Verification and Validation
Toward Guidelines and Good Practices

Presented by David Newman

for

M. Greenwald, J.-N. Leboeuf, G. McKee, D. Mikkelsen, W. Nevins, D. Newman, D. Stotler, and P. Terry

Verification and Validation Task Group, USBPO and TTF

Phys 693 7 Nov 2013

from a Journal club presentation 19 Oct 2007

Why are we building models

We want to understand nature

- Remember models are at best a representation of the the physical world
    - keep their limitations in mind (qualification)
- To be useful models must have either a predictive capability or an explanatory capability
    - Predict something new (regime etc)
    - Clarify some physical process (Occam's razor)

To have confidence in the model one must practice:

Validation - Solving the right equations

Verification - Solving the equations right

"Verification and Validation in Computational science and Engineering", by Patrick Roache 1998 Hermosa Press

Verification of codes

Does discrete solution approach continuum solution

- Spatial resolution convergence studies
- Time resolution convergence studies

- Test problems with analytic solutions to compare to

Validation through comparison

Don't ask for more then can be supplied

Understand the art of asking questions in a foreign culture

Don't expect apples and oranges to have meaningful comparisons at many levels

- Similarities and differences important
- Linear effects
- Nonlinear effects
  - Turbulent dynamics
  - Transport dynamics
    - Different signatures of same dynamics
      - Honey vs water example

A basic procedure is to look for similarities and differences (universality) in characteristic measures

- Model-model comparisons
- Model-experiment comparisons
- Experiment-experiment comparisons

Verification and validation in fusion: a brief history

- Pioneering efforts: Model/experiment comparisons

  Qualitative; limited assessment of uncertainty, sensitivity, error

  Issues with credibility

- Oberkampf (SLC TTF): Standardized procedures for testing models

  Verification: numerical algorithm faithfully solves mathematical model

  Validation: Mathematical model faithfully represents real world

- Practiced in stockpile stewardship, fluid dynamics (engineering performance, software reliability)
- Fusion community: Mostly verification to date

  Orchestrated benchmarking exercises - GEM, CYCLONE

Verification efforts underway; focus here on collective task of validation

Goal of predictive capability drives need for verification and validation

- US 10 year goal: "progress toward predictive understanding"
  - ⇒ Working toward: demonstrably predictive models within tolerances
- Process of getting there: validation under commonly understood standards for what constitutes agreement between models and experiment
- Significant challenges
  - Resource limitations (budget, manpower)
  - Complexity of modeling
    - Complexities of turbulence [multiple scales, nonlinearity, geometry (b.c.)]
    - Different regions - different physics, different models
  - Difficulties with measurement
    - Limited access
    - Limited diagnostic capability
    - Plasma diagnostics involve significant modeling a priori

Fusion community is just starting to think seriously about validation

Setting out guidelines is evolving process - much still to be learned

Hope: validation becomes part of research culture

- We will learn as we go
- "good practices" become better as we learn

Different models will have different levels of validation, guidelines not rigid

- Details will be individualized
- Onus on researcher to make convincing case for validation
- Widely accepted guidelines will build confidence

Outline

- Key concepts
- Approaches to code validation
  - Useful starting points for experiment/model comparison
  - Sources of discrepancy between experiment and models
  - Primacy hierarchy of measured quantities
  - Landscape of model behavior
  - Validation metric
  - Working the primacy hierarchy
- Changing the culture of modeling
- Where we go from here
- Questions for discussion

Validation as collective endeavor $\Rightarrow$ standardized concepts

From glossary, key concepts for validation

- Prediction - use of code outside previously validated domain to foretell state of physical system
- Validation - process of determining degree to which model is accurate representation of real world, given intended uses
- Qualification - theoretical specification of expected domain of applicability of model
- Uncertainty - potential deficiency in modeling process due to lack of knowledge, either in model or in experimental data used for validation
- Sensitivity analysis - study of how output variation is apportioned to different sources of variation
- Primacy hierarchy - ranking of measurable quantity in terms of extent to which other effects integrate to set value of quantity
- Validation metric - assessment, and rating of uncertainties and primacy hierarchies, given sensitivities, to quantify degree to which model is accurate representation of real world

Obvious but not-to-be-forgotten points for experiment/model comparisons

Code validation is a joint enterprise between modeling, experiment, theory

Long term product of US fusion sciences: Validated predictive model or set of models for moving to DEMO, commercialization

- Use of common units

  e.g., SI units (including μ_0 and ε_0)

- Full disclosure of simple (easily overlooked) conventions

  e.g., $\sqrt{2}$ in v_{th}

- Common understanding of what quantities are measured or could be measured

  including limitations, effect of modeling in diagnostic

- Application of experimental resources (runtime) for validation work

  may not be the most interesting runs from physics or fusion perspective

- Application of qualified models appropriate to experimental conditions

Important to identify, understand and quantitatively assess sources of discrepancy between models and experiments

Central to several validation elements:

- Error and Uncertainty

    What are a priori deficiencies in model or experimental measurement?

- Qualification

    Under what conditions would model deficiencies not be expected to affect a comparison, or to affect only within some tolerance?

- Validation metric

    Assign confidence level to results of validation activity

    Confront disagreement in quantitative detail, figure out its source

    Can deficiencies be quantified?

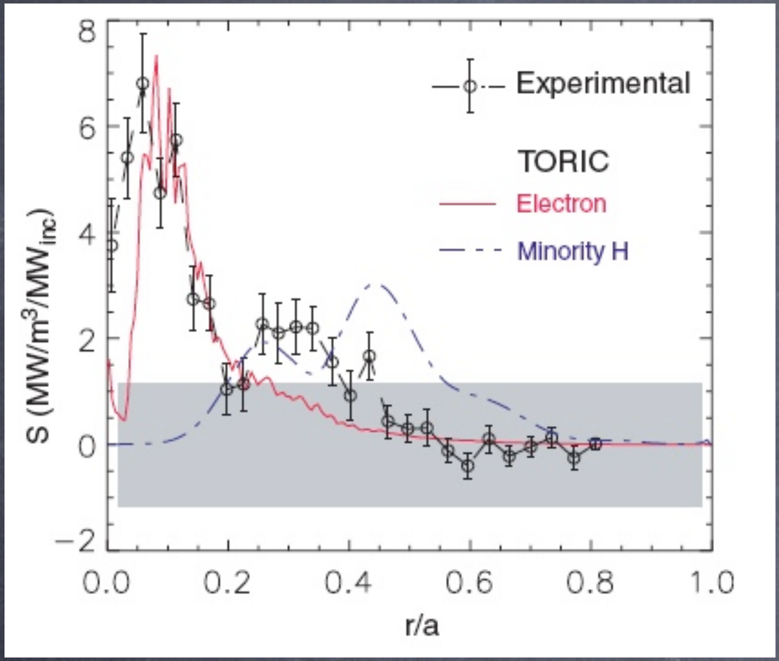
    Can differences in comparison results be reasonably attributed to deficiencies?

    Reasonable: Qualification of model (where and how deficiencies arise)

    Quantitative assessment of deficiencies (magnitude of effect)

    Are there refinements to comparison that could establish source of disagreement between model and experiment?

For validation, "generally in agreement" needs to followed up with quantitative analysis of features not in agreement



- Agreement is generally good
- Qualitative discussion of

  -Shift of peak near magnetic axis

  -Second peak

Need

- Quantitative analysis - demonstrate

  sources of disagreement are identified

- Can systematic deviations be bounded?

Mode converted electron heating profile from ICRF in C-Mod

Modeling from toroidal full-wave ICRF

Discrepancies include statistical error and systematic deficiencies in experiment

- Statistical error
  - Relatively easy to rate; often exclusive content of error bars
  - Important to describe how error bars are arrived at
  - Magnitude relies on statistical assumptions that may not be valid
    - Large ensembles (Markov), sampling $\Rightarrow$ Gaussian
    - Dynamical fluctuations need not obey Gaussian statistics
- Uncertainty in experiment (mostly systematic error)
  - Equilibrium solver
  - Lack of precision in input to equilibrium solver
  - Diagnostic sensitivity
  - Diagnostic resolution
  - Inversions
  - Modeling is intrinsic to diagnostics
  - Processing and interpretation of diagnostic signals

Models and simulations often have numerous uncertainties
Qualification issue

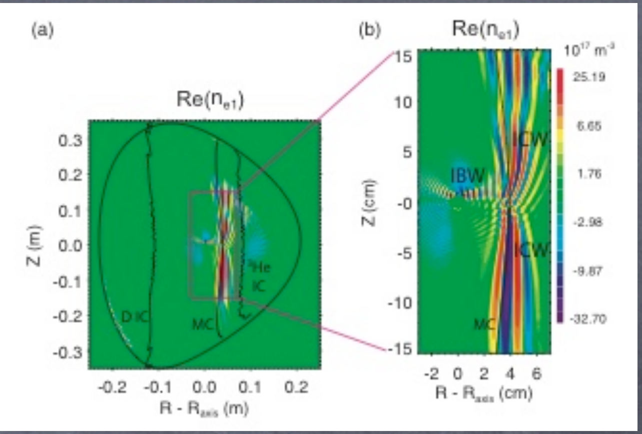
Practical considerations may dictate reduced models even if models with fewer limitations exist $\Rightarrow$ assessing uncertainties unavoidable

- Mapping magnetic topology to coordinates
- Equilibrium specification [fixed or variable; subject to modeling]
- Limitations on physical processes included [missing fields, missing kinetic effects, boundary representation, inhomogeneities not included (flow)]
- Limitations on sampling [in singular layers; scale ranges]
- Integration time [long time correlations, coupling of transport to turbulent time scale]
- Artificial constraints [fixed profile, flux tube, missing or imprecise experimental data for input parameters]
- Resolution [large scale, small scale, time step]
- Representation of dissipative processes

Discrepancies associated with diagnostics can be handled with synthetic diagnostics in simulation

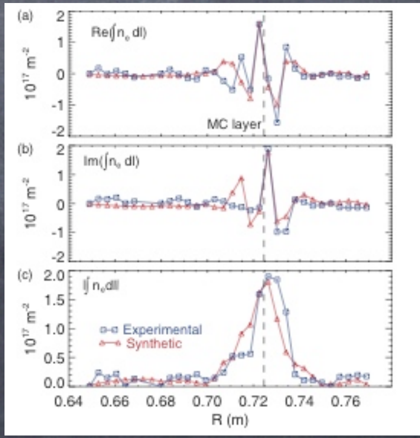
Synthetic diagnostics emulate experimental diagnostics in processing of raw input data

- Include spatial and temporal transfer functions
- Mimic Resolution and sensitivity limitations
- Replicate plasma modeling inherent in diagnostic signal interpretation



Useful for sensitivity studies of experimental data:

Can distinct inputs to diagnostic yield indistinguishable output signals?



Useful for quantifying modeling effects, physics uncertainties in experimental diagnostics

Important to understand factors in experiment and models affecting fidelity and significance of validation comparisons

- Some measured quantities are more sensitive discriminators between different models
- Some measured quantities are poor discriminators

  Very different models seem to do about as well

- Some measured quantities can be susceptible to false positives
- Some measured quantities have model assumptions folded into them

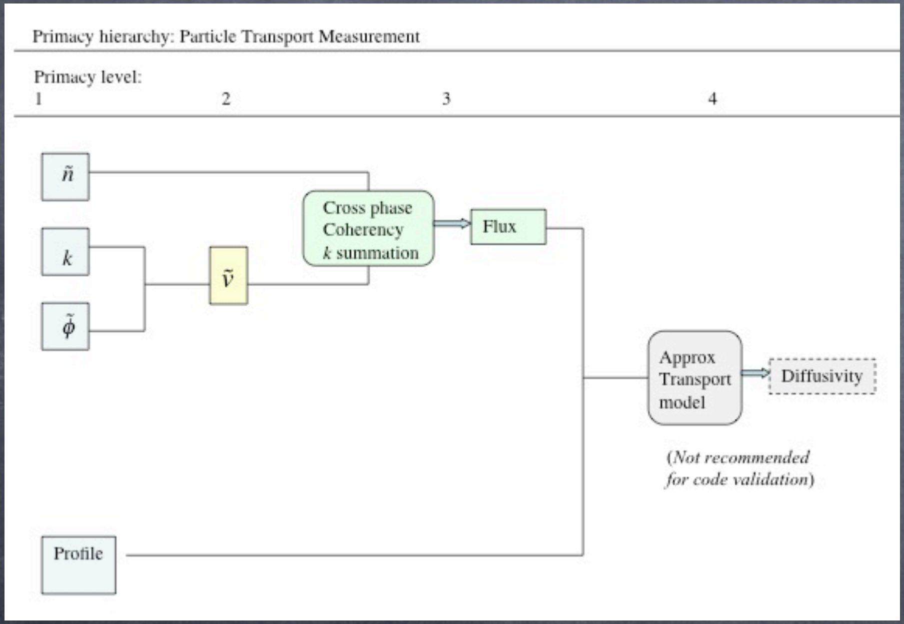
$\Rightarrow$ Not all measured quantities and comparisons are equally meaningful in validation

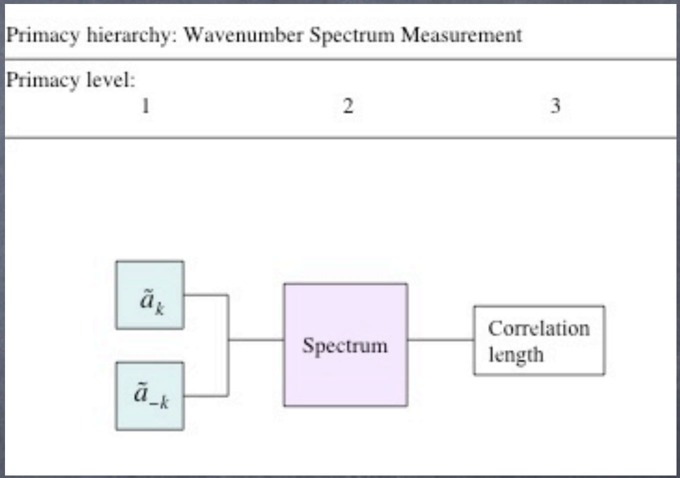
To quantify these effects:

- Primacy hierarchy (mostly measured quantities)
- Sensitivity analysis (mostly models)

Primacy hierarchy: ranking of measured quantities in terms of extent to which other effects integrate to set value of quantity

Can be constructed in various ways for various types of comparisons





Lower primacy level: fewer effects integrated

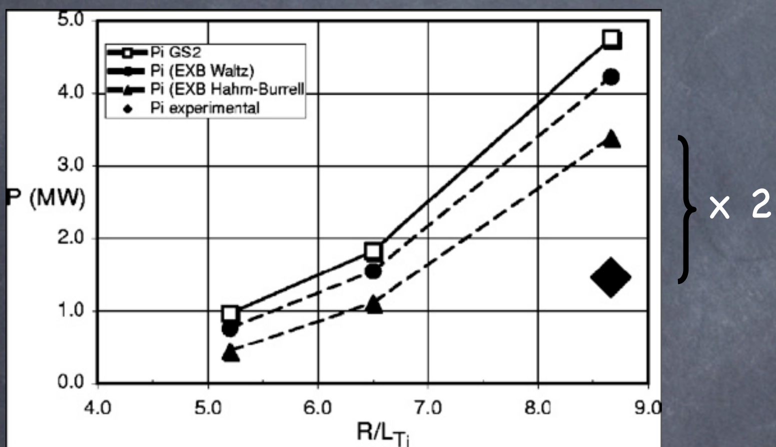
Measurements at multiple levels recommended, with awareness of hierarchy

Primacy hierarchy evident in comparisons with gyrokinetic models

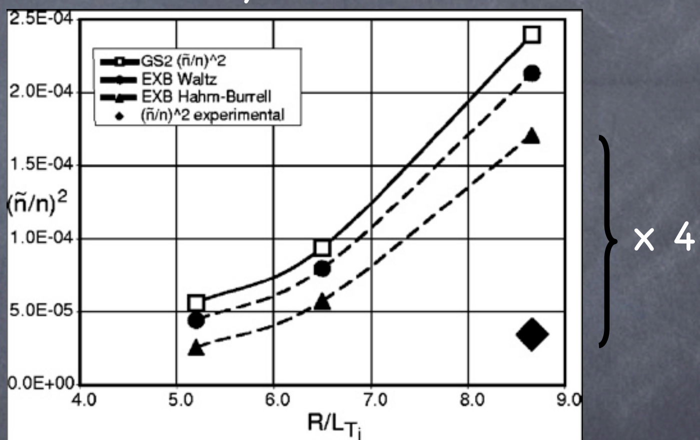
Fluxes (level 3) are in closer agreement than fluctuations (level 1)

⇒ higher level - reduced capability for discrimination between models

flux



density fluctuation



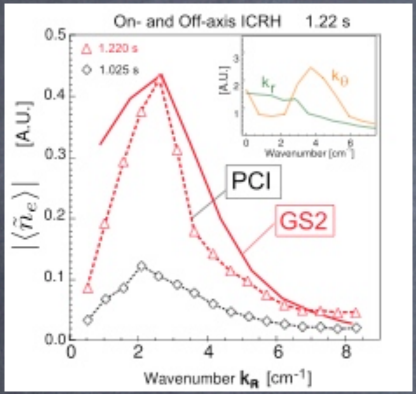
GS2 - Ross and Dorland, Phys. Plasmas **9**, 5031 (2002)

Understanding how effects integrate physically is also useful in assessing comparisons

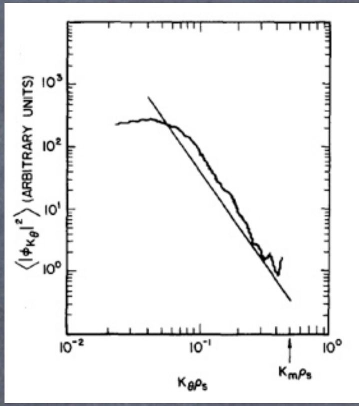
Historically: *k* spectrum agreement easier to get than other quantities

decreasing model complexity, analysis sophistication →

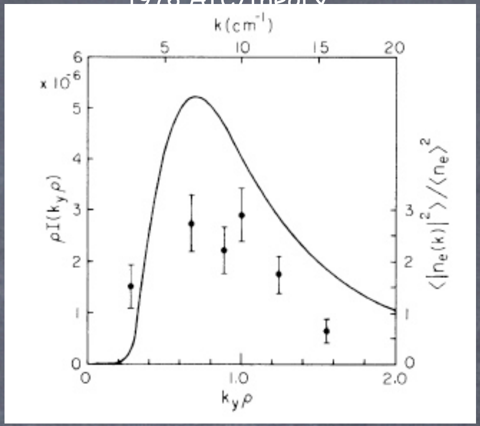
2006 CMod/GS2



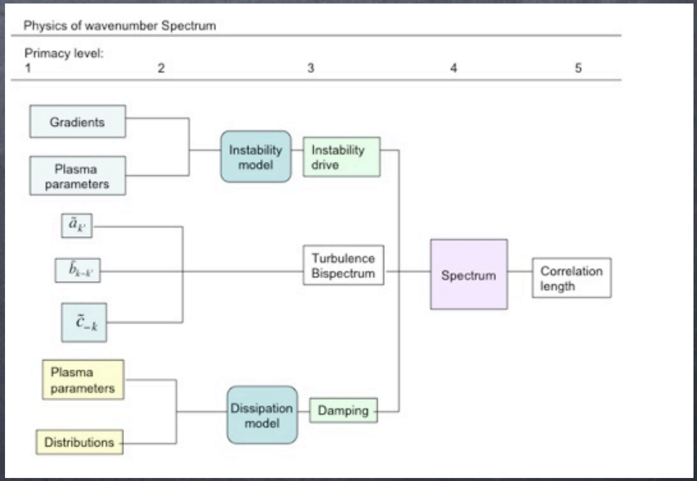
1985 TEXT/theory



1976 ATC/theory



Spectrum is amalgam of lower-order processes



Most significant physics at lower level
- Goes into calculation of spectrum
- But folding makes spectrum a poor discriminator between models

Primacy hierarchies are useful in assigning confidence level to validation activities and tracing effects of uncertainties

- Identify possibility that errors/uncertainties are canceling
- Sort out error/uncertainty propagation

  Holistic view of error/uncertainty sources and folding paths

  Tracing backwards through hierarchy helps identify most important uncertainties

- Assess ability of measurements to discriminate between different models

  Synthetic diagnostics applied at higher levels might further degrade ability to discriminate between models $\rightarrow$ apply to lower levels

- Hierarchies not necessarily unique in form

  Important to make comparisons at multiple levels

  Grappling with way effects integrate in comparisons more important than detailed from of hierarchy

Complexity of plasma (or other system) dynamics must be confronted in validation

Plasma dynamics is nonlinear and complex:

- Bifurcations

  e.g., transitions to enhanced confinement regimes

- Stiffness

  e.g., dependence of fluctuations, fluxes on profiles

- Many parameters
- Extreme sensitivity to certain parameters

  e.g., edge heat flux at L-H transition

- Different behavior in different parameter regimes

  e.g., collisionality switches nonlinear behavior on/off in electron dynamics

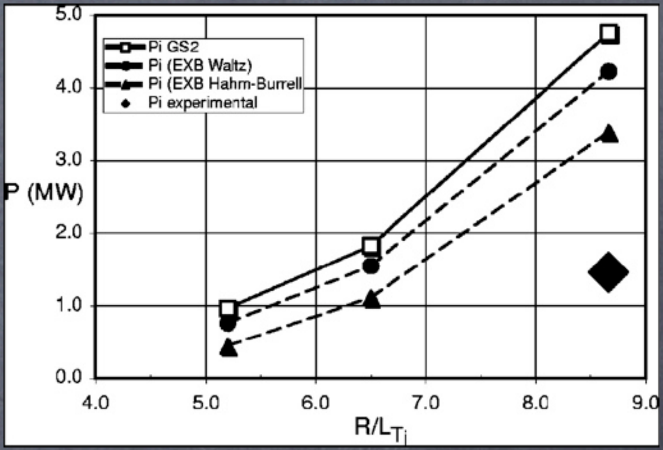
Any of above can pose serious problems for validation

How to deal with it:

- Basic theory understanding
- Sensitivity analysis

Theory understanding is crucial in validation
Again, Qualification issue

- Identifies features of dynamical landscape
- Lays out workings of processes creating landscape
- Provides qualitative and quantitative description of dynamics
  - Basic scalings
    - Which parameters crucial
    - Where most extreme sensitivities are
  - Morphology of dynamical behavior
- Identifies previously unknown effects
- Creates conceptual framework



- Example: EXB shear
  - Effect that cannot be ignored
  - Scalings for effect on fluctuations, transport
  - Must be accounted for in validation, doesn't fully close gap in GS2 comparison
- Validation will fail or lack credibility if done in theoretical vacuum
- Commensurate development of theoretical understanding essential

Validation will not be credible without sensitivity analysis

Certain measurable quantities vary more strongly with certain parameters on which they depend than on other parameters

- Sensitivity of fluctuations, fluxes to profiles is problem in every comparison to date

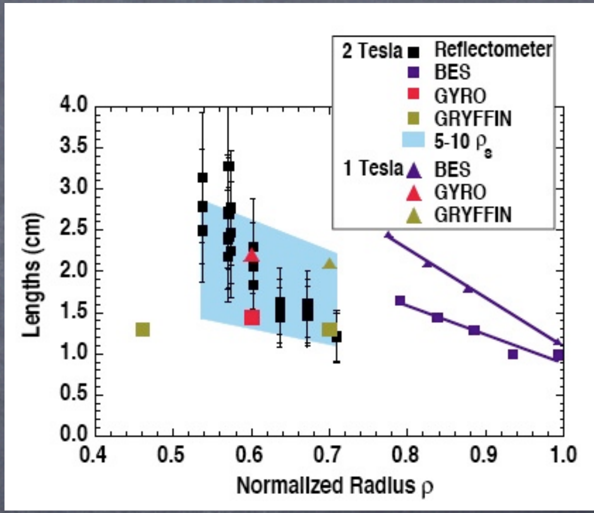
Difficulty:

- Agreement extremely difficult in some quantities
- Agreement too easy in others

Recommendations:

- Must map out sensitivity of all parameters

    Use theory for guidance



- Looking at quantities that remove sensitivity may help agreement, but may limit ability to discriminate

    Example: radial correlation length

- Sensitivity to computational effects also important

    Particle noise

    Simulation time

    Resolution

Uncertainties, primacy hierarchies, and sensitivities are rated in a validation metric

- Assign confidence level to results of validation activity
- Confront disagreement in quantitative detail, figure out its source

- Uncertainty (and error) - how to grade it:
  - Which uncertainties have been subjected to quantitative testing? Which have not?
  - Are there bounds associated with reasonable variation?
    - Use synthetic diagnostics to bound uncertainties associated with resolution, sensitivity
  - Are there nonlinear effects from combinations of uncertainties?
  - What are their bounds?
- Researcher develops grading scheme
  - Low score - higher confidence level
  - High score - lower confidence level

Construction of a Composite Validation metric V

We attempt to construct an admittedly non-unique composite validation metric. The idea behind this is to build an objective, reproducible validation metric which is a composite of individual metrics used to validate a model. The composite metric will be constructed from a combination of individual metrics weighted by their position on the primacy hierarchy and their sensitivity to parameters. This will allow an overall assessment of goodness of validation consistent with our insistence that multiple measures be used, spanning the primacy hierarchy.

Construct an individual Validation metric by:

1) For each individual measure take the normalized measure(B) * normalized value on primacy hierarchy(P) * normalized sensitivity index(S)*repetition weight(W?). (individual metric can have ensemble weighting (perhaps in primacy factor?? Not to exceed 1…should not count each element as a separate metric)

2) Sum the individual weighted metrics

$$M_s = \sum_i B_i * P_i * S_i * W_i \frac{1}{10}$$

< 1 is a poor score

$1 < M_s < 5$ is OK score

> 10 is a good score

3) Divide that sum by the number of elements

$$M_n = \frac{1}{n} \sum_i B_i * P_i * S_i * W_i \frac{1}{10}$$

< 0.3 is a poor score

$0.3 < M_n < 0.7$ is OK score

> 0.7 is good score

The actual final metric is then a vector with $V = (M_s, M_n)$

- normalized measure(B): scale of 0-1 can be a variety of measures, (1-normalized deviation), pass fail measure, normalized Bayes factor, (individual metric can have ensemble weighting (or perhaps in primacy or sensitivity factor?? Not to exceed 1... should not count each element as a separate metric)
- normalized primacy hierarchy (P): scored from 1-5? (5 being lowest on the hierarchy and 1 being highest)
- sensitivity index (S): Rank metric sensitivity and normalize from 1-2
- repetition weight(W): two measures at the same level on the same branch of the Primacy Hierarchy should not get double counted but perhaps should be given a discount weight of 0.5^k where k is the number of measures on that level. For example a cross correlation measure and a cross phase measure might both be fairly low on the primacy hierarchy but the second of the 2 would be weighted with a 0.5 multiplier because it is at the same level.

Note: M_s is not normalized and therefore is not, in principle, bounded. This is intentional as we wish to give higher scores (encourage) for more comparisons.

- Taylor diagrams (Taylor, 2001) provide a way of graphically summarizing how closely a pattern (or a set of patterns) matches observations. The similarity between two patterns is quantified interms of their correlation, their centered root-mean-square difference and the amplitude of their variations (represented by their standard deviations). These diagrams are especially useful in evaluating multiple aspects of complex models or in gauging the relative skill of many different models (e.g., IPCC, 2001). Figure 1 is a sample Taylor diagram which shows how it can be used to summarize the relative skill with which several global climate models simulate the spatial pattern of annual mean precipitation. Statistics for eight models were computed, and a letter was assigned to each model considered. The position of each letter appearing on the plot quantifies how closely that model's simulated precipitation pattern matches observations. Consider model F, for example.

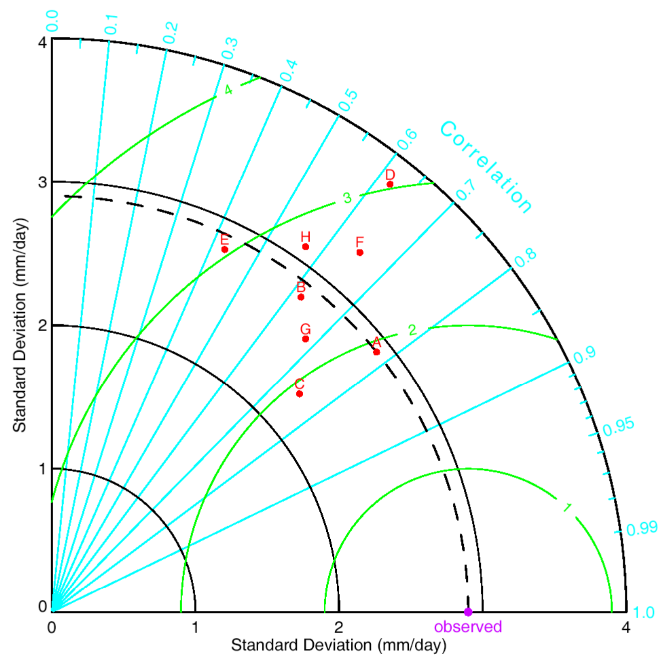
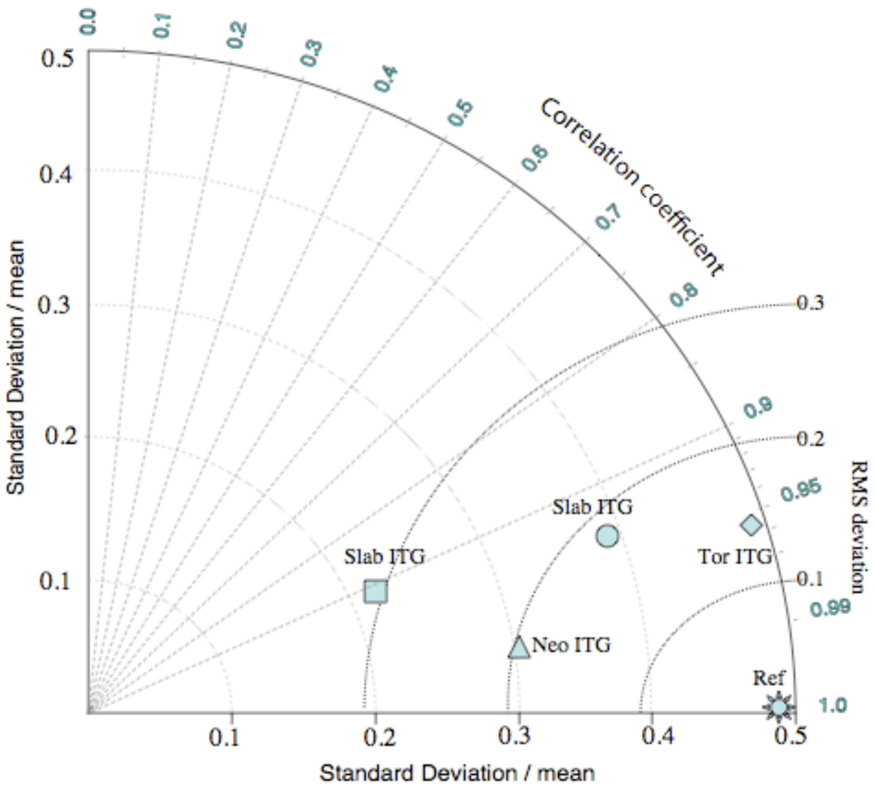


Figure 1: Sample Taylor diagram displaying a statistical comparison with observations of eight model estimates of the global pattern of annual mean precipitation.

Taylor Diagrams

$$E'^2 = \sigma_f^{\ 2} + \sigma_r^{\ 2} - 2\sigma_f \sigma_r R,$$



Validation metric - primacy hierarchy and sensitivity

Primacy hierarchies have ratings associated with primacy levels

Measurement and comparison at multiple levels better than single level

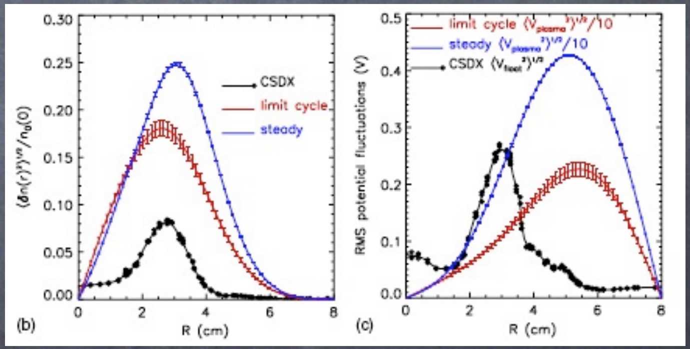
Sensitivity:

- Agreement in quantities with high degree of sensitivity is not rated as favorably as agreement in quantities with low sensitivity
- May be able to use robust predictions to remove sensitivity
  - Examples: χ_i/χ_e , wavenumber spectrum peaks, have low sensitivity
  - But these may remove ability to discriminate between different models
  - Agreement in quantities with poor ability to discriminate is not rated as favorably as agreement in quantities with good ability to discriminate
- Are there robust predictions that also discriminate?
- High sensitivity: large output uncertainties even for validated models within validation domain
- May be possible to beat down sensitivity problem by reducing uncertainty in source parameters

Special experimental conditions can remove complicating factors or probe lower levels of primacy hierarchy

Special experiments

- Simplified geometry/magnetic topology
- Freeze quantities that vary in general
- Parameters in regime of simpler physics
- Fewer disparate effects integrated
- Enhanced diagnostic access



CSDX: linear geometry, controlled turbulence level

Collisional, passing particle drift wave regime

Hasegawa-Wakatani model not optimal for comparison

Comparison with appropriate gyrokinetic model?

Other examples: LAPD, Helimak . . . .

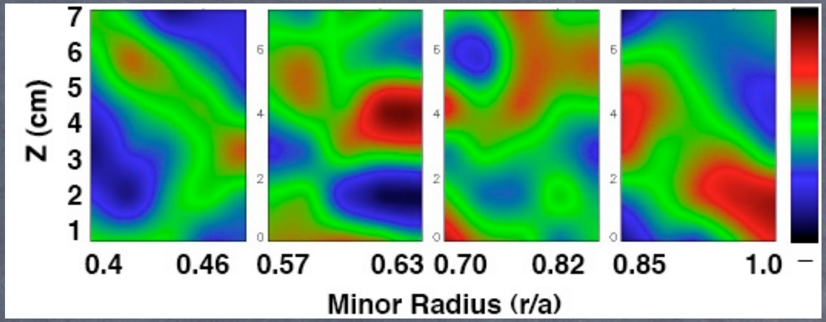
New experiments to propose?

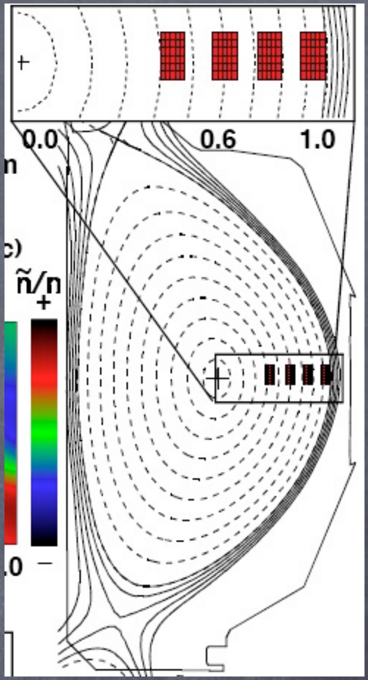
Uses for alternates in model validation?

Enhanced diagnostic capability, special discharges expand comparison possibilities

Examples of payoffs from enhanced capability

- BES sensitivity improvements: fluctuations over wider range of *r/a*





High wavenumber diagnostics: probe electron scale fluctuations

Future development: welcome anything in direction of

- More fluctuating fields
- Bispectra, bicoherence
- Direct sampling of wavenumber

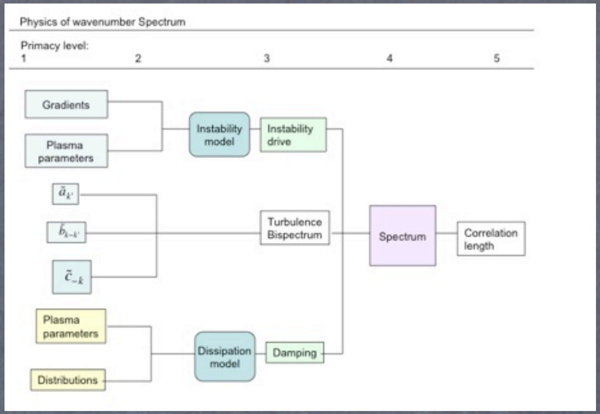
Special discharges: boring for showcasing expt, crucial for verification

- L mode
- Long duration, steady state

Develop, use techniques to undo integration of effects

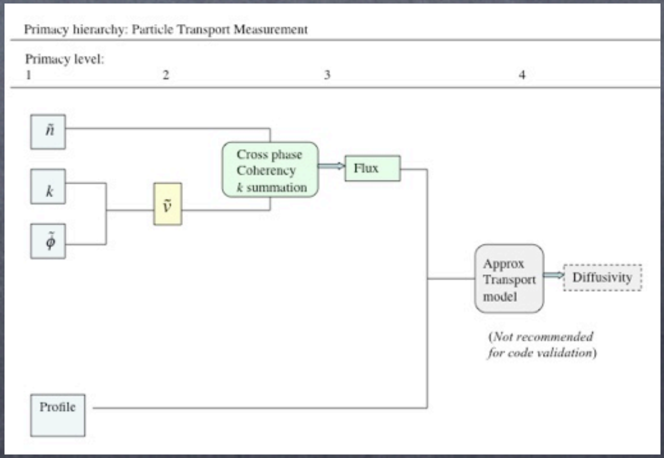
Wavenumber spectrum is poor discriminator between models

Many effects integrate



Measure bispectrum - infer underlying instability drive (bispectral deconvolution)

Diffusivities impose extreme model assumptions



Model fluxes with fractional derivatives

Seek better analysis tools

Change culture of modeling

- Joint activity between modelers, experimentalists, theorists
  - TTF has developed right forum for reporting validation efforts
- Run codes in predictive mode
  - Blind, double blind comparison
- Validation as important scientific activity
  - Pursue independently of code building
  - We are working with journals (editors, referees) to welcome V&V papers
- Open reporting of difficulties, shortcomings in comparisons
  - Remove stigma of reporting imperfect results
  - Skepticism about favorable results: hallmark of good science
  - Don't stop tweaking when agreement obtained (is it really agreement?)

Where we go from here

- Creating guidelines and good practices
  - Initial proposals
  - Feedback
  - Refinement
  - Iteration
- Technical development
  - Robust quantities, sensitivity and discriminating between models
  - Ideas for validation experiments
  - Diagnostic and analysis technique development
  - Do validation with validation metric
- Programmatic opportunities
  - Fusion Simulation Project - impacting way it is set up
  - 5 year planning for major facilities - including validation activities