectral indices α, obtained by fitting portions of the spectrum as $S(f) \propto f^{-\alpha}$ can be calculated and compared as well.

In addition to the single-field (e.g., electron density or temperature) comparisons, comparisons of correlations between different fluctuation fields have proved to be of significant utility when possible. Physically, the turbulent cross-field fluxes depend upon the correlation of density, velocity, and temperature fluctuations with radial velocity fluctuations, and as such can be represented spectrally in terms of coherency $\gamma(f)$, cross-phase $\Theta(f)$, and the autospectra of the individual fluctuation fields as follows (using the normalized electron particle flux $\Gamma = \langle \delta n \delta v_r \rangle$ as a specific example):

$$\begin{aligned} \Gamma(f) &= \mathrm{Re}\{\langle \delta n^*(f) \delta v_r(f) \rangle\}, \\ &= \gamma(f) \langle |\delta n(f)|^2 \rangle^{1/2} \langle |\delta v_r(f)|^2 \rangle^{1/2} \cos\Theta(f), \end{aligned} \tag{23}$$

$$\gamma(f) = \frac{|\langle \delta n^*(f) \delta v_r(f) \rangle|}{\sqrt{\langle |\delta n(f)|^2 \rangle \langle |\delta v_r(f)|^2 \rangle}}, \tag{24}$$

$$\Theta(f) = \tan^{-1}\left(\frac{Im\{\langle \delta n^*(f) \delta v_r(f) \rangle\}}{Re\{\langle \delta n^*(f) \delta v_r(f) \rangle\}} \right). \tag{25}$$

Therefore, testing the specific predictions of the coherency and cross-phase against measurements is greatly desirable for assessing whether the specific nonlinear dynamics of the turbulence that determine the cross-field fluxes are being accurately captured in the simulation. Unfortunately, measurements of correlations between fluctuations fields and radial velocity fluctuations (either the generally dominant $\vec{E} \times \vec{B}$ component or the magnetic flutter component $v_r^{\delta B} = v_{||} \delta B_r$) are very rarely available on closed flux surfaces in high-power tokamaks due to the lack of diagnostic techniques available for measuring either component of v_r; they can sometimes be obtained in the plasma edge and scrape-off layer regions via Langmuir probes. However, more broadly gyrokinetic theory predicts unique phase relationships between any two fields for each instability of interest (ITG, TEM, ETG, *etc.*). Therefore, comparisons of the measured and predicted cross-phase of any two fluctuations (such as n_e and T_e) provide at minimum a test of whether the mix of underlying instabilities predicted by the simulation is consistent with observations. Such comparisons have been documented in some validation studies,[111,138] which found that not only did nonlinear gyrokinetic simulations quantitatively predict the cross-phases and their variations with heating power, but also these results were close to the predictions of cross-phases from linear stability calculations. These results provide support for the quasilinear transport modeling approach described above in Sec. III, which describes the transport in terms of a variety of small-amplitude fluctuations that retain many of the linear dispersion and phasing properties.

One practical challenge for many of these frequency spectra-based comparisons is that they can be highly sensitive to uncertainties in the toroidal rotation of the plasma, which translates into uncertainties in the Doppler shift that often dominates the lab-frame spectra. For example, the spectra shown in Fig. 17 are taken from the discharge documented in Refs. 113 and 114, which has an approximately 10% uncertainty in the toroidal rotation velocity V_{tor}, which in this discharge dominates the local $\vec{E} \times \vec{B}$ velocity that determines the Doppler shift between the plasma and lab reference frames. The lab-frame frequency can be expressed as

$$f_{lab} = f_{plasma} + \frac{\vec{k} \cdot \vec{V}_{ExB}}{2\pi} = f_{plasma} + f_{Doppler}, \tag{26}$$

where f_{plasma} is the plasma-frame frequency of the fluctuation in question. Considering simply the $k_y \rho_s = 0.3$ fluctuation (where the wavenumber spectrum peaks in the simulation), and estimating f_{plasma} as the linear mode frequency f_{lin} calculated from a linear gyrokinetic simulation, which is consistent with comparisons of linear and nonlinear calculations shown in Refs. 114 and 138, we find $f_{lin} = 12.7\,\mathrm{kHz}$ and $f_{Doppler} = 191\,\mathrm{kHz}$. Thus even a 10% uncertainty in V_{ExB} yields a uncertainty in $f_{Doppler}$ comparable with f_{lin}. Given this sensitivity of the predicted spectra to the Doppler shift, significant values of which are common for high-power tokamak plasma conditions, the uncertainties in $f_{Doppler}$ must be carefully considered before strong conclusions can be derived from comparisons of these frequency-based quantities with experiment.

C. Fluctuation Comparisons Based on Spatial Correlation Properties

Beyond the frequency-based model–experiment comparisons described above, comparisons which utilize the local spatial correlation and wavenumber properties of the turbulence can provide extremely useful tests of the predicted nonlinear dynamics of the system. At a high level, this point can be understood by noting that the gyrokinetic-Maxwell equations are more naturally expressed in terms of couplings between different spatial wavenumbers than frequencies, and so wavenumber-based comparisons can more directly connect to the underlying theory. More specifically, although one can often draw connections between the poloidal wavenumber and lab-frame frequency via the Doppler shift and linear dispersion properties of the turbulence in question, much of the important nonlinear dynamics involves couplings of different radial wavenumbers which in general cannot be easily mapped back to frequency space. In particular, one of the primary saturation mechanisms for ion-scale turbulence is now known to be nonlinear energy transfer from unstable to stable modes mediated by radially sheared axisymmetric $\vec{E} \times \vec{B}$ flows in the plasma. These flows can be both part of the equilibrium plasma (as a bulk rotation of the plasma), or nonlinearly generated by the turbulence itself, in which case they are often referred to as "zonal flows" reflecting their axisymmetric character. A full review of the details of these flows, their generation, and back-reaction on the turbulence is beyond the scope of this paper; the interested reader is refereed to the extensive literature on the topic for more information (see, e.g., Ref. 160). The key point for this discussion is that in wavenumber space, one can express this shearing process in the nonlinear term of the gyrokinetic equation as (using a simple Cartesian representation for clarity)

$$\frac{\partial \tilde{f}(k_x,k_y,k_z)}{\partial t} = -\sum_{k'_x} i k_y V^{SF}_{ExB}(k'_x,0,0)\,\tilde{f}(k_x-k'_x,k_y,k_z)+\ldots . \tag{27}$$

Thus, the radially sheared axisymmetric shear flow V^{SF}_{ExB} transfers energy between fluctuations with different values of radial wavenumber k_x but same poloidal (k_y) and toroidal (k_z) wavenumbers. Through this and other nonlinear processes, energy is transferred from linearly unstable modes (generally with small k_x) to stable modes at a variety of different wavenumbers,[161–163] saturating the turbulence and generally resulting in a wavenumber spectrum broad in both k_x and k_y. Therefore, if we want to test whether a given model is accurately capturing the nonlinear dynamics of the turbulence at an even deeper level than the frequency-based comparisons above allow, we should first examine the fidelity of the model in capturing this wavenumber spectrum.

By analogy to the frequency-based comparisons, these wavenumber comparisons can be formulated in terms of comparing mean wavenumbers and spectral widths in each spatial dimension. Alternatively, the wavenumber spectrum can be Fourier-transformed into a correlation function, and comparisons formulated in terms of mean wavenumber and correlation length, obtained analogously to the correlation time via a fit to the envelope of the correlation function. An example of such a comparison is shown in Fig. 18, taken from Shafer *et al.*[164] These results illustrate the strong impact the BES PSF can have on the results, which is not surprising since the spatial extent of the PSF is comparable with the size of the turbulent eddies, as shown in Fig. 15. In Fig. 18 the wavenumber spectra obtained from the unfiltered gyrokinetic results are compared against measured wavenumber spectrum from which the k-space representation of the PSF has been deconvolved, as well as comparisons of the synthetic spectrum to the unfiltered measured spectrum. Their differences can be quantified in terms of wavenumber peaks and widths, as shown in Table IV. Most notable is the clear overprediction of mean k_r at $\rho_{tor} = 0.75$, corresponding to the simulation eddies being more "titled" in configuration space than what is observed.

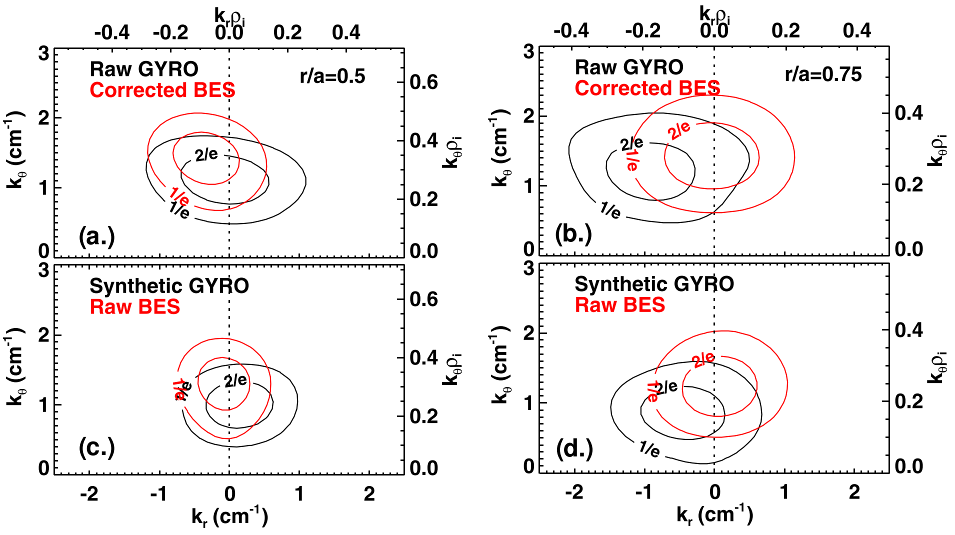

FIG. 18. Comparison of GYRO-predicted (—) and BES-measured (red line) density fluctuation wavenumber spectra at (a) and (c) $r/a = \rho_{tor} = 0.5$ and (b) and (d) $r/a = 0.75$. Raw (unfiltered) GYRO spectra are compared with "corrected" BES spectra from which the k-space structure of the PSF has been deconvolved in (a) and (b), while the synthetic GYRO spectra are compared with raw BES measurements in (b) and (d). Reprinted with permission from Phys. Plasmas **19**, 032504 (2012). Copyright 2012 AIP Publishing LLC.[164]

TABLE IV. Radial and poloidal wavenumber peaks and widths calculated for the spectra shown in Figs. 18(a) and 18(b). Adapted with permission from Phys. Plasmas **19**, 032504 (2012). Copyright 2012 AIP Publishing LLC.[164]

	k_r width (cm^{-1})	k_θ width (cm^{-1})	k_r peak (cm^{-1})	k_θ peak (cm^{-1})
Raw BES ($r/a = 0.5$)	1.3	1.4	–0.1	1.3
Syn. GYRO ($r/a = 0.5$)	1.6	1.2	0.2	1.0
Raw BES ($r/a = 0.75$)	2.0	1.5	–0.5	1.3
Syn. GYRO ($r/a = 0.75$)	2.2	1.5	0.1	0.9

D. Fluctuation sensitivity plots and validation metrics

Once a particular fluctuation characteristic has been chosen and quantified for comparison, it is straightforward to generate local sensitivity plots and error metrics for this quantity analogous to those for the turbulent fluxes discussed in Sec. IV. In all of the discharges considered, BES and CECE data are available at $\rho_{tor} = 0.5$ and 0.75. Using these data, the same fractional difference approach as described Sec. IV E is applied to the RMS normalized electron density and temperature fluctuation amplitudes to generate the results shown in Fig. 19. Similar plots for other quantities discussed above could also be readily generated, but the use of RMS fluctuation amplitude comparisons is sufficient for the illustrative goals of this paper. Comparing the results shown in Fig. 19 with Fig. 12, one can draw the qualitative conclusion that the predicted fluctuation levels exhibit similar levels of agreement with the measured levels as did the predicted turbulent fluxes and power balance calculations, namely, broad consistency between simulation and experiment at $\rho_{tor} = 0.5$, and systematic underprediction of the experimental observations at $\rho_{tor} = 0.75$.

Quantitative fractional error metrics for the fluctuations can be defined analogously to the fractional error metric for Q_e (Eq. (10)) E_{Q_e} as

$$E_{\delta n_e} = \Delta_{\delta n_e}|_{\Delta_{Q_i}=0}, \tag{28}$$

$$E_{\delta T_e} = \Delta_{\delta T_e}|_{\Delta_{Q_i}=0}, \tag{29}$$

with errors defined equivalently to $\sigma_{E_{Q_e}}$ (Eq. (11)). These error metrics are plotted in Fig. 20, along with their uncertainty-weighted average values. Significant scatter is seen in the results for both fields and both radii. Focusing on the ensemble-averaged values, we see in Fig. 20(a) that the predicted flux-matching density fluctuations at both radii match the measured values within the uncertainty, albeit just barely with, e.g., $E_{\delta n_e}$ only slightly smaller than $\sigma_{E_{\delta n_e}}$. This result supports the straightforward expectation from the underlying gyrokinetic model of a close correlation between the amplitude of the ρ_i-scale density fluctuations measured by BES and the ion temperature and velocity fluctuations at those scales that set Q_i. Therefore, by matching the power balance and turbulent values of Q_i, one would expect the predicted ρ_i-scale fluctuations to match the measured values. On the other hand, and somewhat surprisingly, we see that (on average) the predicted ρ_i-scale T_e fluctuations match experimental levels at $\rho_{tor} = 0.5$, but are larger than experiment levels at $\rho_{tor} = 0.75$ when the simulations match the power balance Q_i values, even though the power balance Q_e is overpredicted by the gyrokinetic model at $\rho_{tor} = 0.5$ and (modestly) underpredicted at $\rho_{tor} = 0.75$ (as shown in Fig. 14). Assuming that the electron thermal transport is also dominated by ρ_i-scale fluctuations, one would generally expect the electron energy flux and temperature fluctuation error metrics E_{Q_e} and $E_{\delta T_e}$ to exhibit similar trends with radius.

How to resolve these findings with simple theoretical expectations remains an open question. One possibility would be to investigate whether a more sophisticated synthetic CECE diagnostic, such as the one presented in Görler *et al.*,[138] which utilizes the perpendicular electron temperature fluctuation $\tilde{T}_{e\perp} = \int d^3v(m_e v_\perp^2/2 - 1)\tilde{f}_e$ that more

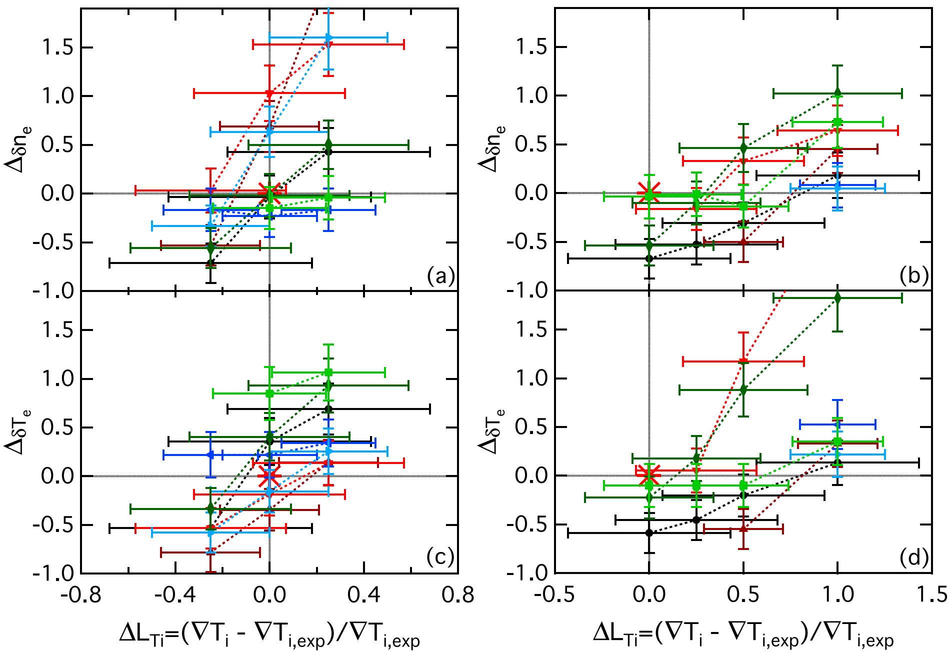

FIG. 19. Local sensitivity plots for (a) and (b) $\Delta_{\delta n_e}$ and (c) and (d) $\Delta_{\delta T_e}$ at (a) and (c) $\rho_{tor} = 0.5$ and (b) and (d) 0.75 calculated from same simulations as utilized in Fig. 12. The broken lines indicate where $\Delta_{\delta n_e}$, $\Delta_{\delta T_e}$, and $\Delta L_{Ti} = 0$ and the origin is indicated by (red asterisk).

closely corresponds to the actual CECE measurement than the total fluctuation $\tilde{T}_e = \int d^3v(m_e v^2/2 - 1)\tilde{f}_e$ used in the GYRO analysis, changes these findings significantly. Similarly, one could investigate more sophisticated synthetic BES algorithms that incorporate the impact of ion density fluctuations on the atomic physics of the collisional radiation processes used to calculate the predicted emission levels, or the neglect of parallel localization in both synthetic diagnostics employed here. And as with the discussion of the findings for E_{Q_e} in Sec. IV E, a second natural avenue to pursue (if tractable) would be to transition from varying only R/L_{Ti} to match Q_i to simultaneous variations of R/L_{Ti} and R/L_{Te} to simultaneously match Q_i and Q_e. Calculating $E_{\delta n_e}$ and $E_{\delta T_e}$ for these simulations and contrasting them to the results presented in Fig. 14 and 20 would provide insight into whether it is meaningful to consider the turbulence in these discharges as dominated by a single instability with a single strong control parameter, or whether they in fact must be considered a tightly coupled mix of different instabilities (e.g., ITG and TEM) with multiple, equally important drives in determining the transport and turbulence in each channel and field. A third, even more computationally expensive avenue of approach would be to investigate the impact of multiscale simulations which self-consistently incorporate ρ_e-scale ETG fluctuations into these ρ_i-scale simulations.[65–68,140–146] Other research avenues are possible as well. However, for the goals of this paper, it should be clear how utilizing a variety of local validation metrics for multiple predicted quantities can be employed to test model fidelity and our physical understanding at a level not possible by earlier global metrics.

VI. USING COMPOSITE METRICS FOR ASSESSING OVERALL MODEL FIDELITY

While the various individual metrics described above provide extremely useful and detailed insight into how well a specific aspect of turbulence dynamics and transport is captured by a model, it is also desirable to formulate composite metrics which integrate the information contained in the individual metrics into assessments overall model fidelity. These composite metrics should be used to complement the insights gained from the individual metrics, and reports of validation activities should include tables and figures documenting both the single metric results as well as composite metric results. In general, these composite metrics will take the form of weighted sums

$$M = \sum_i W_i R_i \qquad (30)$$

where W_i is the weight of the i-th comparison included and R_i the value representing the level of model–experiment agreement found for that comparison. As discussed below, R_i will generally be a function of the various individual metrics such as E_{LT_i}, but need not be exactly equal to them or even a linear function of them. Beyond formulating R_i, a second challenge lies in deciding how to weight different individual metrics, e.g., should more weight be given to tests of fluctuations than fluxes, or the comparisons with smallest uncertainties?

In formulating R_i, the first challenge is to ensure that individual metrics are combined in such as way as to avoid cancellations which would yield overly favorable assessments of total model fidelity. Thus, direct linear sums of signed metrics such as $E_{comp} = E_{LT_i} + E_{Q_e}$ should not be used, to avoid a case where combining, e.g., $E_{LT_i} \simeq 1$ and $E_{Q_e} \simeq -1$ yields a composite metric value $E_{comp} \simeq 0$, which would suggest very close model–experiment agreement based on two individual metrics indicating significant model–experiment disagreement. Terry *et al.* recommend utilizing normalized goodness ratings B_i for R_i, where B_i is bounded and varies from 0 (no agreement) to 1 (perfect agreement). Such a bounding can be achieved through a number of different means, the tanh function being a common one (as seen in formulation of the R in Eq. (16) by Ricci *et al.*[147,148]). For the local flux-matching metrics defined in this paper, using E_{LT_i} as an example, a simple approach would be to define

$$B_{T_i} = 1 - \tanh(|E_{LT_i}|), \qquad (31)$$

or by analogy to a suggestion from Greenwald[15]

$$B_{T_i} = 1 - \tanh(|E_{LT_i}| + |\sigma_{E_{LT_i}}|), \qquad (32)$$

if we desired to incorporate the uncertainty into the goodness rating. Alternatively one could use

$$B_{T_i} = 1 - \tanh(|E_{LT_i}/\sigma_{E_{LT_i}}|^\alpha), \qquad (33)$$

where $\alpha = 2$ would be a natural choice by analogy to chi-squared statistics. A transition parameter λ analogous to that of Eq. (16) could also be included, although one would need to carefully assess its impact on the interpretation of the results.

Terry *et al.* recommend using several criteria in choosing how to weight the goodness measure associated with each metric. The most important of these is what they term the primacy hierarchy. The primacy hierarchy is a way of ranking different quantities based upon the number of experimental measurements it integrates. For instance, in edge turbulence studies using Langmuir probes, at the lowest (most "fundamental") level of the primacy hierarchy would be measurements of equilibrium electrostatic potential and electron density and temperature profiles, as well as their fluctuations. At the next level of the primacy hierarchy would be equilibrium profile gradients (derived from the measured profiles one level below) as well as auto- and cross-correlations of fluctuation measurements, including cross-field fluxes based calculated using correlations of density or temperature and electric field fluctuations (derived from finite differencing the measured potential fluctuations). And at the highest level of the primacy hierarchy would be calculations of particle and thermal diffusivities that combined the fluxes with equilibrium profile gradients. The higher a quantity is in the primacy hierarchy, the lower a weight it is given. In addition to weighting by the level in the primacy hierarchy, Terry *et al.* also recommend including weighting factors accounting for the sensitivity of the

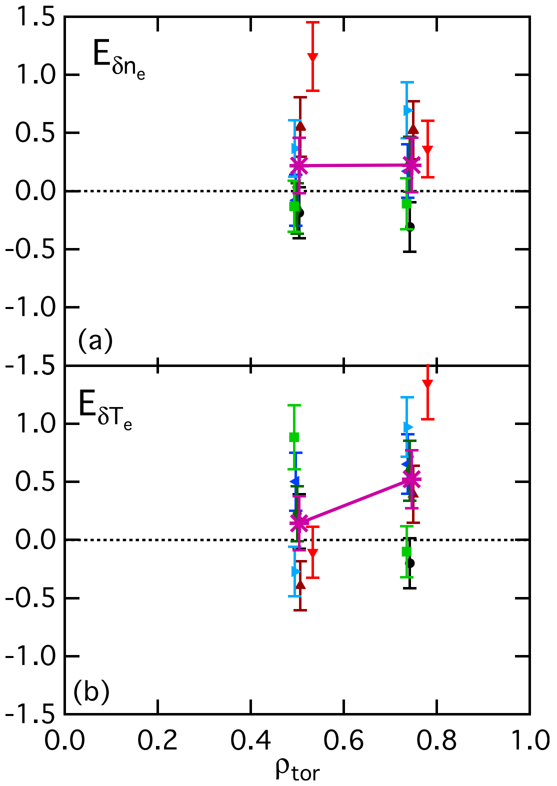

FIG. 20. Dependence of (a) $E_{\delta n_e}$ and (b) $E_{\delta T_e}$ on ρ_{tor} for all seven discharges listed in Table III. The uncertainty-weighted ensemble mean values are plotted as (pink asterisk).

comparison and whether it repeats similar measures (to encourage testing of many different quantities, rather than many tests of effectively the same quantity in slightly different conditions). An application of this methodology can be found in the studies of Ricci *et al.*,[147,148,165] where composite metrics of the form

$$\chi = \frac{\sum_i R_i H_i S_i}{\sum_i H_i S_i} \tag{34}$$

are used. Here R_i is the measure of agreement for a single comparison as defined in Eq. (16), H_i is a weighting for place in the primacy hierarchy, and S_i the sensitivity weight. The hierarchy weight H_i is defined as $H_i = 1/h_i$, where h_i is the level of the comparison in the primacy hierarchy. An interesting refinement of the initial primacy hierarchy proposal used in Ref. 148 is the definition of separate hierarchies for the experimental data, simulation data, and comparisons, depending on the number of assumptions or integrations required for each case. The sensitivity weight S_i is defined as

$$S_i = \exp\left(-\frac{\sum_j \Delta e_{ij} + \sum_j \Delta s_{ij}}{\sum_j |e_{ij}| + \sum_j |s_{ij}|}\right), \tag{35}$$

where $e_{i,j}$ and $s_{i,j}$ are the j-th instance of the experimental and simulation values, respectively, of the i-th comparison quantity; $\Delta e_{i,j}$ and $\Delta s_{i,j}$ are their associated uncertainties. Through the use of such a weighting, one is able to account for experimental simulation uncertainties in a composite metric while still using formulations R_i that do not explicitly reference these uncertainties.

Following these works, the values of several different composite metrics calculated for the DIII-D modeling results presented in Secs. IV E (Fig. 14) and V D (Fig. 20) are given in Table V and plotted in Fig. 21. At each radius, a single composite metric value is calculated from the four individual metrics $E_i = \{E_{LT_i}, E_{Q_e}, E_{\delta n_e}, E_{\delta T_e}\}$ calculated and all seven discharges considered. Using the definition of $B_{i,j} = 1 - \tanh(|E_{i,j}|)$ from Eq. (31), three composite metrics of increasing complexity are formulated

$$M_0 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{7} B_{i,j}}{7N}, \tag{36}$$

$$M_1 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{7} H_i B_{i,j}}{7\sum_{i=1}^{N} H_i}, \tag{37}$$

$$M_2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{7} H_i S_{ij} B_{i,j}}{\sum_{i=1}^{N}\sum_{j=1}^{7} H_i S_{ij}}. \tag{38}$$

The quantity H_i is a primacy hierarchy weight; we use a value of $H_i = 1$ for $E_{\delta n_e}$ and $E_{\delta T_e}$, and $H_i = 0.5$ for E_{LT_i} and E_{Q_e}. Since each $\sigma_{E_{ij}}$ represents the uncertainty in a fractional error metric E_{ij}, they are all treated equivalently in a simple sensitivity weighting $S_{ij} = \exp(-\sigma_{E_{ij}})$. At $\rho_{tor} = 0.5$ and 0.75, $N = 4$ is used, corresponding to the four individual metrics in E_i discussed in Secs. IV and V. However, since there are no fluctuation measurements available at $\rho_{tor} = 0.25$, we use $N = 2$ there, corresponding to $E_i = \{E_{LT_i}, E_{Q_e}\}$. Alternatively one could set either B_{ij} or S_{ij} equal to zero for $E_{\delta n_e}$ and $E_{\delta T_e}$ at this location, reflecting this absence relative to the other radii. Examining the results, one can see that each formulation gives a very similar answer, namely, the highest values at $\rho_{tor} = 0.25$ and 0.5 ($M \geq 0.7$), and lower values at $\rho_{tor} = 0.75$ ($M \leq 0.63$). Thus, these composite metrics provide a compact representation of the trends consistently identified in earlier sections—that the experimental fidelity of the code is significantly lower at $\rho_{tor} = 0.75$ than 0.25 and 0.5 in the modeled neutral beam heated DIII-D L-mode discharges, and that this trend is robust regardless of whether one

TABLE V. Values of composite metrics M_0, M_1, and M_2.

ρ_{tor}	M_0	M_1	M_2
0.25	0.75	0.75	0.74
0.5	0.70	0.70	0.70
0.75	0.60	0.63	0.63

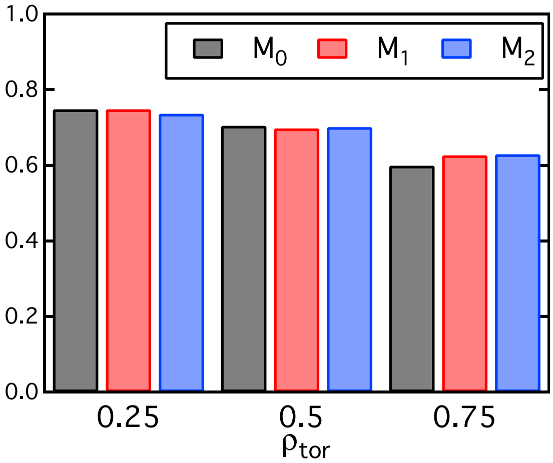

FIG. 21. Comparison of composite metrics M_0 (gray bar), M_1 (red bar), and M_2 (blue bar), calculated using Eqs. (36)–(38) and fractional error metrics shown in Figs. 14 and 20. The numerical values of each metric can be found in Table V.

considers predictions of just gradients and fluxes, or includes comparisons with measured fluctuation levels as well. However, these particular composite metric formulations are only intended to be illustrative, and the results can be sensitive to the specific mathematical formulations used. Future validation studies should more thoroughly investigate other formulations motivated by the particular physics and intended uses of the models under consideration.

VII. CONCLUSIONS AND FUTURE DIRECTIONS

Validated predictive models of plasma dynamics will play an increasingly important role for fusion research, both to ensure that planned future devices such as ITER can be operated safely and efficiently, and to help identify new and innovative configurations and operating scenarios that accelerate the realization of fusion as an economically viable commercial energy source. Validation is essential not only for building confidence in these predictions, but also for identifying parameter regimes where current models do not perform acceptably and improvements are needed. Building upon a number of previous reviews of validation best practices,[10,14,15] this paper illustrates some of the practical validation challenges for MFE validation research through the development of validation metrics suitable for testing of turbulent transport models. In order to go beyond earlier global transport metrics, these new metrics utilize the local sensitivity plots developed by the community for both verification[80] and validation[17,18] to quantify model–experiment agreement in terms of fractional flux-matching gradient errors, and associated residual discrepancies in other quantities such as fluctuation amplitudes. This choice of comparison quantities and metric formulation follows directly from the intended end use of the transport models—predicting equilibrium profiles and gradients in MFE plasmas as a function of magnetic configuration and auxiliary applied heating, fueling, and torque. Since the cross-field fluxes associated with these sources are never directly measured in the core of high-temperature MFE plasmas, simultaneous comparisons of predicted and measured fluctuation characteristics (through the use of synthetic diagnostics) in conjunction with comparisons with independent power balance flux calculations is identified as a key component for rigorous validation of plasma microturbulence models. Equally important is the explicit incorporation of uncertainties in measured profiles and fluctuation levels, power balance analyses, simulation outputs, and synthetic diagnostics into these validation metrics. The challenges in quantifying each of these uncertainties are discussed in detail in Secs. IV and V. Sec. VI discusses challenges and recommendations for combining the many individual validation metrics developed in Secs. IV and V into composite metrics that provided an integrated overall assessment of model fidelity, including formulation of bounded measures of agreement and use of primacy hierarchies as a means of weighting different comparisons. The key result found for the example experimental cases and code considered is that the code fidelity is significantly lower at $\rho_{tor} = 0.75$ than 0.25 and 0.5 in the modeled neutral beam heated DIII-D L-mode discharges, and that this trend is robust regardless of whether one considers predictions of just gradients and fluxes, or includes comparisons with measured fluctuation levels as well. These trends and results are common across each metric (E_{LT_i}, E_{Q_e}, $E_{\delta n_e}$ and $E_{\delta T_e}$) considered, and identify a need to improve model performance in this region for this class of plasmas.

Many of the challenges illustrated in this paper require further study and attention. Perhaps most important is to develop more rigorous and extensive methods for UQ appropriate for MFE plasmas. To do so, the MFE community should draw from the extensive current, on-going work on this topic by researchers in many other fields. Advances in UQ are needed on both the experimental and modeling sides, to better quantify and characterize the uncertainties in the equilibrium profiles and gradients, and then efficiently propagate these uncertainties through both power balance and turbulent transport models. On the experimental side, using techniques such as Gaussian process regression[119,120] or integrated data assessment[121–124] should be investigated more widely. For simulations, propagation of uncertainties requires a transition from single, deterministic simulations at a given point to ensembles of calculations as the default workflow. Such ensemble calculations are possible for conventional ion-scale gyrokinetic and gyrofluid simulations, but require large numbers of processor-hours to complete. As such, only a modest number of such ensemble calculations are practically feasible on current high performance computing (HPC) platforms on useful timescales. However, for robust validation across the full parameter space of interest for fusion, one must perform many such ensemble calculations, more than would currently be tractable. It is to be hoped that the next generation of exascale computing platforms currently being developed supports the workflows needed to perform larger numbers of such ensemble calculations. On the other hand, many-realization ensemble simulations of the reduced transport models commonly used in current predictive and interpretive transport modeling workflows are already viable on currently available HPC platforms. Since these models execute fairly quickly (on the core-minute or less timescale), reduced-model ensemble calculations offer an enormous opportunity to begin studying how to optimally propagate uncertainties in an MFE-relevant and specific context. For instance, techniques such as polynomial chaos expansions[166,167] should be investigated as means of more efficiently propagating uncertainties through plasma turbulence models than brute-force Monte Carlo simulation.

Finally, while the metrics that have been discussed in this paper are appropriate for testing predictions of continuously varying quantities, binary classification tests are also quite common and important in a variety of settings. In the context of MFE plasmas, these tests arise most frequently in tests of predicted global mode stability made by ideal or resistive magnetohydrodynamic (MHD) calculations. In these cases, the test is whether or not the plasma exhibits a specific behavior (e.g., onset of a specific global mode or disruption[168,169]) at the time or condition predicted by a given model. In MFE studies, such tests are generally plotted in terms of parameter space visualizations which indicate the predicted regions of (in)stability, combined with data points indicating where the measurements of whether the plasma is observed to be (un)stable. Although such plots are widely utilized in the fusion community and provide an analogous function to the local sensitivity plot approach describe here, formal validation metrics explicitly quantifying the fidelity of these models in predicting the onset of the dynamics in question have not been widely utilized to date. Drawing from the broad literature on binary outcome validation metrics in other fields such as medical research and machine learning, a number of different metric formulations could be pursued within the MFE community. At the simplest level, one could simply quantify model performance in terms of fraction correct (i.e., for N experimental measurements or conditions, how many are correctly predicted to be (un)stable?). More sophisticated widely used binary classification metrics incorporate information about relative amounts of true and false positive and negative predictions (such as the F_1-score[170]) or are derived from receiver–operator characteristic (ROC) plots.[171,172] However, while specific metrics to test this aspect of MHD theory have not been widely applied, many other aspects of MHD instability predictions, particular of mode structure and growth rate, have been tested extensively in a variety of machines. Examples include tests of mode growth rate,[173,174] and resistive wall mode,[173] neoclassical tearing mode,[175] and Alfvén eigenmode structure.[176–180] The same advances in data analysis, computing, and workflows required for improved validation of plasma microturbulence models be invaluable for improving our understanding of these equally important macroscopic phenomena.

ACKNOWLEDGMENTS

The author has benefitted greatly from discussion on this topic with many members of the research community over the years, including (in alphabetical order): E. M. Bass, R. V. Bravenec, K. H. Burrell, J. Candy, T. Carter, J. C. DeBoo, P. H. Diamond, A. Dimits, W. Dorland, T. Görler, B. A. Grierson, W. Guttenfelder, G. W. Hammett, D. R. Hatch, W. W. Heidbrink, N. T. Howard, F. Jenko, S. M. Kaye, J. E. Kinsey, T. Luce, G. R. McKee, D. R. Mikkelsen, D. E. Newman, W. M Nevins, S. E. Parker, C. C. Petty, M. J. Puschel, T. L. Rhodes, L. Schmitz, M. W. Shafer, S. P. Smith, P. Snyder, G. M. Staebler, P. Terry, G. R. Tynan, R. E. Waltz, and A. E. White. He thanks them all for their contributions, and apologizes in advance if any were inadvertently omitted. This research was supported by the Office of Science of the U.S. Department of Energy under Contract Nos. DE-FG02-06ER54871 and DE-SC0006957. It used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

[1] AIAA, "Guide for the verification and validation of computational fluid dynamics simulations," Technical Report No. AIAA G-077-1998(2002), The American Institute of Aeronautics and Astronautics, 1998.
[2] Fusion Energy Sciences Advisory Committee, Report on Strategic Planning, 2014.
[3] U.S. Department of Energy, *The Office of Science's Fusion Energy Sciences Program: A Ten-Year Perspective* (U.S. Dept. of Energy, 2015).
[4] ITER Physics Basis Editors and ITER Physics Expert Group Chairs and Co-Chairs and ITER Joint Central Team and Physics Integration Unit, Nucl. Fusion **39**, 2137 (1999).
[5] ASME, "Standard for verification and validation in computational fluid dynamics and heat transfer," Technical Report No. ASME Standard V&V 20-2009, The American Society of Mechanical Engineers, 2009.
[6] P. J. Roache, Ann. Rev. Fluid Mech. **29**, 123 (1997).
[7] F. Stern, R. V. Wilson, H. W. Coleman, and E. G. Paterson, J. Fluids Eng. **123**, 793 (2001).
[8] W. L. Oberkampf and T. Trucano, Prog. Aero. Sci. **38**, 209 (2002).
[9] W. Oberkampf, T. Trucano, and C. Hirsch, App. Mech. Rev. **57**, 345 (2004).
[10] W. L. Oberkampf and M. F. Barone, J. Comput. Phys. **217**, 5 (2006).
[11] Y. Liu, W. Chen, P. Arendt, and H.-Z. Huang, J. Mech. Design **133**, 071005 (2011).
[12] P. J. Roache, *Fundamentals of Verification and Validation* (Hermosa Publishers, 2009).
[13] W. L. Oberkampf and C. J. Roy, *Verification and Validation in Scientific Computing* (Cambridge University Press, 2012).
[14] P. W. Terry, M. Greenwald, J.-N. Leboeuf, G. R. McKee, D. R. Mikkelsen, W. M. Nevins, D. E. Newman, and D. Stotler, Phys. Plasmas **15**, 062503 (2008).
[15] M. Greenwald, Phys. Plasmas **17**, 058101 (2010).
[16] ITER Physics Expert Group on Confinement and Transport and ITER Physics Expert Group on Confinement Modelling and Database and ITER Physics Basis Editors, Nucl. Fusion **39**, 2175 (1999).
[17] D. W. Ross, R. V. Bravenec, W. Dorland, M. A. Beer, G. W. Hammett, G. R. McKee, R. J. Fonck, M. Murakami, K. H. Burrell, G. L. Jackson, and G. M. Staebler, Phys. Plasmas **9**, 177 (2002).
[18] D. W. Ross and W. Dorland, Phys. Plasmas **9**, 5031 (2002).
[19] G. W. Hammett and F. W. Perkins, Phys. Rev. Lett. **64**, 3019 (1990).
[20] E. J. Doyle, W. A. Houlberg, Y. Kamada, V. Mukhovatov, T. H. Osborne, A. Polevoi, G. Bateman, J. W. Connor, J. G. Cordey, T. Fujita, X. Garbet, T. S. Hahm, L. D. Horton, A. E. Hubbard, F. Imbeaux, F. Jenko, J. Kinsey, Y. Kishimoto, J. Li, T. C. Luce, Y. Martin, M. Ossipenko, V. Parail, A. Peeters, T. L. Rhodes, J. E. Rice, C. M. Roach, V. Rozhansky, F. Ryter, G. Saibene, R. Sartori, A. C. C. Sips, J. A. Snipes, M. Sugihara, E. J. Synakowski, H. Takenaga, T. Takizuka, K. Thomsen, M. R. Wade, H. R. Wilson, and ITPA Transport Physics Topical Group and ITPA Confinement Database and Modelling Topical Group and ITPA Pedestal and Edge Topical Group, Nucl. Fusion **47**, S18 (2007).
[21] See http://juq.siam.org/cgi-bin/main.plex for SIAM/ASA Journal on Uncertainty Quantification (JUQ).
[22] D. Borland and R. Taylor, IEEE Comput. Graphics Appl. **27**, 14 (2007).
[23] J. G. Cordey, B. Balet, D. Campbell, C. D. Challis, J. P. Christiansen, C. Gormezano, C. Gowers, D. Muir, E. Righi, G. R. Saibene, P. M. Stubberfield, and K. Thomsen, Plasma Phys. Controlled Fusion **38**, A67 (1996).

[24] C. C. Petty, Phys. Plasmas **15**, 080501 (2008).

[25] T. C. Luce, C. C. Petty, and J. G. Cordey, Plasma Phys. Controlled Fusion **50**, 043001 (2008).

[26] B. J. Green and ITER International Team and Participant Teams, Plasma Phys. Controlled Fusion **45**, 687 (2003).

[27] C. Gormezano, A. Sips, T. Luce, S. Ide, A. Becoulet, X. Litaudon, A. Isayama, J. Hobirk, M. Wade, T. Oikawa, R. Prater, A. Zvonkov, B. Lloyd, T. Suzuki, E. Barbato, P. Bonoli, C. Phillips, V. Vdovin, E. Joffrin, T. Casper, J. Ferron, D. Mazon, D. Moreau, R. Bundy, C. Kessel, A. Fukuyama, N. Hayashi, F. Imbeaux, M. Murakami, A. Polevoi, and H. S. John, Nucl. Fusion **47**, S285 (2007).

[28] A. W. Leonard, Phys. Plasmas **21**, 090501 (2014).

[29] T. E. Evans, Plasma Phys. Controlled Fusion **57**, 123001 (2015).

[30] E. Joffrin, M. Baruzzo, M. Beurskens, C. Bourdelle, S. Brezinsek, J. Bucalossi, P. Buratti, G. Calabro, C. D. Challis, M. Clever, J. Coenen, E. Delabie, R. Dux, P. Lomas, E. de la Luna, P. de Vries, J. Flanagan, L. Frassinetti, D. Frigione, C. Giroud, M. Groth, N. Hawkes, J. Hobirk, M. Lehnen, G. Maddison, J. Mailloux, C. Maggi, G. Matthews, M. Mayoral, A. Meigs, R. Neu, I. Nunes, T. Puetterich, F. Rimini, M. Sertoli, B. Sieglin, A. Sips, G. van Rooij, I. Voitsekhovitch, and J. Contributors, Nucl. Fusion **54**, 013011 (2014).

[31] H. Sugama and W. Horton, Phys. Plasmas **5**, 2560 (1998).

[32] S. Braginskii, *Review of Plasma Physics* (Consultants Bureau, New York, 1965), p. 214.

[33] F. Hinton and R. Hazeltine, Rev. Mod. Phys. **48**, 239 (1976).

[34] P. Helander and D. Sigmar, *Collisional Transport in Magnetized Plasmas* (Cambridge University Press, Cambridge, 2002).

[35] W. Horton, Rev. Mod. Phys. **71**, 735 (1999).

[36] J. Weiland, *Collective Modes in Inhomogeneous Plasmas* (Institute of Physics Publishing, London, 1999).

[37] E. Frieman and L. Chen, Phys. Fluids **25**, 502 (1982).

[38] G. W. Hammett, W. Dorland, and F. W. Perkins, Phys. Fluids B **4**, 2052 (1992).

[39] R. E. Waltz, R. R. Dominguez, and G. W. Hammett, Phys. Fluids B **4**, 3138 (1992).

[40] W. Dorland and G. W. Hammett, Phys. Fluids B **5**, 812 (1993).

[41] M. A. Beer and G. W. Hammett, Phys. Plasmas **3**, 4046 (1996).

[42] P. B. Snyder and G. W. Hammett, Phys. Plasmas **8**, 3199 (2001).

[43] G. M. Staebler, J. E. Kinsey, and R. E. Waltz, Phys. Plasmas **12**, 102508 (2005).

[44] B. Scott, Phys. Plasmas **17**, 102306 (2010).

[45] J. Madsen, Phys. Plasmas **20**, 072301 (2013).

[46] G. M. Staebler, J. E. Kinsey, and R. E. Waltz, Phys. Plasmas **14**, 055909 (2007).

[47] R. J. Hastie, Astrophys. Space Sci. **256**, 177 (1997).

[48] H. P. Furth, J. Killeen, and M. N. Rosenbluth, Phys. Fluids **6**, 459 (1963).

[49] R. Fitzpatrick, Nucl. Fusion **33**, 1049 (1993).

[50] D. A. Spong, Phys. Plasmas **22**, 055602 (2015).

[51] J. Scoville, D. Humphreys, J. Ferron, and P. Gohil, in *Proceedings of the 24th Symposium on Fusion Technology SOFT-24* [Fusion Eng. Des. **82**, 1045 (2007)].

[52] M. Kotschenreuther, G. Rewoldt, and W. Tang, Comput. Phys. Commun. **88**, 128 (1995).

[53] F. Jenko, W. Dorland, M. Kotschenreuther, and B. Rogers, Phys. Plasmas **7**, 1904 (2000).

[54] J. Candy and R. Waltz, J. Comput. Phys. **186**, 545 (2003).

[55] A. G. Peeters, Y. Camenen, F. J. Casson, W. A. Hornsby, A. P. Snodin, D. Strintzi, and G. Szepesi, Comput. Phys. Commun. **180**, 2650 (2009).

[56] S. Maeyama, T. Watanabe, Y. Idomura, M. Nakata, M. Nunami, and A. Ishizawa, Plasma Fusion Res. **8**, 1403150 (2013).

[57] Z. Lin, T. Hahm, W. Lee, W. Tang, and R. White, Science **281**, 1835 (1998).

[58] W. Wang, G. Rewoldt, W. Tang, F. Hinton, J. Manickam, L. Zakharov, R. White, and S. Kaye, Phys. Plasmas **13**, 082501 (2006).

[59] Y. Chen and S. E. Parker, J. Comput. Phys. **220**, 839 (2007).

[60] V. Grandgirard, Y. Sarazin, P. Angelino, A. Bottino, N. Crouseilles, G. Darmet, G. Dif-Pradalier, X. Garbet, P. Ghendrih, S. Jolliet, G. Latu, E. Sonnendrücker, and L. Villard, Plasma Phys. Controlled Fusion **49**, B173 (2007).

[61] S. Jolliet, A. Bottino, P. Angelino, R. Hatzky, T. Tran, B. Mcmillan, O. Sauter, K. Appert, Y. Idomura, and L. Villard, Comput. Phys. Commun. **177**, 409 (2007).

[62] S. Ku, C. S. Chang, and P. H. Diamond, Nucl. Fusion **49**, 115021 (2009).

[63] J. Candy, C. Holland, R. E. Waltz, M. R. Fahey, and E. Belli, Phys. Plasmas **16**, 060704 (2009).

[64] M. Barnes, I. G. Abel, W. Dorland, T. Goerler, G. W. Hammett, and F. Jenko, Phys. Plasmas **17**, 056109 (2010), APS Invited Paper DI3.00001.

[65] N. Howard, A. White, M. Greenwald, C. Holland, and J. Candy, Phys. Plasmas **21**, 032308 (2014).

[66] N. T. Howard, C. Holland, A. E. White, M. Greenwald, and J. Candy, Phys. Plasmas **21**, 112510 (2014).

[67] S. Maeyama, Y. Idomura, T.-H. Watanabe, M. Nakata, M. Yagi, N. Miyato, A. Ishizawa, and M. Nunami, Phys. Rev. Lett. **114**, 255002 (2015).

[68] N. T. Howard, C. Holland, A. E. White, M. Greenwald, and J. Candy, Nucl. Fusion **56**, 014004 (2016).

[69] A. Fukuyama, K. Itoh, S. I. Itoh, M. Yagi, and M. Azumi, Plasma Phys. Controlled Fusion **37**, 611 (1995).

[70] C. Bourdelle, X. Garbet, F. Imbeaux, A. Casati, N. Dubuit, R. Guirlet, and T. Parisot, Phys. Plasmas **14**, 112501 (2007).

[71] T. Rafiq, A. H. Kritz, J. Weiland, A. Y. Pankin, and L. Luo, Phys. Plasmas **20**, 032506 (2013).

[72] M. Kotschenreuther, W. Dorland, M. A. Beer, and G. W. Hammett, Phys. Plasmas **2**, 2381 (1995).

[73] R. E. Waltz, G. M. Staebler, W. Dorland, G. W. Hammett, M. Kotschenreuther, and J. A. Konings, Phys. Plasmas **4**, 2482 (1997).

[74] X. Garbet, P. Mantica, F. Ryter, G. Cordey, F. Imbeaux, C. Sozzi, A. Manini, E. Asp, V. Parail, R. Wolf, and the JET EFDA Contributors, Plasma Phys. Controlled Fusion **46**, 1351 (2004).

[75] F. Ryter, Y. Camenen, J. C. DeBoo, F. Imbeaux, P. Mantica, G. Regnoli, C. Sozzi, U. Stroth, A. Upgrade, DIII-D, FTU, J.-E. contributors, TCV, T. Supra, and W.-A. Teams, Plasma Phy. Controlled Fusion **48**, B453 (2006).

[76] P. Mantica, D. Strintzi, T. Tala, C. Giroud, T. Johnson, H. Leggate, E. Lerche, T. Loarer, A. G. Peeters, A. Salmi, S. Sharapov, D. Van Eester, P. C. de Vries, L. Zabeo, and K.-D. Zastrow, Phys. Rev. Lett. **102**, 175002 (2009).

[77] J. C. DeBoo, C. C. Petty, A. E. White, K. H. Burrell, E. J. Doyle, J. C. Hillesheim, C. Holland, G. R. McKee, T. L. Rhodes, L. Schmitz, S. P. Smith, G. Wang, and L. Zeng, Phys. Plasmas **19**, 082518 (2012).

[78] J. C. Hillesheim, J. C. DeBoo, W. A. Peebles, T. A. Carter, G. Wang, T. L. Rhodes, L. Schmitz, G. R. McKee, Z. Yan, G. M. Staebler, K. H. Burrell, E. J. Doyle, C. Holland, C. C. Petty, S. P. Smith, A. E. White, and L. Zeng, Phys. Rev. Lett. **110**, 045003 (2013).

[79] C. Roach, M. Walters, R. Budny, F. Imbeaux, T. Fredian, M. Greenwald, J. Stillerman, D. Alexander, J. Carlsson, J. Cary, F. Ryter, J. Stober, P. Gohil, C. Greenfield, M. Murakami, G. Bracco, B. Esposito, M. Romanelli, V. Parail, P. Stubberfield, I. Voitsekhovitch, C. Brickley, A. Field, Y. Sakamoto, T. Fujita, T. Fukuda, N. Hayashi, G. Hogeweij, A. Chudnovskiy, N. Kinerva, C. Kessel, T. Aniel, G. Hoang, J. Ongena, E. Doyle, W. Houlberg, A. Polevoi, and ITPA Confinement Database and Modelling Topical Group and ITPA Transport Physics Topical Group, Nucl. Fusion **48**, 125001 (2008).

[80] A. M. Dimits, G. Bateman, M. A. Beer, B. I. Cohen, W. Dorland, G. W. Hammett, C. Kim, J. E. Kinsey, M. Kotschenreuther, A. H. Kritz, L. L. Lao, J. Mandrekas, W. M. Nevins, S. E. Parker, A. J. Redd, D. E. Shumaker, R. Sydora, and J. Weiland, Phys. Plasmas **7**, 969 (2000).

[81] G. L. Falchetto, B. D. Scott, P. Angelino, A. Bottino, T. Dannert, V. Grandgirard, S. Janhunen, F. Jenko, S. Jolliet, A. Kendl, B. F. McMillan, V. Naulin, A. H. Nielsen, M. Ottaviani, A. G. Peeters, M. J. Pueschel, D. Reiser, T. T. Ribeiro, and M. Romanelli, Plasma Phys. Controlled Fusion **50**, 124015 (2008).

[82] D. R. Ernst, P. T. Bonoli, P. J. Catto, W. Dorland, C. L. Fiore, R. S. Granetz, M. Greenwald, A. E. Hubbard, M. Porkolab, M. H. Redi, J. E. Rice, K. Zhurovich, and A. C.-M. Group, Phys. Plasmas **11**, 2637 (2004).

[83] P. Mantica, C. Angioni, C. Challis, G. Colyer, L. Frassinetti, N. Hawkes, T. Johnson, M. Tsalas, P. C. deVries, J. Weiland, B. Baiocchi, M. N. A. Beurskens, A. C. A. Figueiredo, C. Giroud, J. Hobirk, E. Joffrin, E. Lerche, V. Naulin, A. G. Peeters, A. Salmi, C. Sozzi, D. Strintzi, G. Staebler, T. Tala, D. Van Eester, and T. Versloot, Phys. Rev. Lett. **107**, 135004 (2011).

[84] W. Guttenfelder, J. Candy, S. M. Kaye, W. M. Nevins, E. Wang, J. Zhang, R. E. Bell, N. A. Crocker, G. W. Hammett, B. P. LeBlanc, D. R. Mikkelsen, Y. Ren, and H. Yuh, Phys. Plasmas **19**, 056119 (2012).

[85] C. Holland, J. Kinsey, J. DeBoo, K. Burrell, T. Luce, S. Smith, C. Petty, A. White, T. Rhodes, L. Schmitz, E. Doyle, J. Hillesheim, G. McKee, Z. Yan, G. Wang, L. Zeng, B. Grierson, A. Marinoni, P. Mantica, P. Snyder, R. Waltz, G. Staebler, and J. Candy, Nucl. Fusion **53**, 083027 (2013).

[86] J. Citrin, F. Jenko, P. Mantica, D. Told, C. Bourdelle, J. Garcia, J. W. Haverkort, G. M. D. Hogeweij, T. Johnson, and M. J. Pueschel, Phys. Rev. Lett. **111**, 155001 (2013).

[87] J. Citrin, F. Jenko, P. Mantica, D. Told, C. Bourdelle, R. Dumont, J. Garcia, J. Haverkort, G. Hogeweij, T. Johnson, M. Pueschel, and JET-EFDA contributors, Nucl. Fusion **54**, 023008 (2014).

[88] D. Ernst, K. Burrell, W. Guttenfelder, T. Rhodes, L. Schmitz, A. Dimits, E. Doyle, B. Grierson, M. Greenwald, C. Holland, G. McKee, R. Perkins, C. Petty, J. Rost, D. Truong, G. Wang, L. Zeng, and the DIII-D and Alcator C-Mod Teams, in *Proceedings of the 2014 IAEA FEC Conference* (2014), Paper No. EX/2.

[89] N. T. Howard, A. E. White, M. Greenwald, C. Holland, J. Candy, and J. E. Rice, Plasma Phys. Controlled Fusion **56**, 124004 (2014).

[90] J. Citrin, J. Garcia, T. Görler, F. Jenko, P. Mantica, D. Told, C. Bourdelle, D. R. Hatch, G. M. D. Hogeweij, T. Johnson, M. J. Pueschel, and M. Schneider, Plasma Phys. Controlled Fusion **57**, 014032 (2015).

[91] A. Bañón Navarro, T. Happel, T. Görler, F. Jenko, J. Abiteboul, A. Bustos, H. Doerk, D. Told, and ASDEX Upgrade Team, Phys. Plasmas **22**, 042513 (2015).

[92] N. Bonanomi, P. Mantica, G. Szepesi, N. Hawkes, E. Lerche, P. Migliano, A. Peeters, C. Sozzi, M. Tsalas, D. V. Eester, and JET Contributors, Nucl. Fusion **55**, 113016 (2015).

[93] S. Smith, C. Petty, A. White, C. Holland, R. Bravenec, M. Austin, L. Zeng, and O. Meneghini, Nucl. Fusion **55**, 083011 (2015).

[94] A. E. White, N. T. Howard, A. J. Creely, M. A. Chilenski, M. Greenwald, A. E. Hubbard, J. W. Hughes, E. Marmar, J. E. Rice, J. M. Sierchio, C. Sung, J. R. Walk, D. G. Whyte, D. R. Mikkelsen, E. M. Edlund, C. Kung, C. Holland, J. Candy, C. C. Petty, M. L. Reinke, and C. Theiler, Phys. Plasmas **22**, 056109 (2015).

[95] R. Prater, D. Farina, Y. Gribov, R. Harvey, A. Ram, Y.-R. Lin-Liu, E. Poli, A. Smirnov, F. Volpe, E. Westerhof, A. Zvonkov, and the ITPA Steady State Operation Topical Group, Nucl. Fusion **48**, 035006 (2008).

[96] R. Budny, Nucl. Fusion **34**, 1247 (1994).

[97] R. V. Budny, D. R. Ernst, T. S. Hahm, D. C. McCune, J. P. Christiansen, J. G. Cordey, C. G. Gowers, K. Guenther, N. Hawkes, O. N. Jarvis, P. M. Stubberfield, K.-D. Zastrow, L. D. Horton, G. Saibene, R. Sartori, K. Thomsen, and M. G. von Hellermann, Phys. Plasmas **7**, 5038 (2000).

[98] W. Heidbrink and G. Sadler, Nucl. Fusion **34**, 535 (1994).

[99] Y. Peysson and J. Decker, Physics of Plasmas **15**, 092509 (2008).

[100] O. Meneghini, S. Shiraiwa, I. Faust, R. R. Parker, A. Schmidt, and G. Wallace, Fusion Sci. Technol. **60**, 40 (2011).

[101] J. C. Wright, A. Bader, L. A. Berry, P. T. Bonoli, R. W. Harvey, E. F. Jaeger, J.-P. Lee, A. Schmidt, E. D'Azevedo, I. Faust, C. K. Phillips, and E. Valeo, Plasma Phys. Controlled Fusion **56**, 045007 (2014).

[102] Y. Lin, S. J. Wukitch, P. T. Bonoli, E. Marmar, D. Mossessian, E. Nelson-Melby, P. Phillips, M. Porkolab, G. Schilling, S. Wolfe, and J. Wright, Plasma Phys. Controlled Fusion **45**, 1013 (2003).

[103] R. Prater, Phys. Plasmas **11**, 2349 (2004).

[104] P. T. Bonoli, J. Ko, R. Parker, A. E. Schmidt, G. Wallace, J. C. Wright, C. L. Fiore, A. E. Hubbard, J. Irby, E. Marmar, M. Porkolab, D. Terry, S. M. Wolfe, S. J. Wukitch, the Alcator C-Mod Team, J. R. Wilson, S. Scott, E. Valeo, C. K. Phillips, and R. W. Harvey, Phys. Plasmas **15**, 056117 (2008)

[105] P. T. Bonoli, Phys. Plasmas **21**, 061508 (2014).

[106] W. W. Heidbrink, Rev. Sci. Instrum. **81**, 10D727 (2010).

[107] Y. Lin, S. Wukitch, A. Parisot, J. C. Wright, N. Basse, P. Bonoli, E. Edlund, L. Lin, M. Porkolab, G. Schilling, and P. Phillips, Plasma Phys. Controlled Fusion **47**, 1207 (2005).

[108] N. Tsujii, M. Porkolab, P. T. Bonoli, E. M. Edlund, P. C. Ennever, Y. Lin, J. C. Wright, S. J. Wukitch, E. F. Jaeger, D. L. Green, and R. W. Harvey, Phys. Plasmas **22**, 082502 (2015).

[109] I. H. Hutchinson, *Principles of Plasma Diagnostics* (Cambridge University Press, 2005).

[110] A. Donné, A. Costley, R. Barnsley, H. Bindslev, R. Boivin, G. Conway, R. Fisher, R. Giannella, H. Hartfuss, M. von Hellermann, E. Hodgson, L. Ingesson, K. Itami, D. Johnson, Y. Kawano, T. Kondoh, A. Krasilnikov, Y. Kusama, A. Litnovsky, P. Lotte, P. Nielsen, T. Nishitani, F. Orsitto, B. Peterson, G. Razdobarin, J. Sanchez, M. Sasao, T. Sugie, G. Vayakis, V. Voitsenya, K. Vukolov, C. Walker, K. Young, and the ITPA Topical Group on Diagnostics, Nucl. Fusion **47**, S337 (2007).

[111] A. E. White, W. A. Peebles, T. L. Rhodes, C. Holland, G. Wang, L. Schmitz, T. A. Carter, J. C. Hillesheim, E. J. Doyle, L. Zeng, G. R. McKee, G. M. Staebler, R. E. Waltz, J. C. DeBoo, C. C. Petty, and K. H. Burrell, Phys. Plasmas **17**, 056103 (2010).

[112] H. E. S. John, T. S. Taylor, Y. R. Lin-Liu, and A. D. Turnbull, Plasma Phys. Controlled Nucl. Fusion Res. **3**, 603 (1994).

[113] A. E. White, L. Schmitz, G. R. McKee, C. Holland, W. A. Peebles, T. A. Carter, M. W. Shafer, M. E. Austin, K. H. Burrell, J. Candy, J. C. DeBoo, E. J. Doyle, M. A. Makowski, R. Prater, T. L. Rhodes, G. M. Staebler, G. R. Tynan, R. E. Waltz, and G. Wang, Phys. Plasmas **15**, 056116 (2008).

[114] C. Holland, A. White, G. McKee, M. Shafer, J. Candy, R. Waltz, L. Schmitz, and G. Tynan, Phys. Plasmas **16**, 052301 (2009).

[115] W. W. Heidbrink, Phys. Plasmas **15**, 055501 (2008).

[116] C. S. Chang and F. L. Hinton, Phys. Fluids **25**, 1493 (1982).

[117] W. A. Houlberg, K. C. Shaing, S. P. Hirshman, and M. C. Zarnstorff, Phys. Plasmas **4**, 3230 (1997).

[118] E. A. Belli and J. Candy, Plasma Phys. Controlled Fusion **50**, 095010 (2008); **54**, 015015 (2012).

[119] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA, 2006).

[120] M. A. Chilenski, M. Greenwald, Y. Marzouk, N. T. Howard, A. E. White, J. E. Rice, and J. R. Walk, Nucl. Fusion **55**, 023012 (2015).

[121] R. Fischer, C. j. Fuchs, B. Kurzan, W. Suttrop, E. Wolfrum, and ASDEX Upgrade Team, Fusion Sci. Technol. **58**, 675 (2010).

[122] B. P. van Milligen, T. Estrada, E. Ascasíbar, D. Tafalla, D. López-Bruna, A. L. Fraguas, J. A. Jiménez, I. García-Cortés, A. Dinklage, and R. Fischer, Rev. Sci. Instrum. **82**, 073503 (2011).

[123] J. Svensson, O. Ford, D. C. McDonald, A. Meakins, A. Werner, M. Brix, A. Boboc, M. Beurskens, and JET EFDA Contributors, Contrib. Plasma Phys. **51**, 152 (2011).

[124] M. Galante, L. Reusch, D. D. Hartog, P. Franz, J. Johnson, M. McGarry, M. Nornberg, and H. Stephens, Nucl. Fusion **55**, 123016 (2015).

[125] K. H. Burrell, W. P. West, E. J. Doyle, M. E. Austin, J. S. deGrassie, P. Gohil, C. M. Greenfield, R. J. Groebner, R. Jayakumar, D. H. Kaplan, L. L. Lao, A. W. Leonard, M. A. Makowski, G. R. McKee, W. M. Solomon, D. M. Thomas, T. L. Rhodes, M. R. Wade, G. Wang, J. G. Watkins, and L. Zeng, Plasma Phys. Controlled Fusion **46**, A165 (2004).

[126] H. Bindslev and D. Bartlett, A Technique for Improving the Relative Accuracy of JET ECE Temperature Profiles, Report No. JET-R-88-04, 1988.

[127] D. R. Mikkelsen and W. Dorland, Bull. Am. Phys. Soc. **50**, 196 (2005).

[128] K. H. Burrell, Phys. Plasmas **4**, 1499 (1997).

[129] P. W. Terry, Rev. Mod. Phys. **72**, 109 (2000).

[130] C. Holland, L. Schmitz, T. Rhodes, W. Peebles, J. Hillesheim, G. Wang, L. Zeng, E. Doyle, S. Smith, R. Prater, K. Burrell, J. Candy, R. Waltz, J. Kinsey, G. Staebler, J. DeBoo, C. Petty, G. McKee, Z. Yan, and A. White, Phys. Plasmas **18**, 056113 (2011).

[131] T. L. Rhodes, C. Holland, S. P. Smith, A. E. White, K. H. Burrell, J. Candy, J. C. DeBoo, E. J. Doyle, J. C. Hillesheim, J. E. Kinsey, G. R. McKee, D. Mikkelsen, W. A. Peebles, C. C. Petty, R. Prater, S. Parker, Y. Chen, L. Schmitz, G. M. Staebler, R. E. Waltz, G. Wang, Z. Yan, and L. Zeng, Nucl. Fusion **51**, 063022 (2011).

[132] J. Callen, R. Colchin, R. Fowler, D. McAlees, and J. Rome, in *Proceedings of the 5th International Conference on Plasma Physics and Controlled Nuclear Fusion Research, Tokyo* (1974), Vol. 1, p. 645.

[133] K. Matsuda, IEEE Trans. Plasma Sci. **17**, 6 (1989).

[134] The GYRO source code is available at https://github.com/gafusion/gacode. Version ID r4-864-g6ea4 was used in this work.

[135] J. Candy, Plasma Phys. Controlled Fusion **51**, 105009 (2009).

[136] R. V. Bravenec, J. Candy, M. Barnes, and C. Holland, Phys. Plasmas **18**, 122505 (2011).

[137] J. Chowdhury, W. Wan, Y. Chen, S. E. Parker, R. J. Groebner, C. Holland, and N. T. Howard, Phys. Plasmas **21**, 112503 (2014).

[138] T. Görler, A. E. White, D. Told, F. Jenko, C. Holland, and T. L. Rhodes, Phys. Plasmas **21**, 122307 (2014).

[139] T. Görler, A. E. White, D. Told, F. Jenko, C. Holland, and T. L. Rhodes, Fusion Sci. Technol. **69**, 537 (2015).

[140] F. Jenko, J. Plasma Fusion Res. **6**, 11 (2004).

[141] R. E. Waltz, J. Candy, and M. Fahey, Phys. Plasmas **14**, 056116 (2007).

[142] J. Candy, R. E. Waltz, M. R. Fahey, and C. Holland, Plasma Phys. Controlled Fusion **49**, 1209 (2007).

[143] T. Görler and F. Jenko, Phys. Rev. Lett. **100**, 185002 (2008).

[144] T. Görler and F. Jenko, Phys. Plasmas **15**, 102508 (2008).

[145] L. Schmitz, C. Holland, T. L. Rhodes, G. Wang, L. Zeng, A. E. White, J. C. Hillesheim, W. A. Peebles, S. P. Smith, R. Prater, G. R. McKee, Z. Yan, W. M. Solomon, K. H. Burrell, C. T. Holcomb, E. J. Doyle, J. C. DeBoo, M. E. Austin, J. S. deGrassie, and C. C. Petty, Nucl. Fusion **52**, 023003 (2012).

[146] N. T. Howard, C. Holland, A. E. White, M. Greenwald, and J. Candy, Plasma Phys. Controlled Fusion **57**, 065009 (2015).
[147] P. Ricci, C. Theiler, A. Fasoli, I. Furno, K. Gustafson, D. Iraji, and J. Loizu, Phys. Plasmas **18**, 032109 (2011).
[148] P. Ricci, F. Riva, C. Theiler, A. Fasoli, I. Furno, F. D. Halpern, and J. Loizu, Phys. Plasmas **22**, 055704 (2015).
[149] K. E. Taylor, J. Geo. Res. **106**, 7183 (2001).
[150] T. L. Rhodes, J.-N. Leboeuf, R. D. Sydora, R. J. Groebner, E. J. Doyle, G. R. McKee, W. A. Peebles, C. L. Rettig, L. Zeng, and G. Wang, Phys. Plasmas **9**, 2141 (2002).
[151] P. Ricci, F. D. Halpern, S. Jolliet, J. Loizu, A. Mosetto, A. Fasoli, I. Furno, and C. Theiler, Plasma Phys. Controlled Fusion **54**, 124047 (2012).
[152] A. Fasoli, B. Labit, M. McGrath, S. H. Müller, G. Plyushchev, M. Podestá, and F. M. Poli, Phys. Plasmas **13**, 055902 (2006); A. Fasoli, A. Burckel, L. Federspiel, I. Furno, K. Gustafson, D. Iraji, B. Labit, J. Loizu, G. Plyushchev, P. Ricci, C. Theiler, A. Diallo, S. H. Mueller, M. Podestá, and F. Poli, Plasma Phys. Controlled Fusion **52**, 124020 (2010).
[153] R. V. Bravenec and W. M. Nevins, Rev. Sci. Instrum. **77**, 015101 (2006).
[154] D. A. Russell, J. R. Myra, D. A. D'Ippolito, T. L. Munsat, Y. Sechrest, R. J. Maqueda, D. P. Stotler, S. J. Zweben, and T. N. Team, Phys. Plasmas **18**, 022306 (2011).
[155] G. R. McKee, R. J. Fonck, D. K. Gupta, D. J. Schlossberg, M. W. Shafer, and R. L. Boivin, Rev. Sci. Instrum. **77**, 10F104 (2006).
[156] M. W. Shafer, R. J. Fonck, G. R. McKee, and D. J. Schlossberg, Rev. Sci. Instrum. **77**, 10F110 (2006).
[157] C. Holland, J. Candy, R. E. Waltz, A. E. White, G. R. McKee, M. W. Shafer, L. Schmitz, and G. R. Tynan, J. Phys.: Conf. Ser. **125**, 012043 (2008).
[158] J. S. Bendat and A. G. Piersol, *Engineering Applications of Correlation and Spectral Analysis* (John Wiley & Sons, Inc., 1993).
[159] J. S. Bendat and A. G. Piersol, *Random Data: Analysis and Measurement Procedures* (John Wiley & Sons, Inc., 2010).
[160] P. H. Diamond, S.-I. Itoh, K. Itoh, and T. S. Hahm, Plasma Phys. Controlled Fusion **47**, R35 (2005).
[161] P. W. Terry, D. A. Baver, and S. Gupta, Phys. Plasmas **13**, 022307 (2006).
[162] K. D. Makwana, P. W. Terry, J.-H. Kim, and D. R. Hatch, Phys. Plasmas **18**, 012302 (2011).
[163] D. R. Hatch, P. W. Terry, F. Jenko, F. Merz, and W. M. Nevins, Phys. Rev. Lett. **106**, 115003 (2011).
[164] M. W. Shafer, R. J. Fonck, G. R. McKee, C. Holland, A. E. White, and D. J. Schlossberg, Phys. Plasmas **19**, 032504 (2012).
[165] P. Ricci, C. Theiler, A. Fasoli, I. Furno, B. Labit, S. H. Müller, M. Podestá, and F. M. Poli, Phys. Plasmas **16**, 055703 (2009).
[166] N. Wiener, Am. J. Math. **60**, 897 (1938).
[167] H. N. Najm, Ann. Rev. Fluid Mech. **41**, 35 (2009).
[168] ITER Physics Expert Group on Disruptions, Plasma Control, and MHD and ITER Physics Basis Editors, Nucl. Fusion **39**, 2251 (1999).
[169] T. Hender, J. Wesley, J. Bialek, A. Bondeson, A. Boozer, R. Buttery, A. Garofalo, T. Goodman, R. Granetz, Y. Gribov, O. Gruber, M. Gryaznevich, G. Giruzzi, S. Günter, N. Hayashi, P. Helander, C. Hegna, D. Howell, D. Humphreys, G. Huysmans, A. Hyatt, A. Isayama, S. Jardin, Y. Kawano, A. Kellman, C. Kessel, H. Koslowski, R. L. Haye, E. Lazzaro, Y. Liu, V. Lukash, J. Manickam, S. Medvedev, V. Mertens, S. Mirnov, Y. Nakamura, G. Navratil, M. Okabayashi, T. Ozeki, R. Paccagnella, G. Pautasso, F. Porcelli, V. Pustovitov, V. Riccardo, M. Sato, O. Sauter, M. Schaffer, M. Shimada, P. Sonato, E. Strait, M. Sugihara, M. Takechi, A. Turnbull, E. Westerhof, D. Whyte, R. Yoshino, H. Zohm, and the ITPA MHD, Disruption and Magnetic Control Topical Group, Nucl. Fusion **47**, S128 (2007).
[170] See https://en.wikipedia.org/wiki/F1_score for details on the how F_1 score is calculated and used for binary classification tests.
[171] N. A. Macmillan and C. D. Creelman, *Detection Theory: A User's Guide* (Lawrence Erlbaum Associates, Inc., 2005).
[172] T. Fawcett, Pattern Recognit. Lett. **27**, 861 (2006).
[173] A. Turnbull, D. Brennan, M. Chu, L. Lao, J. Ferron, A. Garofalo, P. Snyder, J. Bialek, I. Bogatu, J. Callen, M. Chance, K. Comer, D. Edgell, S. Galkin, D. Humphreys, J. Kim, R. L. Haye, T. Luce, G. Navratil, M. Okabayashi, T. Osborne, B. Rice, E. Strait, T. Taylor, and H. Wilson, Nucl. Fusion **42**, 917 (2002).
[174] A. D. Turbull, D. P. Brennan, M. S. Chu, L. L. Lao, and P. B. Snyder, Fusion Sci. Technol. **48**, 875 (2005).
[175] J. H. Yu, M. A. Van Zeeland, M. S. Chu, V. A. Izzo, and R. J. La Haye, Phys. Plasmas **16**, 056114 (2009).
[176] M. A. Van Zeeland, G. J. Kramer, M. E. Austin, R. L. Boivin, W. W. Heidbrink, M. A. Makowski, G. R. McKee, R. Nazikian, W. M. Solomon, and G. Wang, Phys. Rev. Lett. **97**, 135001 (2006).
[177] I. G. J. Classen, P. Lauber, D. Curran, J. E. Boom, B. J. Tobias, C. W. Domier, N. C. Luhmann, Jr., H. K. Park, M. G. Munoz, B. Geiger, M. Maraschek, M. A. V. Zeeland, S. da Graa, and the ASDEX Upgrade Team, Plasma Phys. Controlled Fusion **53**, 124018 (2011).
[178] B. J. Tobias, I. G. J. Classen, C. W. Domier, W. W. Heidbrink, N. C. Luhmann, R. Nazikian, H. K. Park, D. A. Spong, and M. A. Van Zeeland, Phys. Rev. Lett. **106**, 075003 (2011).
[179] B. J. Tobias, R. L. Boivin, J. E. Boom, I. G. J. Classen, C. W. Domier, A. J. H. Donné, W. W. Heidbrink, N. C. Luhmann, T. Munsat, C. M. Muscatello, R. Nazikian, H. K. Park, D. A. Spong, A. D. Turnbull, M. A. Van Zeeland, G. S. Yun, and D.-D. Team, Phys. Plasmas **18**, 056107 (2011).
[180] D. A. Spong, E. M. Bass, W. Deng, W. W. Heidbrink, Z. Lin, B. Tobias, M. A. Van Zeeland, M. E. Austin, C. W. Domier, and N. C. Luhmann, Phys. Plasmas **19**, 082511 (2012).
[181] J. L. Luxon, Nucl. Fusion **42**, 614 (2002).