INSTITUTE FOR DEFENSE ANALYSES

Space-Filling Experimental Design and Surrogate Models for U.S. Department of Defense Modeling and Simulation Evaluation

John T. Haman, Project Leader

Curtis G. Miller

OED Draft

August 2023

IDA Document 1038122

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Task BD-9-2299(90), “Test Science: Methods Advancements” for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Dr. V. Bram Lillard and consisted of Dr. Emily Heuring, Dr. John T. Haman, Dr. Rebecca M. Medlin, Mr. Chris Dimapasok, Dr. Keyla Pagan-Rivera, and Dr. Addison Adams from the Operational Evaluation Division.

For more information:
Dr. John Haman, Project Leader
jhaman@ida.org • (703) 845-2132

Dr. V. Bram Lillard, Director, Operational Evaluation Division
vlillard@ida.org • (703) 845-2230

INSTITUTE FOR DEFENSE ANALYSES

IDA Document 1038122

Space-Filling Experimental Design and Surrogate Models for U.S. Department of Defense Modeling and Simulation Evaluation

John T. Haman, Project Leader

Curtis G. Miller

Executive Summary

Operational testing (OT) provides important information on the capability of equipment acquired by the United States Department of Defense (DoD). Operational testing has been a vital component of the acquisition process since Congress reformed OT conduct with the Department of Defense Authorization Act of 1984. Computer modeling and simulation (M&S) can help address the small sample size problem in live, operational testing, but M&S must be validated in order for stakeholders to believe M&S makes meaningful predictions about real-world outcomes.

The Institute for Defense Analyses (IDA) has published documents providing guidance to the OT community on statistical validation of M&S, including a handbook and papers on the use of space-filling designs and statistical surrogates (also known as metamodels). However, important questions remain that need to be answered in order to best validate M&S.

In particular, operational testers need better recommendations on sample size selection, determining the number of replicates in a design (if any), validating with small real-world sample sizes, and incorporating statistical surrogates into hypothesis tests that help determine whether M&S outcomes match real-world outcomes or not.

We hope the larger statistical community can help contribute answers to these questions. This presentation was given at the Joint Statistical Meeting 2023 in Toronto, Canada.

Space-Filling Experimental Design and Surrogate Models for U.S. Department of Defense Modeling and Simulation Evaluation

Dr. Curtis Miller

August 9, 2023

Institute for Defense Analyses
730 East Glebe Road ● Alexandria, Virginia 22305

We want statisticians to appreciate...

… why operational testing matters

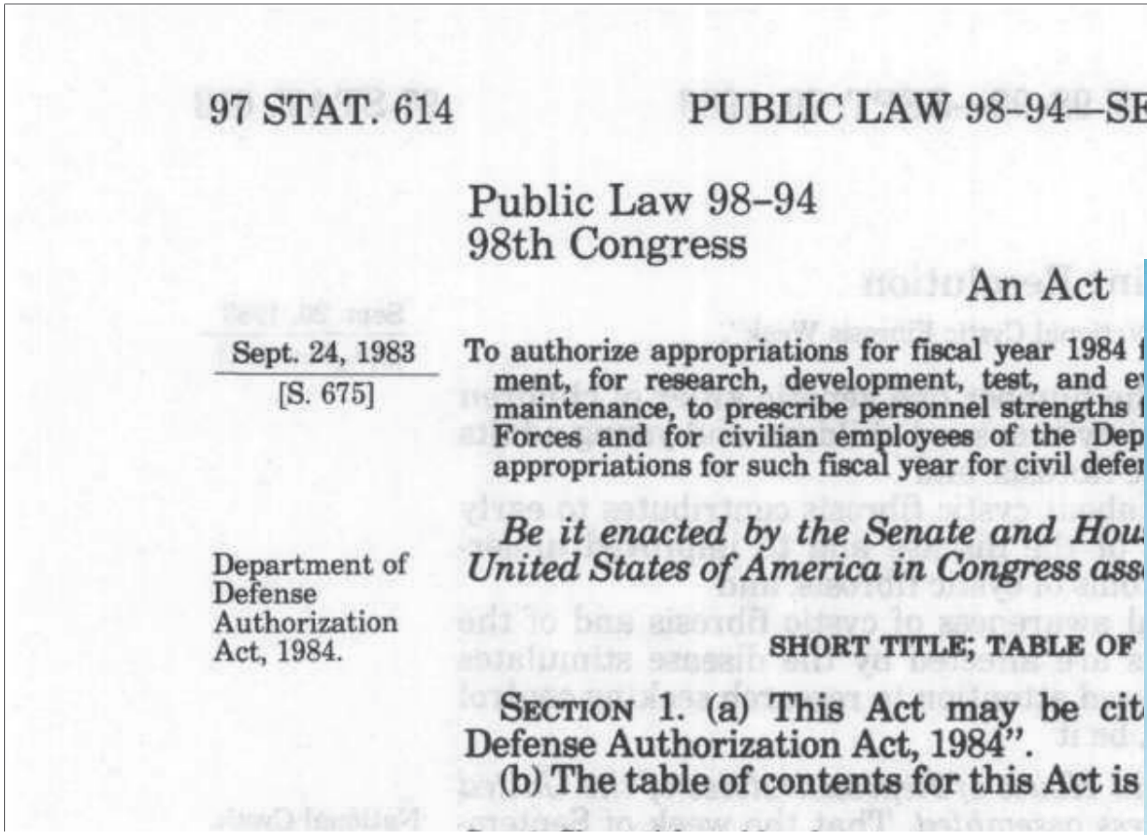
… why computer modeling and simulation matters to operational testing

… the challenges in planning computer modeling and simulation studies

… the challenges in statistically validating a computer model

Why does operational testing matter?

Operational testing studies DOD system effectiveness in warfighting conditions



97 STAT. 614 PUBLIC LAW 98–94—SEPT. 24, 1983

Public Law 98–94
98th Congress

An Act

Sept. 24, 1983
[S. 675]

To authorize appropriations for fiscal year 1984 f
ment, for research, development, test, and e
maintenance, to prescribe personnel strengths f
Forces and for civilian employees of the Dep
appropriations for such fiscal year for civil defe

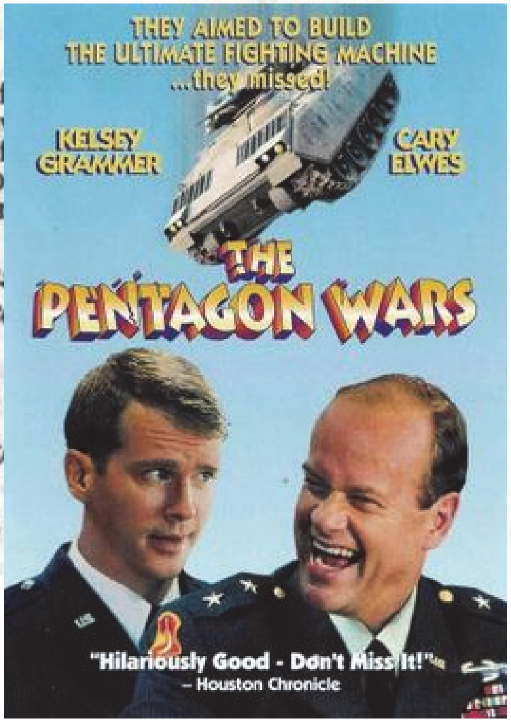
Be it enacted by the Senate and Hou
United States of America in Congress ass

Department of Defense Authorization Act, 1984.

SHORT TITLE; TABLE OF

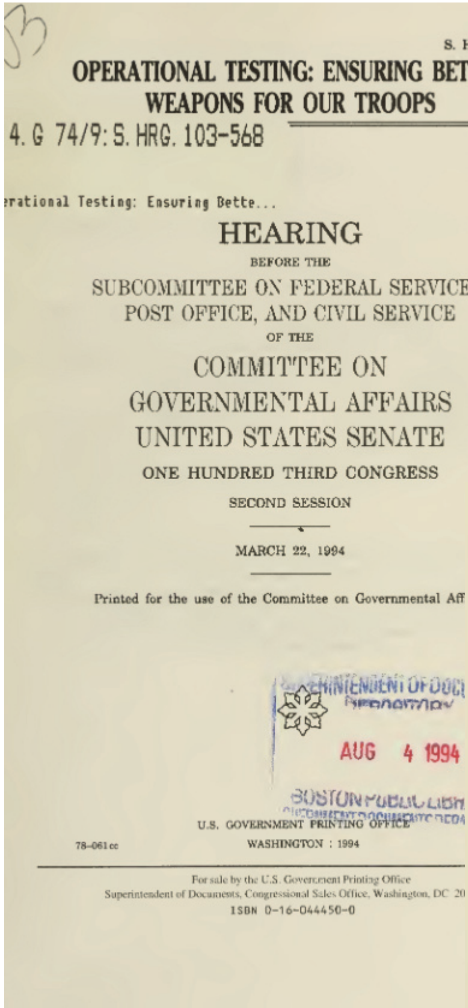
SECTION 1. (a) This Act may be cit
Defense Authorization Act, 1984".

(b) The table of contents for this Act is



DOD – U.S. Department of Defense

OT results matter to both Congress and the warfighter



S. Hrg. 103–568

OPERATIONAL TESTING: ENSURING BETTER WEAPONS FOR OUR TROOPS

HEARING

BEFORE THE

SUBCOMMITTEE ON FEDERAL SERVICES, POST OFFICE, AND CIVIL SERVICE

OF THE

COMMITTEE ON GOVERNMENTAL AFFAIRS UNITED STATES SENATE

ONE HUNDRED THIRD CONGRESS

SECOND SESSION

MARCH 22, 1994

Printed for the use of the Committee on Governmental Aff

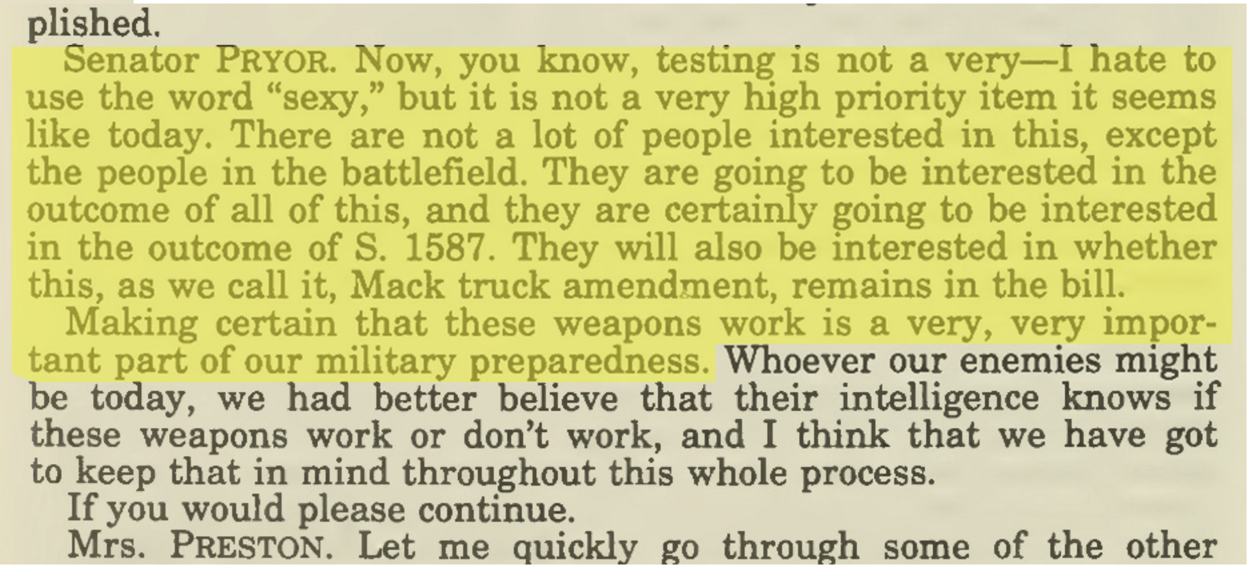
78-061 cc

U.S. GOVERNMENT PRINTING OFFICE
WASHINGTON : 1994

For sale by the U.S. Government Printing Office
Superintendent of Documents, Congressional Sales Office, Washington, DC 20
ISBN 0-16-044450-0

Sen. D. Pryor (D-AK)

Sen. W. Roth (R-DE)



plished.

Senator PRYOR. Now, you know, testing is not a very—I hate to use the word "sexy," but it is not a very high priority item it seems like today. There are not a lot of people interested in this, except the people in the battlefield. They are going to be interested in the outcome of all of this, and they are certainly going to be interested in the outcome of S. 1587. They will also be interested in whether this, as we call it, Mack truck amendment, remains in the bill.

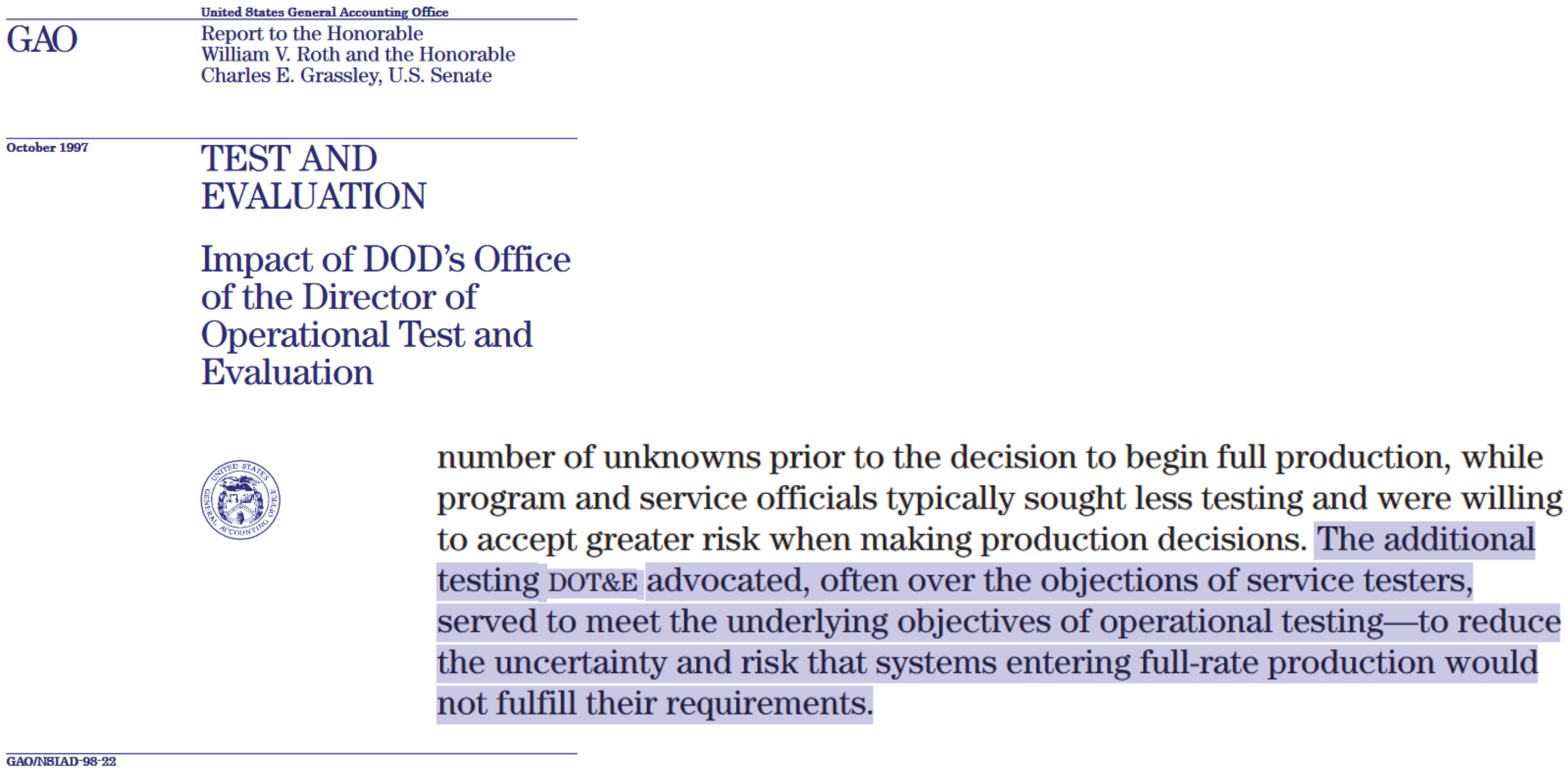
Making certain that these weapons work is a very, very important part of our military preparedness. Whoever our enemies might be today, we had better believe that their intelligence knows if these weapons work or don't work, and I think that we have got to keep that in mind throughout this whole process.

If you would please continue.

Mrs. PRESTON. Let me quickly go through some of the other

OT – Operational Testing

OT&E should reduce risk and uncertainty regarding the performance of systems in wartime



United States General Accounting Office

GAO

Report to the Honorable
William V. Roth and the Honorable
Charles E. Grassley, U.S. Senate

October 1997

TEST AND EVALUATION

Impact of DOD's Office of the Director of Operational Test and Evaluation

number of unknowns prior to the decision to begin full production, while program and service officials typically sought less testing and were willing to accept greater risk when making production decisions. The additional testing DOT&E advocated, often over the objections of service testers, served to meet the underlying objectives of operational testing—to reduce the uncertainty and risk that systems entering full-rate production would not fulfill their requirements.

GAO/NSIAD-98-22

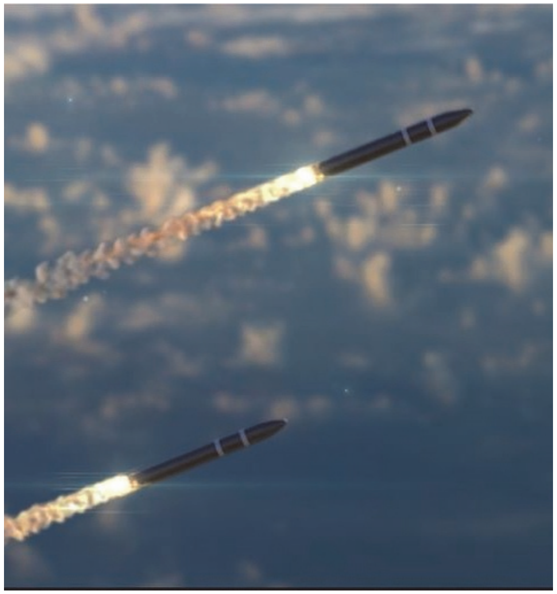
OT&E – Operational Test and Evaluation

Why does computer modeling and simulation matter to OT?

Small data problems are alive and well in operational testing

SM-3 Blk IIA

Next Generation Interceptor



PAC-3

Missile tests can cost $10Ms to $100Ms <u>per shot</u>

https://missiledefenseadvocacy.org/missile-defense-systems-2/missile-defense-systems/missile-interceptors-by-cost/

Small data problems are alive and well in operational testing

SM-3 Blk IIA

Next Generation Interceptor

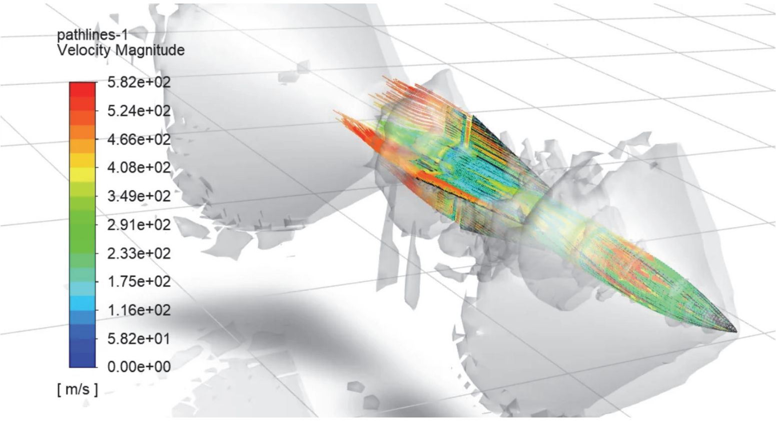
PAC-3

COST AND ASSET AVAILABILITY GENERATE DEMAND FOR M&S

Missile tests can cost $10Ms to $100Ms <u>per shot</u>

https://missiledefenseadvocacy.org/missile-defense-systems-2/missile-defense-systems/missile-interceptors-by-cost/

Modeling and simulation comes in a variety of flavors with unique statistical considerations

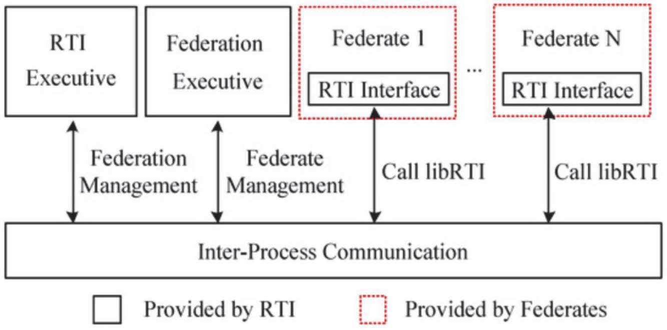


Digital Simulation

https://www.reddit.com/r/dcsworld/comments/mqmvy3/lowquality_steadystate_cfd_simulation_of_an_aim54

Hardware in the Loop

https://testscience.org

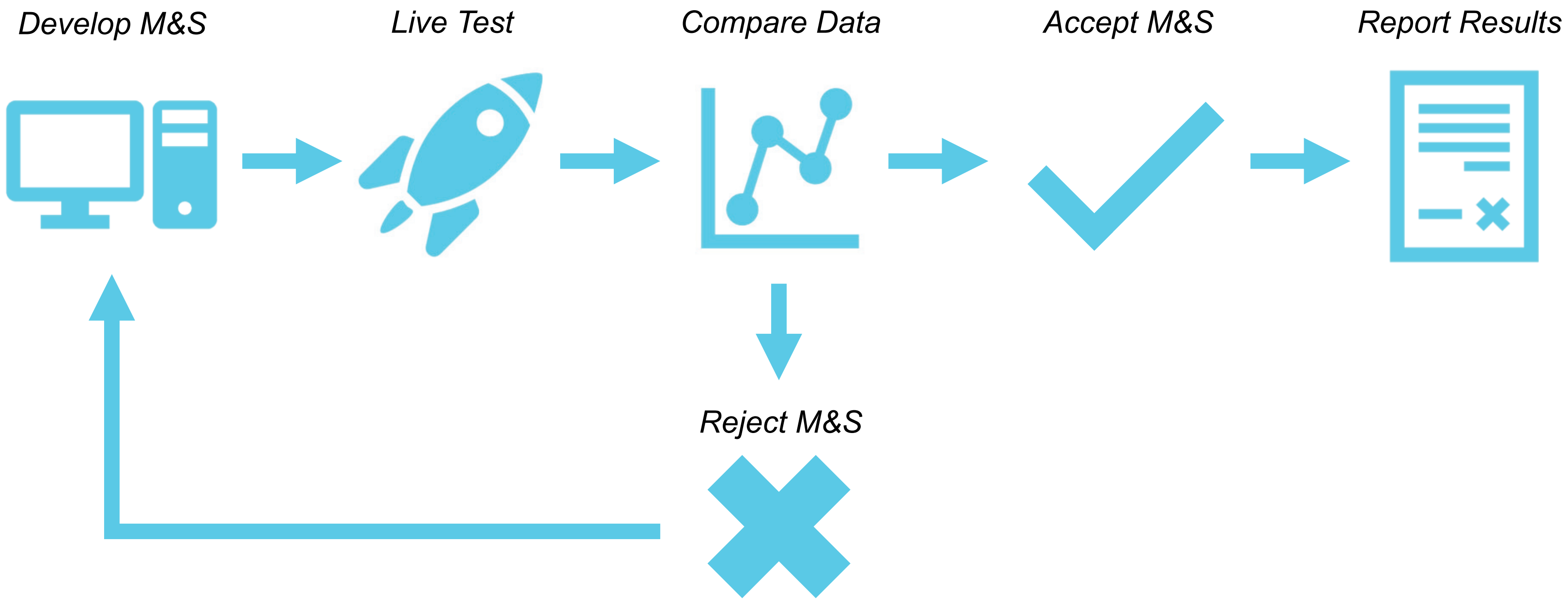


M&S – Modeling and Simulation

Federation

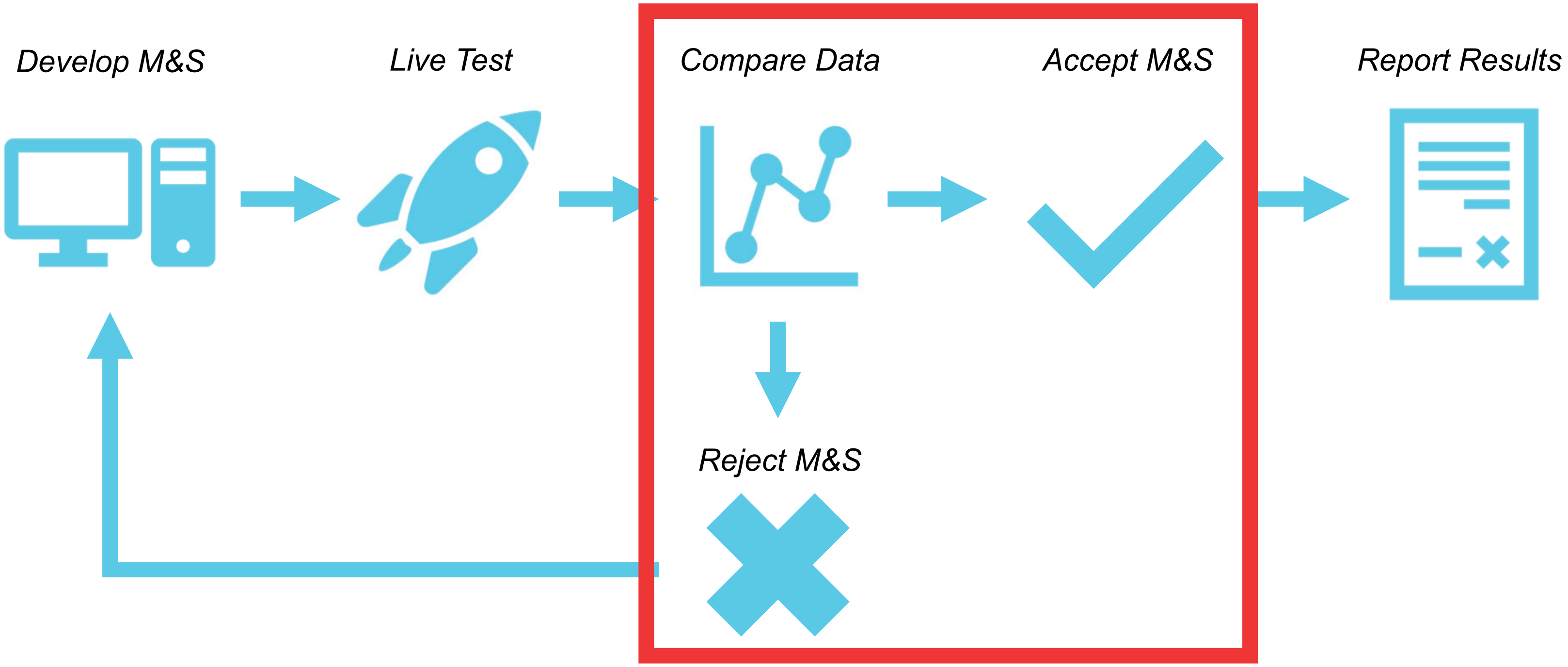
https://doi.org/10.3390/electronics9030540

If M&S outputs will be used for predicting OT results, we must compare M&S outputs to live test data



M&S – Modeling and Simulation; OT – Operational Testing

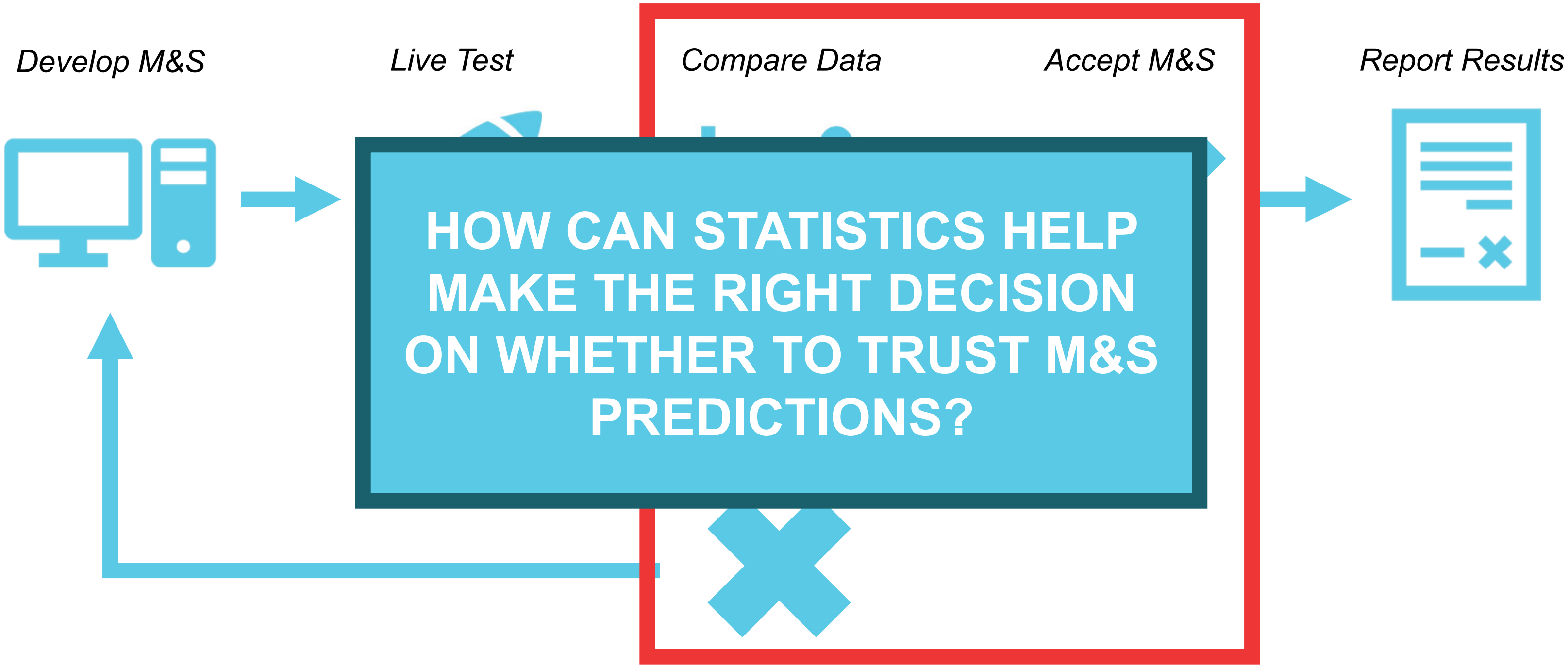
If M&S outputs will be used for predicting OT results, we must compare M&S outputs to live test data



THIS STEP IS STATISTICAL

M&S – Modeling and Simulation; OT – Operational Testing

If M&S outputs will be used for predicting OT results, we must compare M&S outputs to live test data



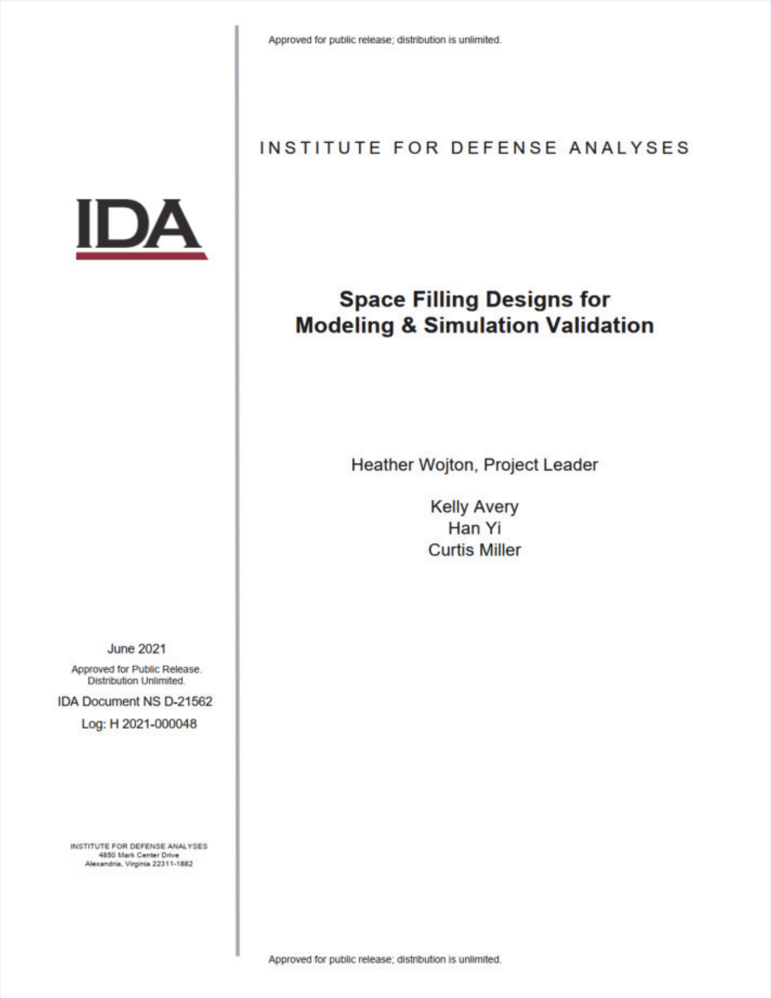
HOW CAN STATISTICS HELP MAKE THE RIGHT DECISION ON WHETHER TO TRUST M&S PREDICTIONS?

THIS STEP IS STATISTICAL

M&S – Modeling and Simulation; OT – Operational Testing

What are the challenges in planning computer M&S studies?

IDA publications introduce and recommend M&S DOE best practices



Approved for public release; distribution is unlimited.

INSTITUTE FOR DEFENSE ANALYSES

Space Filling Designs for
Modeling & Simulation Validation

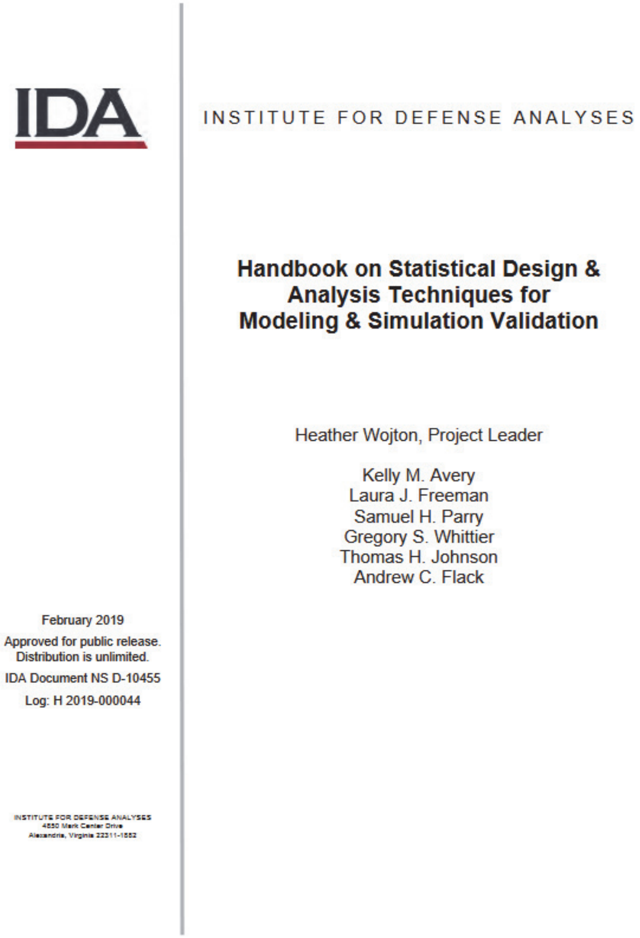
Heather Wojton, Project Leader

Kelly Avery
Han Yi
Curtis Miller

June 2021
Approved for Public Release.
Distribution Unlimited.
IDA Document NS D-21562
Log: H 2021-000048

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

Approved for public release; distribution is unlimited.



INSTITUTE FOR DEFENSE ANALYSES

Handbook on Statistical Design &
Analysis Techniques for
Modeling & Simulation Validation

Heather Wojton, Project Leader

Kelly M. Avery
Laura J. Freeman
Samuel H. Parry
Gregory S. Whittier
Thomas H. Johnson
Andrew C. Flack

February 2019
Approved for public release.
Distribution is unlimited.
IDA Document NS D-10455
Log: H 2019-000044

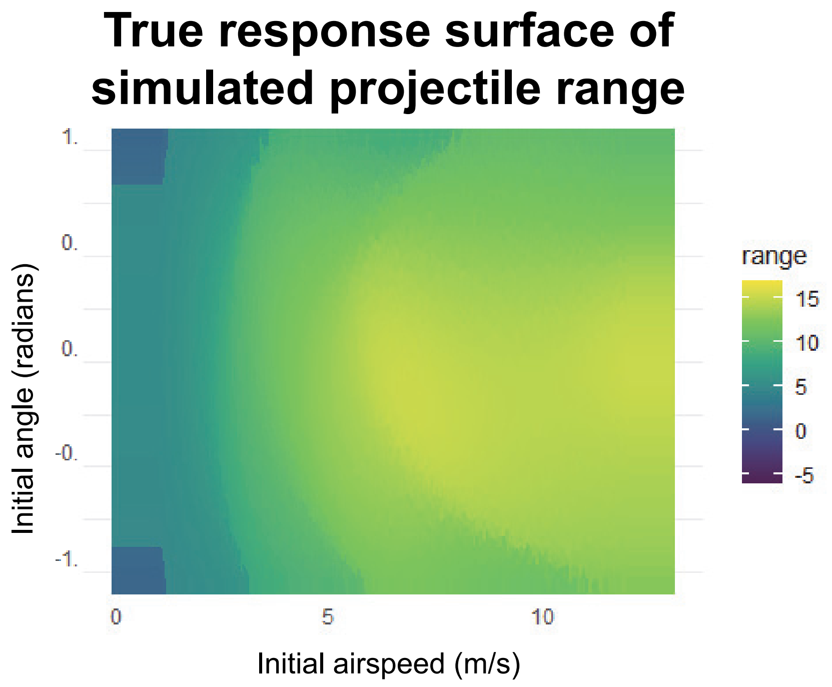
INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

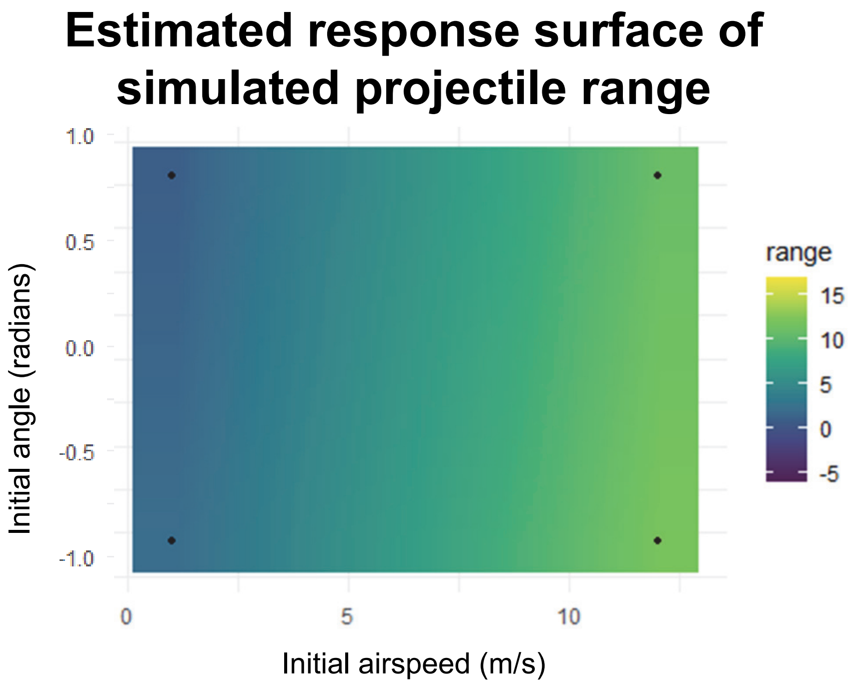
DOE – Design of experiments; IDA – Institute for Defense Analyses; SFD – Space-Filling Design

https://testscience.org/wp-content/uploads/formidable/20/SFD_Literature_Review_Final.html

Space-filling design of experiments helps recover more trends in simulation outputs

A factorial design with a simple linear model fitted will not accurately describe the M&S system's behavior.

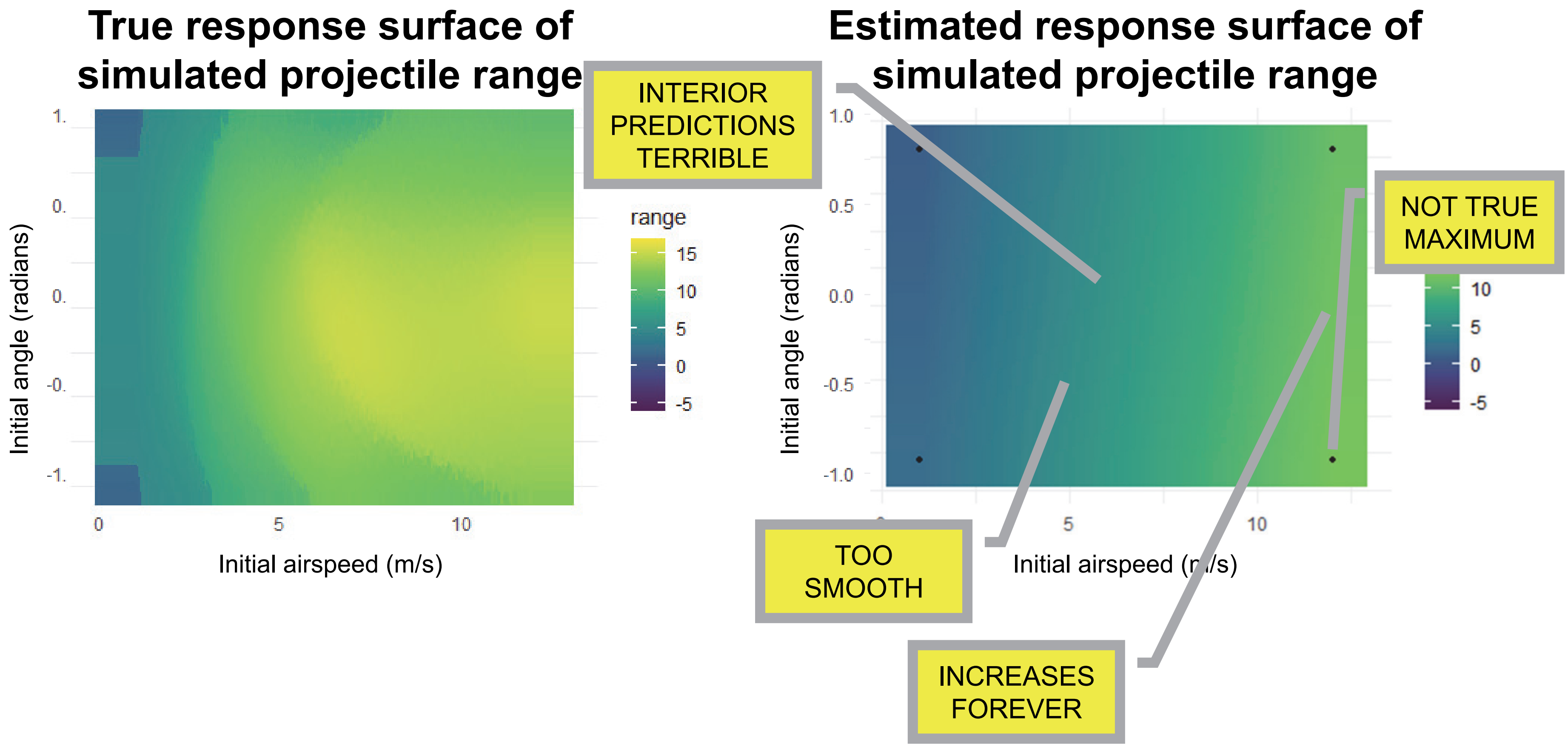

True response surface of simulated projectile range


Estimated response surface of simulated projectile range

M&S – Modeling and Simulation

Space-filling design of experiments helps recover more trends in simulation outputs

A factorial design with a simple linear model fitted will not accurately describe the M&S system's behavior.



True response surface of simulated projectile range

Estimated response surface of simulated projectile range

INTERIOR PREDICTIONS TERRIBLE

NOT TRUE MAXIMUM

TOO SMOOTH

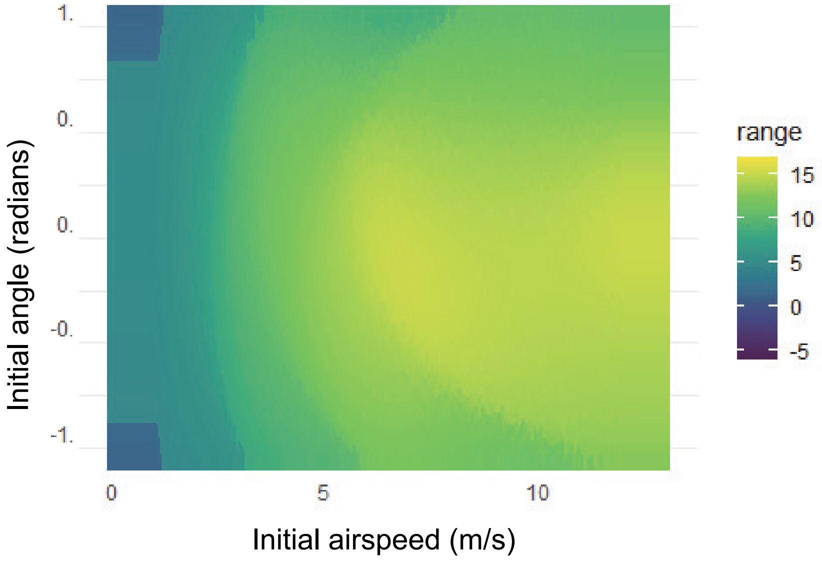
INCREASES FOREVER

M&S – Modeling and Simulation

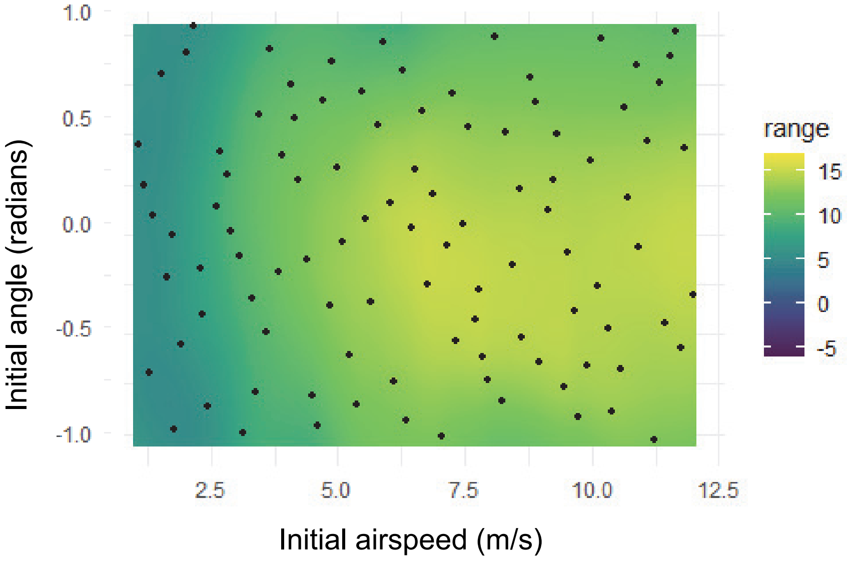
Space-filling design of experiments helps recover more trends in simulation outputs

Analyzing the flights with a Gaussian Process model via a Space-Filling Design yields a good approximation to simulation output.

True response surface of simulated projectile range

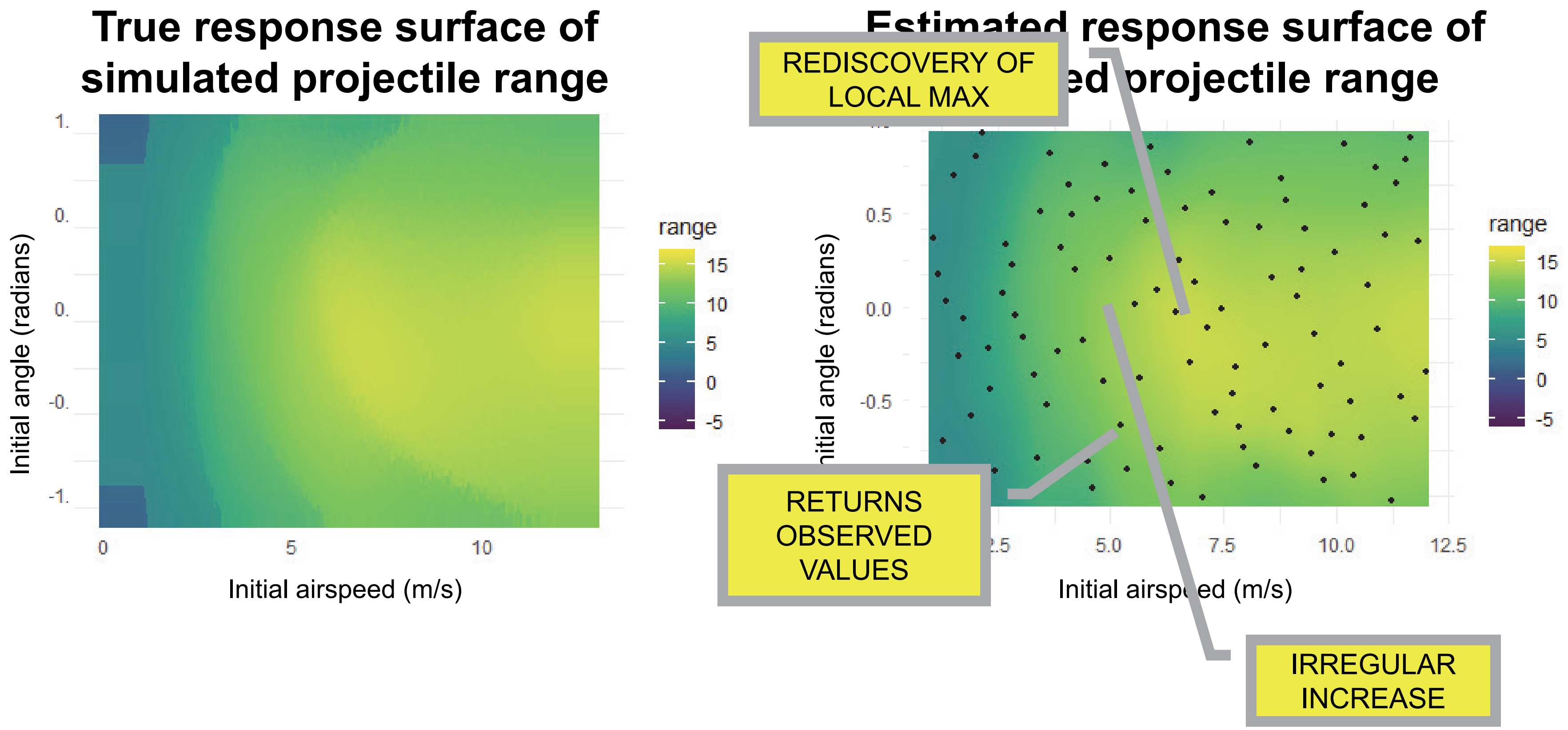


Estimated response surface of simulated projectile range



Space-filling design of experiments helps recover more trends in simulation outputs

Analyzing the flights with a Gaussian Process model via a Space-Filling Design yields a good approximation to simulation output.



True response surface of simulated projectile range

Estimated response surface of simulated projectile range

REDISCOVERY OF LOCAL MAX

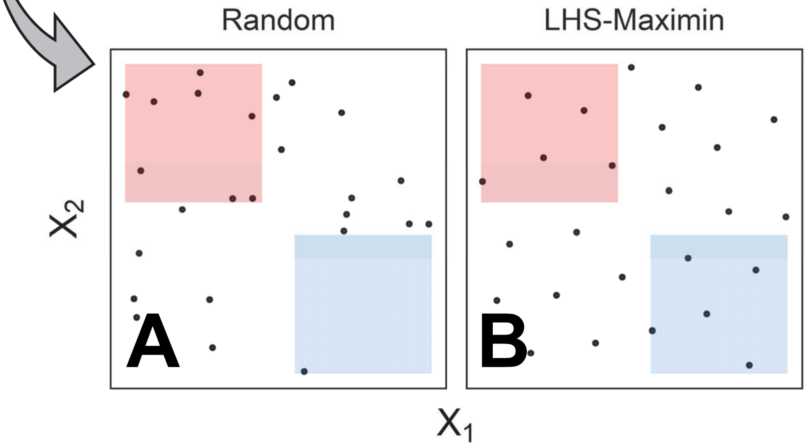
RETURNS OBSERVED VALUES

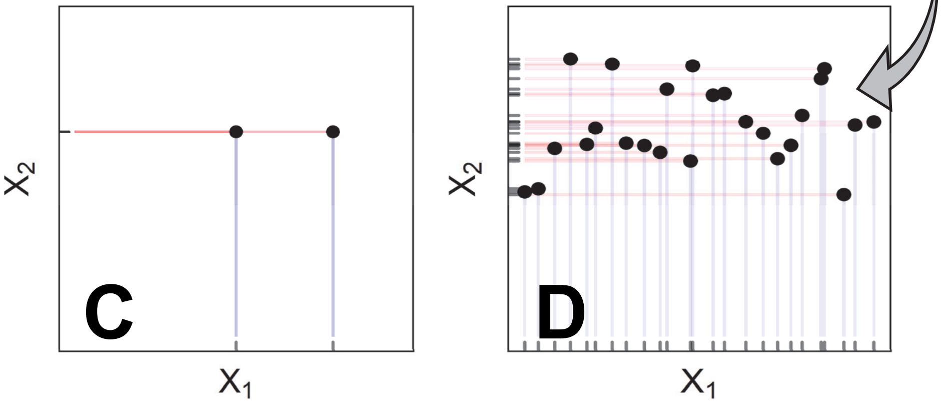
IRREGULAR INCREASE

Just like with classical DOE, there are quantitative ways to evaluate a specific design

Many criteria exist, but it is particularly important that an SFD satisfy the following three criteria in order to be useful:

- Point-distance: Samples are placed as far apart from each other as possible. [Maximin]
- Uniformity: All regions of the design space are equally well-represented. [Center L^2 Discrepancy]
- Projection: The design is robust to variables being collapsed. [MaxPro]

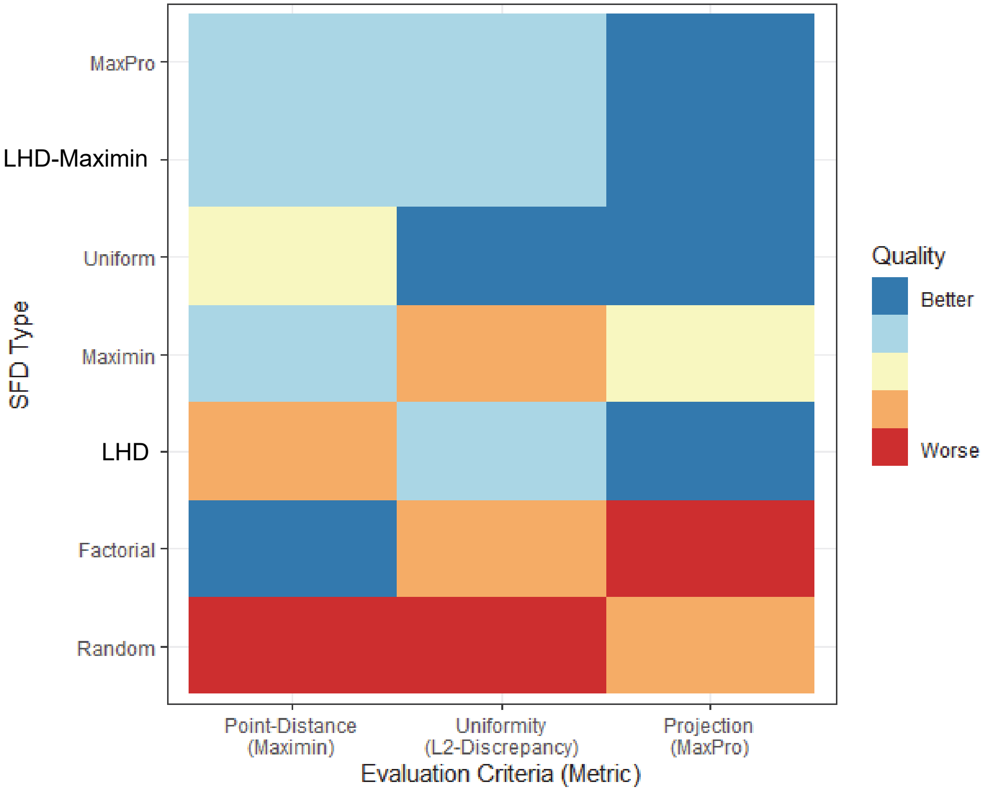


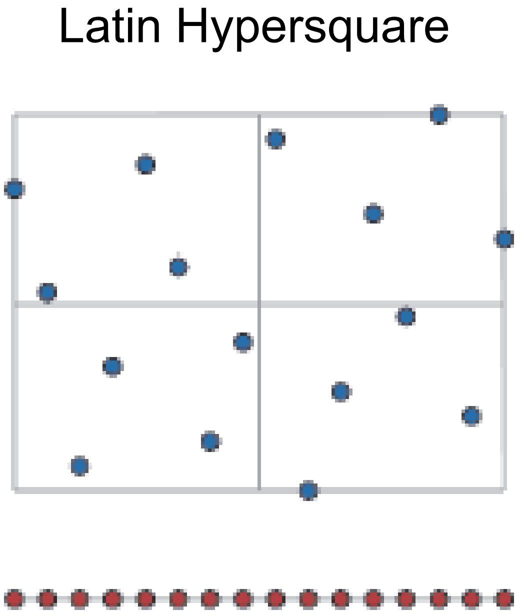


DOE – Design of Experiments; SFD – Space-Filling Design

I prefer maximin sliced Latin hypersquare designs (Maximin SLHD) and MaxPro SFDs

<u>General recommendations</u>:
Maximin (Sliced) LHD or MaxPro





SFD – Space-Filling Design; SLHD – Sliced Latin Hypersquare Design

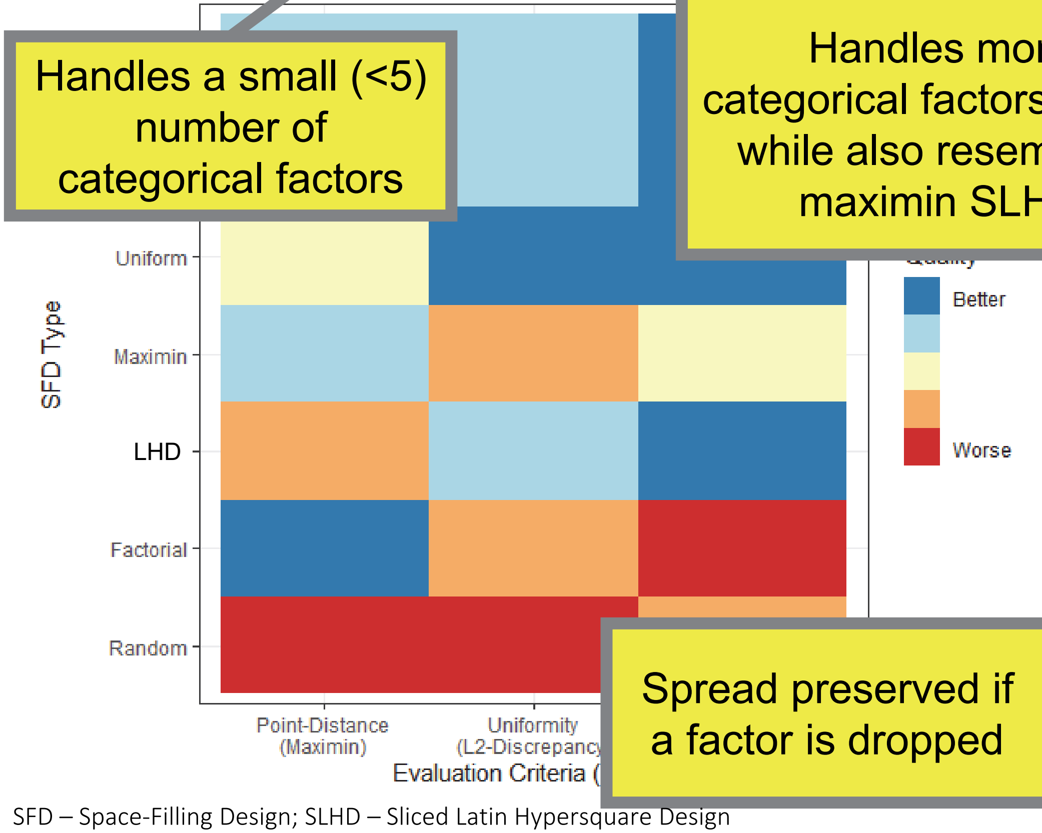
I prefer maximin sliced Latin hypersquare designs (Maximin SLHD) and MaxPro SFDs

General recommendations:

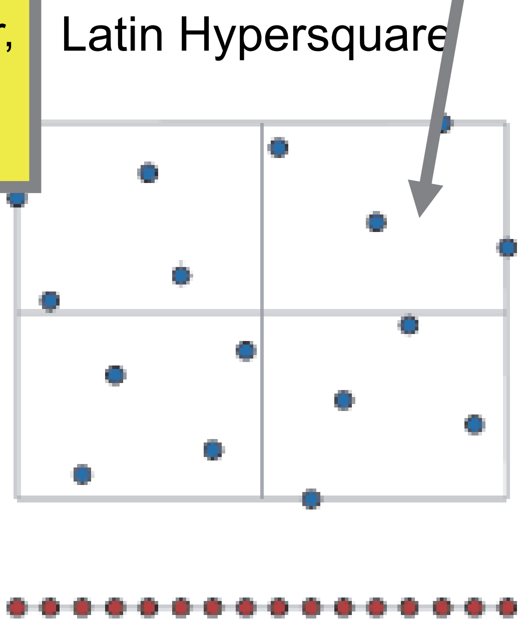
Maximin (Sliced) LHD or MaxPro

Handles a small (<5) number of categorical factors

Handles more categorical factors better, while also resembling maximin SLHD



Good spread in the design space

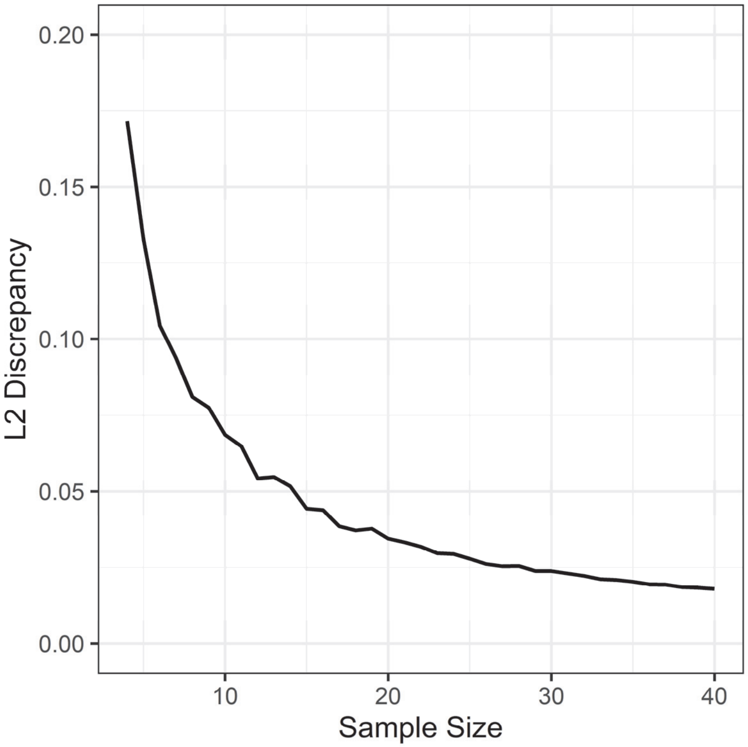


Spread preserved if a factor is dropped

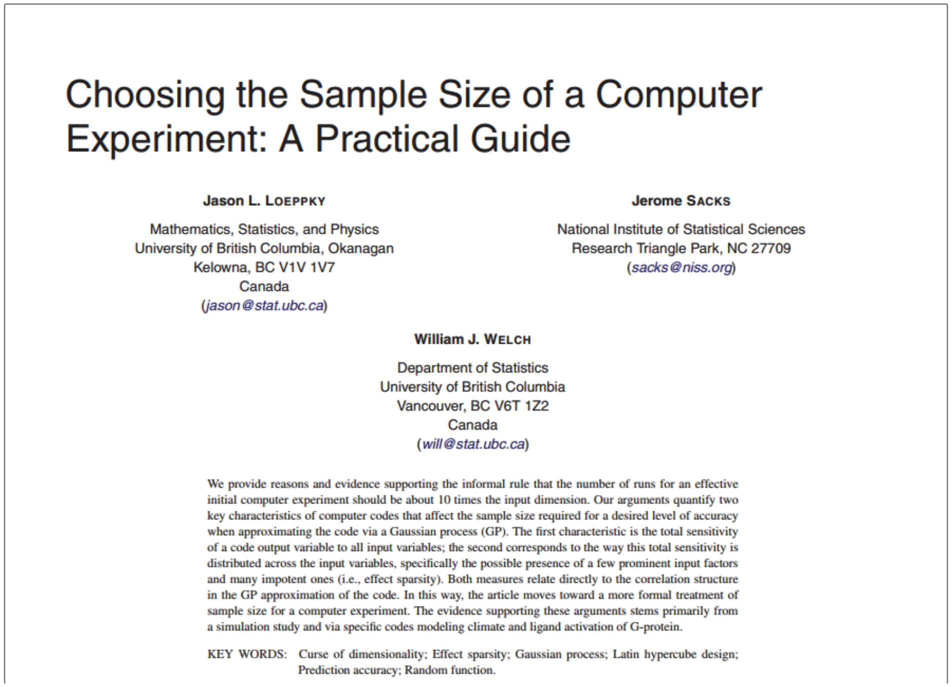
SFD – Space-Filling Design; SLHD – Sliced Latin Hypersquare Design

How should we choose the sample size when generating an SFD?

SCREE PLOT



$$n = 10d$$



Choosing the Sample Size of a Computer Experiment: A Practical Guide

Jason L. Loeppky

Mathematics, Statistics, and Physics
University of British Columbia, Okanagan
Kelowna, BC V1V 1V7
Canada
(*jason@stat.ubc.ca*)

Jerome Sacks

National Institute of Statistical Sciences
Research Triangle Park, NC 27709
(*sacks@niss.org*)

William J. Welch

Department of Statistics
University of British Columbia
Vancouver, BC V6T 1Z2
Canada
(*will@stat.ubc.ca*)

We provide reasons and evidence supporting the informal rule that the number of runs for an effective initial computer experiment should be about 10 times the input dimension. Our arguments quantify two key characteristics of computer codes that affect the sample size required for a desired level of accuracy when approximating the code via a Gaussian process (GP). The first characteristic is the total sensitivity of a code output variable to all input variables; the second corresponds to the way this total sensitivity is distributed across the input variables, specifically the possible presence of a few prominent input factors and many impotent ones (i.e., effect sparsity). Both measures relate directly to the correlation structure in the GP approximation of the code. In this way, the article moves toward a more formal treatment of sample size for a computer experiment. The evidence supporting these arguments stems primarily from a simulation study and via specific codes modeling climate and ligand activation of G-protein.

KEY WORDS: Curse of dimensionality; Effect sparsity; Gaussian process; Latin hypercube design; Prediction accuracy; Random function.

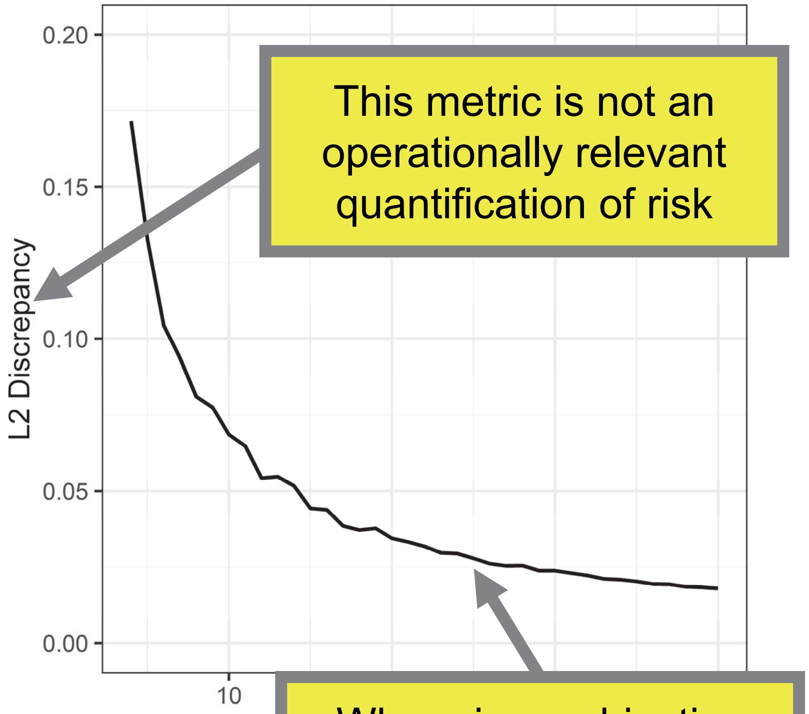
SFD – Space-Filling Design

How should we choose the sample size when generating an SFD?

Seems too small in stochastic cases

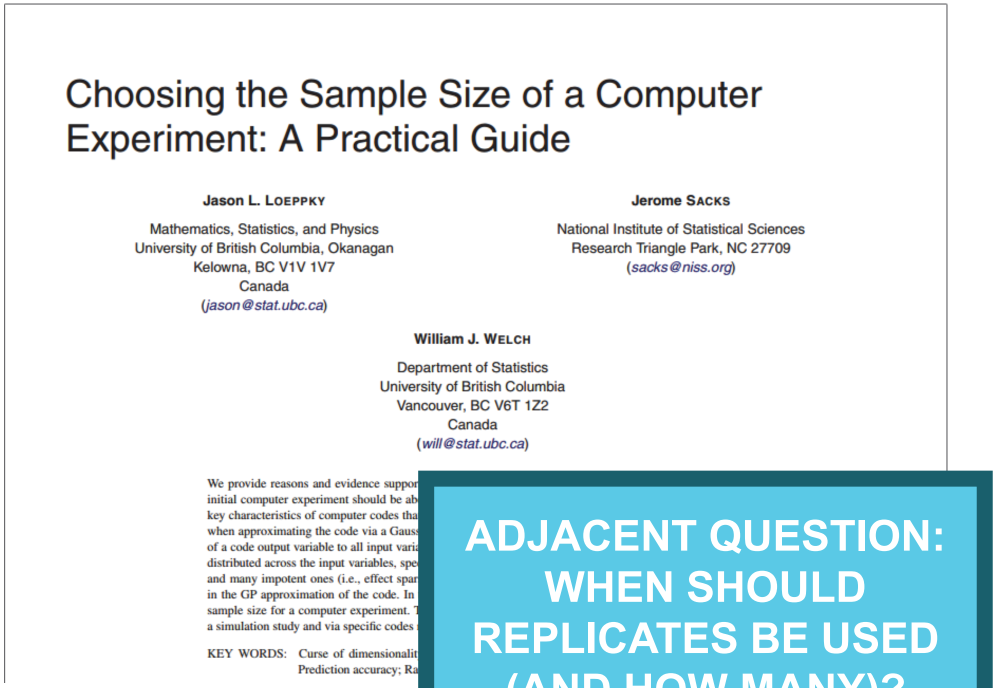
$$n = 10d$$

SCREE PLOT



This metric is not an operationally relevant quantification of risk

Where is an objective cut-off?



Choosing the Sample Size of a Computer Experiment: A Practical Guide

Jason L. Loeppky
Mathematics, Statistics, and Physics
University of British Columbia, Okanagan
Kelowna, BC V1V 1V7
Canada
(jason@stat.ubc.ca)

Jerome Sacks
National Institute of Statistical Sciences
Research Triangle Park, NC 27709
(sacks@niss.org)

William J. Welch
Department of Statistics
University of British Columbia
Vancouver, BC V6T 1Z2
Canada
(will@stat.ubc.ca)

We provide reasons and evidence suppor
initial computer experiment should be ab
key characteristics of computer codes tha
when approximating the code via a Gauss
of a code output variable to all input varia
distributed across the input variables, spe
and many impotent ones (i.e., effect spar
in the GP approximation of the code. In
sample size for a computer experiment. T
a simulation study and via specific codes

KEY WORDS: Curse of dimensionalit
Prediction accuracy; Ra

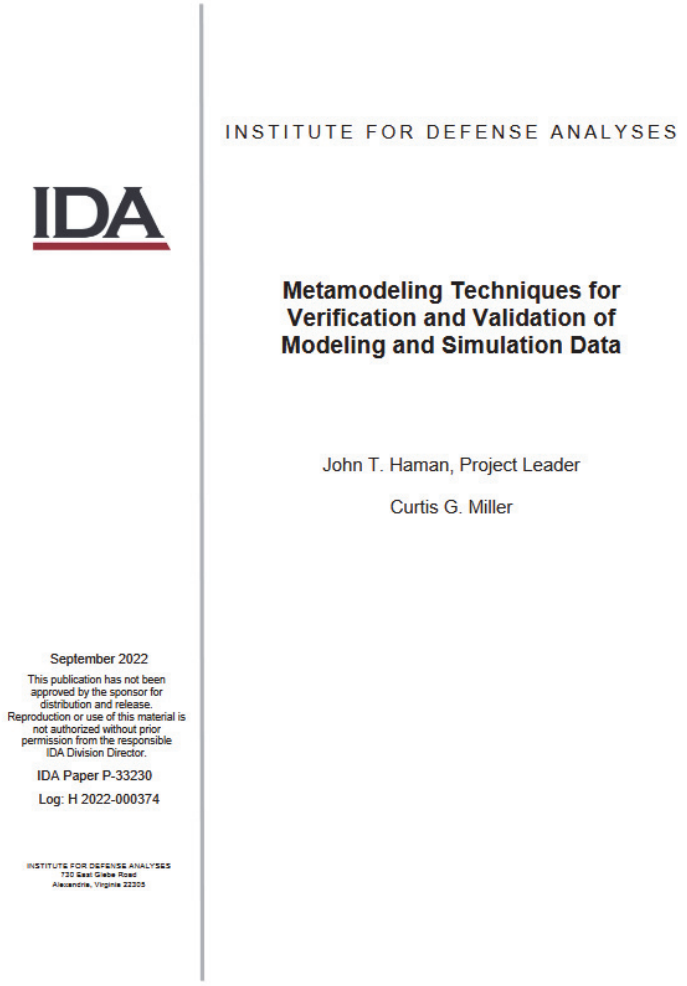
ADJACENT QUESTION: WHEN SHOULD REPLICATES BE USED (AND HOW MANY)?

The minimum may be all you get

SFD – Space-Filling Design

What are the challenges in statistically validating a computer model?

IDA publications and presentations introduce and recommend surrogate modeling best practices



INSTITUTE FOR DEFENSE ANALYSES

Metamodeling Techniques for Verification and Validation of Modeling and Simulation Data

John T. Haman, Project Leader

Curtis G. Miller

September 2022

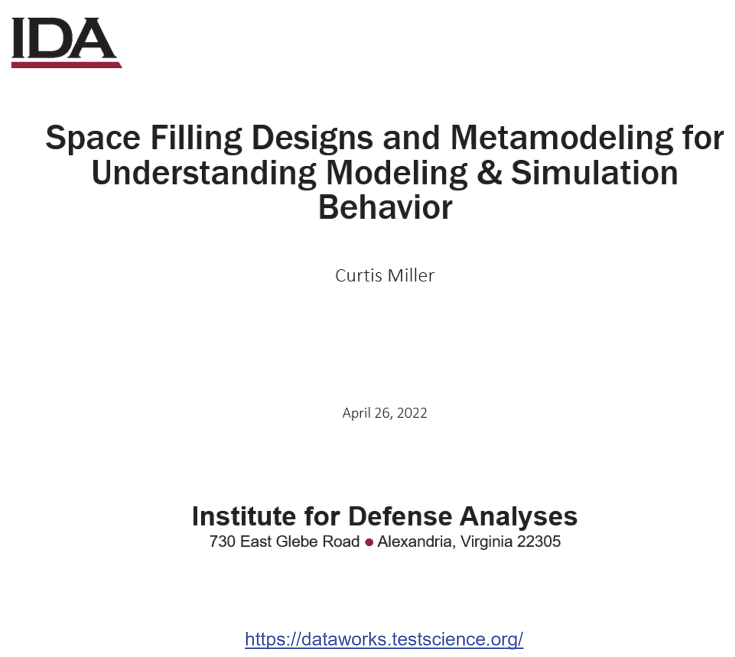
This publication has not been approved by the sponsor for distribution and release. Reproduction or use of this material is not authorized without prior permission from the responsible IDA Division Director.

IDA Paper P-33230

Log: H 2022-000374

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305

https://testscience.org



Space Filling Designs and Metamodeling for Understanding Modeling & Simulation Behavior

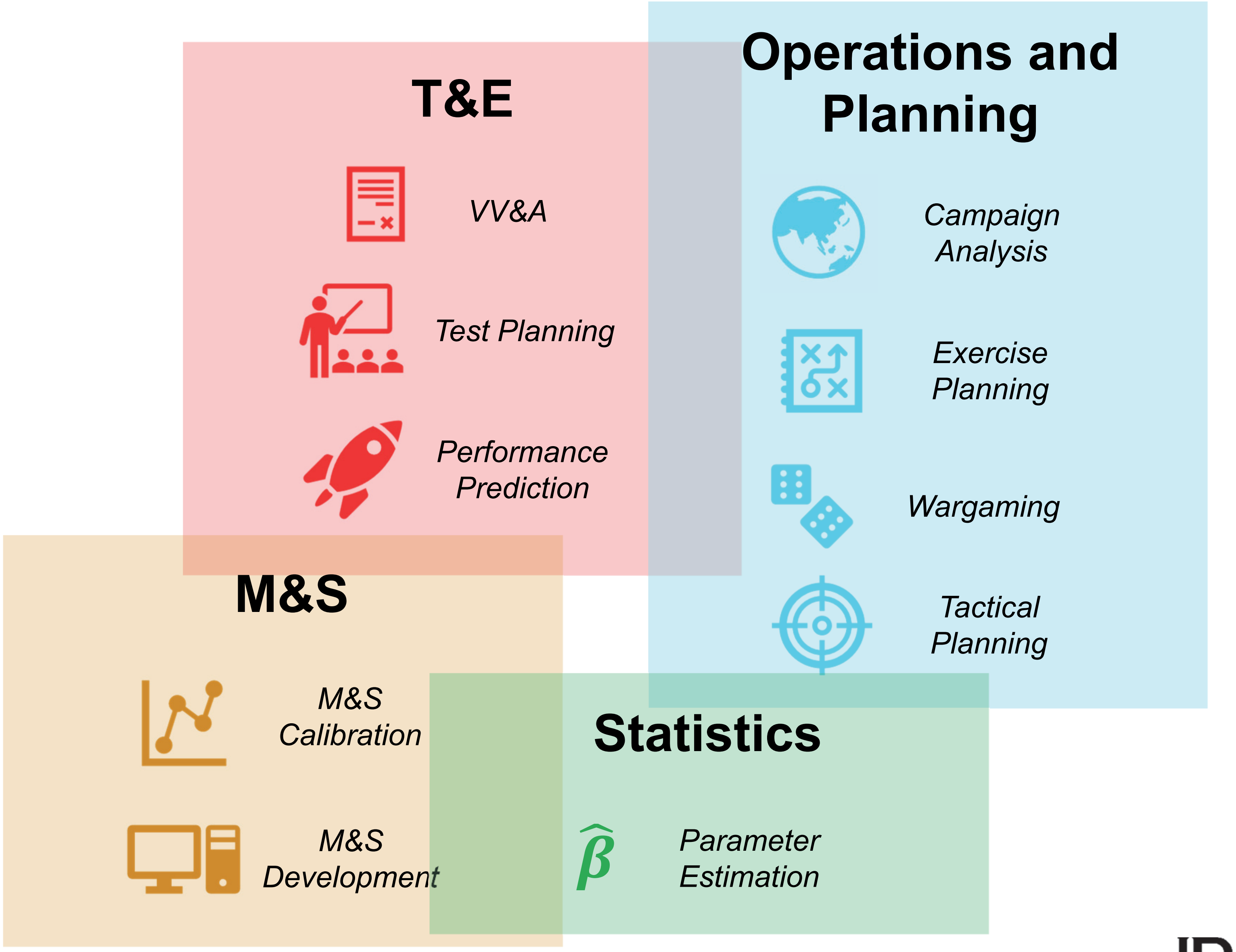
Curtis Miller

April 26, 2022

Institute for Defense Analyses
730 East Glebe Road ● Alexandria, Virginia 22305

https://dataworks.testscience.org/

A statistical surrogate is a useful product in and of itself



T&E

VV&A

Test Planning

Performance Prediction

Operations and Planning

Campaign Analysis

Exercise Planning

Wargaming

Tactical Planning

M&S

M&S Calibration

M&S Development

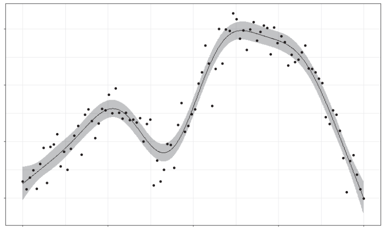
Statistics

$\hat{\beta}$ Parameter Estimation

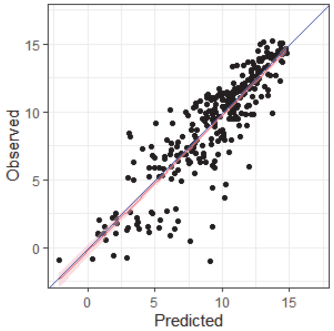
M&S – Modeling and Simulation; T&E – Test and Evaluation; VV&A – Verification, Validation, and Accreditation

Statistical surrogates must allow for relevant assessments of M&S prediction quality

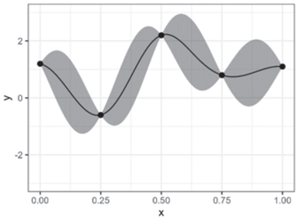
Statistical surrogates should…



… discover unanticipated trends between factors and response variables



… predict observed responses well

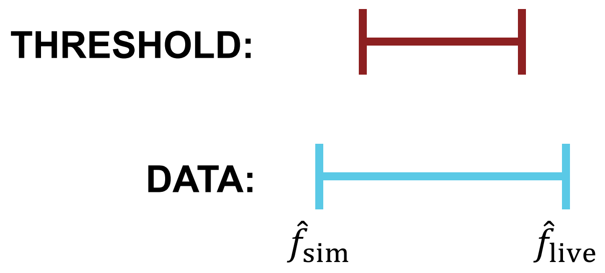


… quantify uncertainty in M&S outputs

M&S – Modeling and Simulation

Statistical surrogates must allow for relevant assessments of M&S prediction quality

Statistical surrogates should…



… allow comparing real-world outcomes to M&S predictions

… allow expert judgement on prediction quality

M&S – Modeling and Simulation

We recommend different statistical surrogate procedures based on the M&S output

	CONTINUOUS OUTPUT	DISCRETE OUTPUT
DETERMINISTIC	GAUSSIAN PROCESS (GP)	NEAREST NEIGHBOR (NN) DECISION TREE
STOCHASTIC	GENERALIZED ADDITIVE MODEL (GAM)	

M&S – Modeling and Simulation

We demonstrate methods using a simple numerical ODE solver

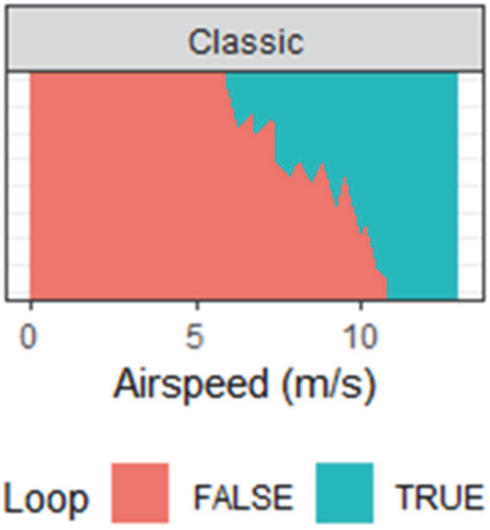
$$\dot{V} = -C_D\left(\rho V^2/2\right)S/m - g\sin(\gamma)$$

$$\dot{\gamma} = \left(C_L\left(\rho V^2/2\right)S/m - g\cos(\gamma)\right)/V$$

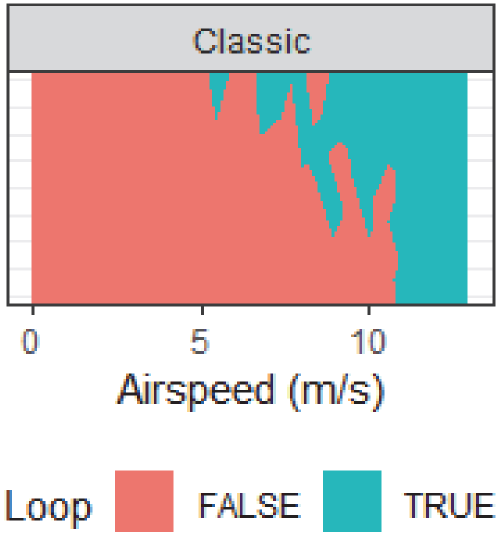
$$\dot{h} = V\sin(\gamma)$$

$$\dot{r} = V\cos(\gamma)$$

Randomize initial conditions for random output (i.e., variance in toss)



DETERMINISTIC

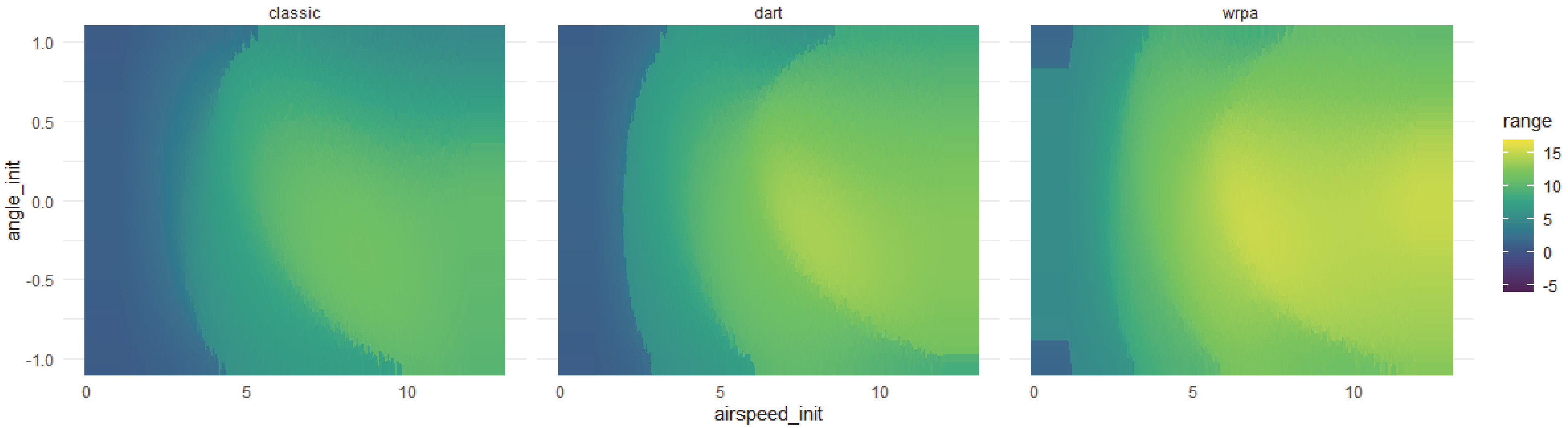


STOCHASTIC

ODE – Ordinary Differential Equation

Model, Airspeed, and Angle Study

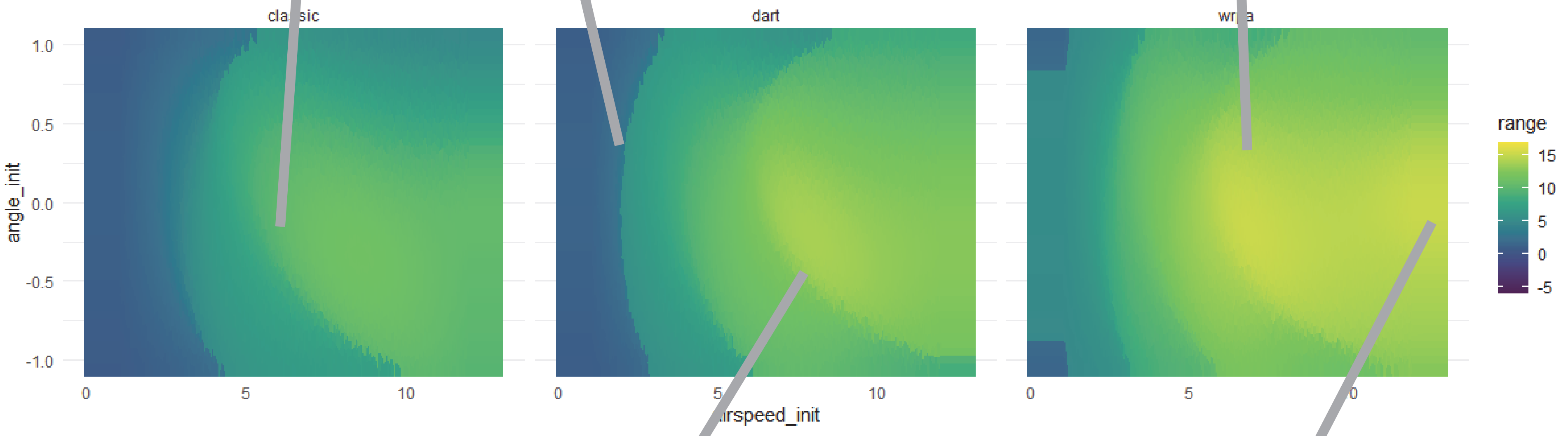
Consider a study of the effects of initial airspeed and angle of flight (continuous factors) on flight terminal range (continuous), while also looking for differences in models (a categorical factor).



WRPA – World Record Paper Airplane

Model, Airspeed, and Angle Study

Consider a study of the effects of initial airspeed and angle of flight (continuous factors) on flight terminal range (continuous), while also looking for differences in models (a categorical factor).



NONLINEAR INCREASE

NUMERICAL ARTIFACTS

GLOBAL MAXIMUM

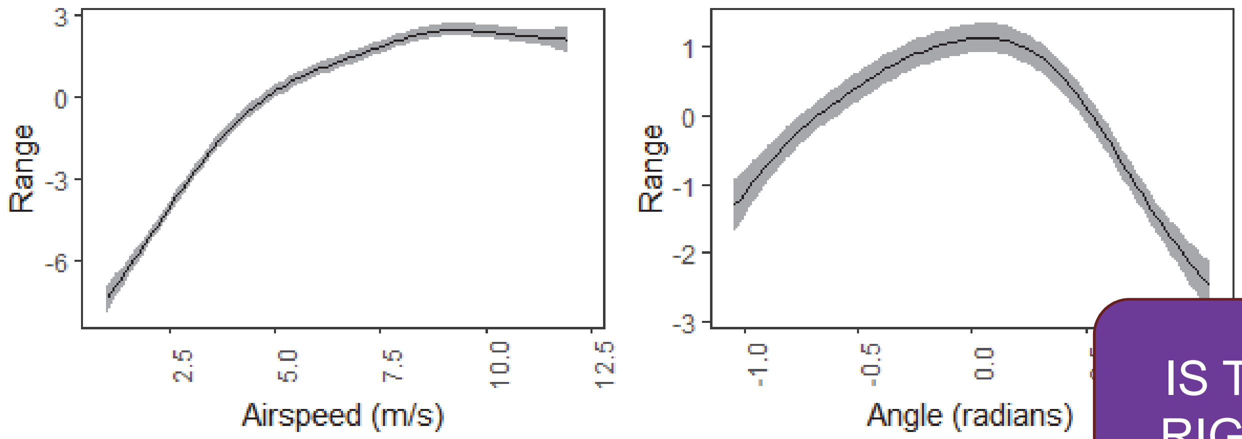
SURFACE "RIDGES"

TWO LOCAL EXTREMA

WRPA – World Record Paper Airplane

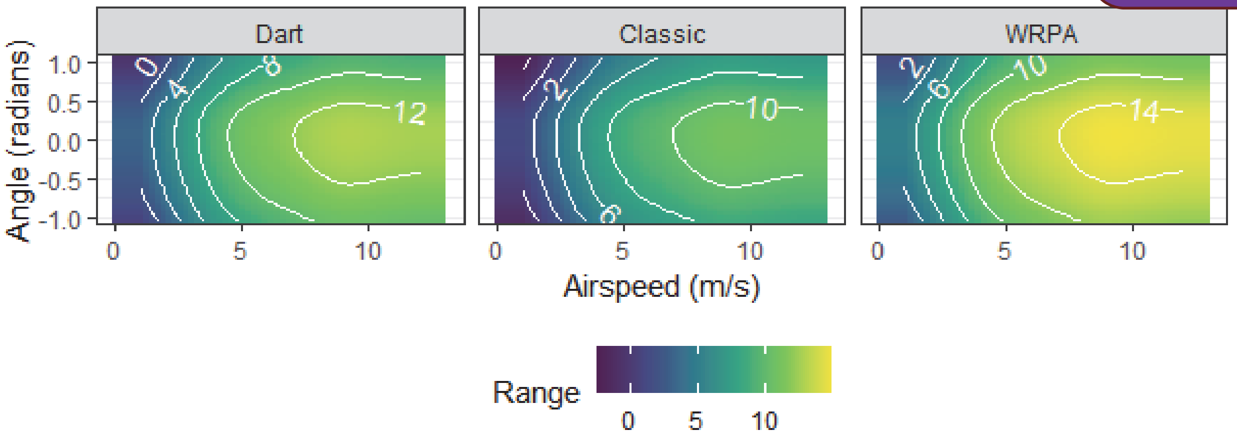
GAMs allow for qualitative assessment of simulation system performance

$$\text{range}_i = \beta_0 + \beta_{\text{cl}}\text{classic}_i + \beta_{\text{wr}}\text{wrpa}_i + f_{\text{as}}(\text{airspeed}_i) + f_{\text{an}}(\text{angle}_i) + \epsilon_i$$



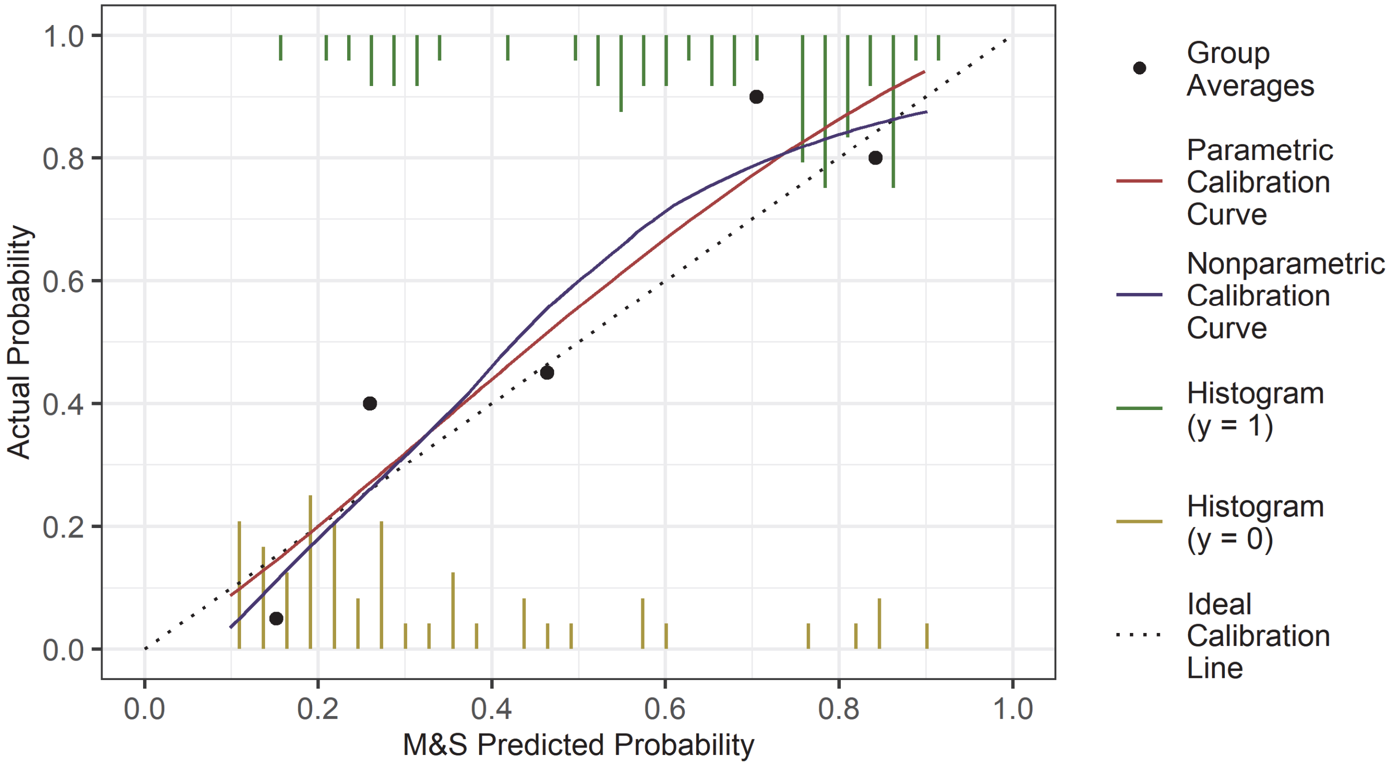
IS THIS RIGHT?

Average Range:
Dart: 9.3 m
Classic: 7.4 m
WRPA: 11.3 m



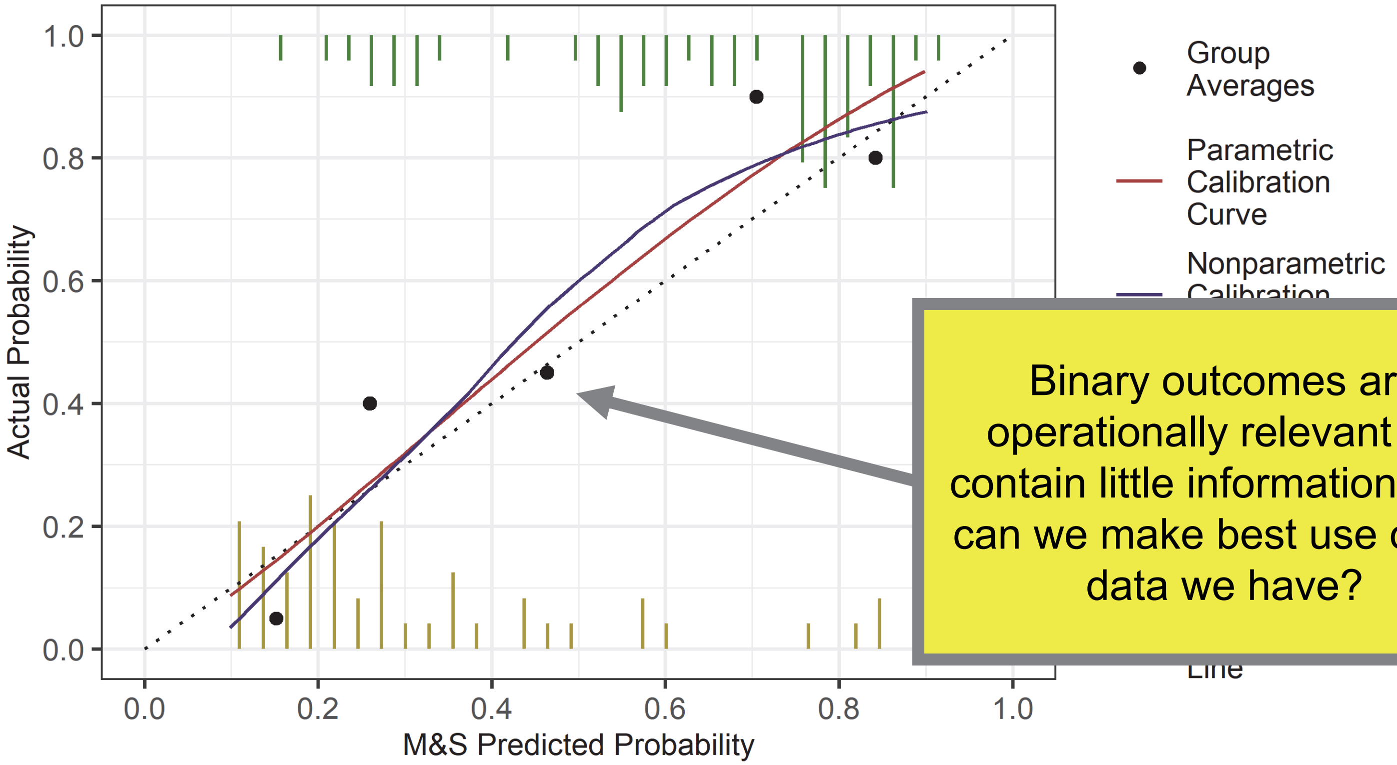
GAM – Generalized Additive Model; WRPA – World Record Paper Airplane

How can statistical surrogates allow for M&S validation with small sample sizes?



M&S – Modeling and Simulation

How can statistical surrogates allow for M&S validation with small sample sizes?



Binary outcomes are operationally relevant but contain little information; how can we make best use of the data we have?

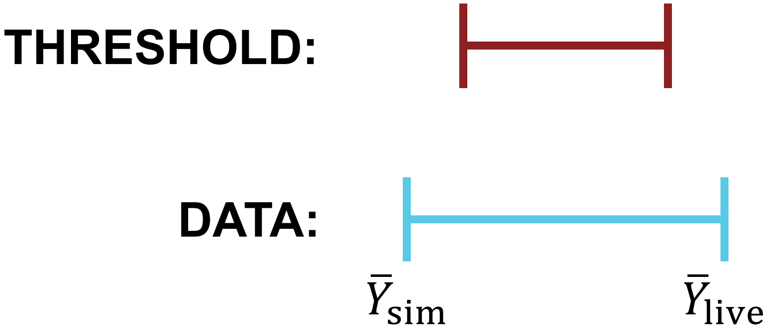
M&S – Modeling and Simulation

How should statistical testing incorporate statistical surrogates?

H_0: Sim=Live

H_A: H_0 is false

$$W = \frac{(\bar{Y}_{\text{sim}} - \bar{Y}_{\text{live}})^2}{\text{Var}(\bar{Y}_{\text{sim}} - \bar{Y}_{\text{live}})}$$



THRESHOLD:

DATA:

$\bar{Y}_{\text{sim}}$ $\bar{Y}_{\text{live}}$

Reject if $W > C_\alpha$.

M&S – Modeling and Simulation

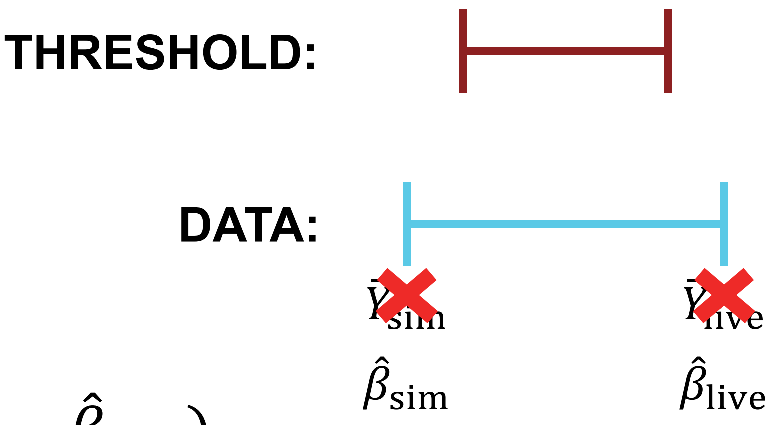
How should statistical testing incorporate statistical surrogates?

H_0: Sim=Live

H_A: H_0 is false

$$W = \xcancel{\frac{(\bar{Y}_{\text{sim}} - \bar{Y}_{\text{live}})^2}{\text{Var}(\bar{Y}_{\text{sim}} - \bar{Y}_{\text{live}})}}$$

$$W = \left(\hat{\beta}_{\text{sim}} - \hat{\beta}_{\text{live}}\right)^{\text{T}} V^{-1} (\hat{\beta}_{\text{sim}} - \hat{\beta}_{\text{live}})$$



THRESHOLD:

DATA:

$\xcancel{\bar{Y}_{\text{sim}}}$ $\xcancel{\bar{Y}_{\text{live}}}$

$\hat{\beta}_{\text{sim}}$ $\hat{\beta}_{\text{live}}$

Reject if $W > C_\alpha$.

M&S – Modeling and Simulation

How should statistical testing incorporate statistical surrogates?

H_0: Sim=Live

H_A: H_0 is false

$$W = \xcancel{\frac{(\bar{Y}_{\text{sim}} - \bar{Y}_{\text{live}})^2}{\text{Var}(\bar{Y}_{\text{sim}} - \bar{Y}_{\text{live}})}}$$

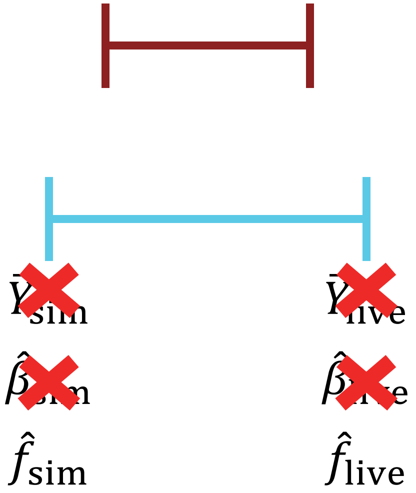
$$W = \xcancel{(\hat{\beta}_{\text{sim}} - \hat{\beta}_{\text{live}})^{\text{T}} V^{-1} (\hat{\beta}_{\text{sim}} - \hat{\beta}_{\text{live}})}$$

$$W = \left\| \hat{f}_{\text{sim}} - \hat{f}_{\text{live}} \right\|_{V^{-1}}^2$$

Reject if $W > C_\alpha$.

THRESHOLD:

DATA:


$\xcancel{\bar{Y}_{\text{sim}}}$
$\xcancel{\bar{Y}_{\text{live}}}$
$\xcancel{\hat{\beta}_{\text{sim}}}$
$\xcancel{\hat{\beta}_{\text{live}}}$
$\hat{f}_{\text{sim}}$
$\hat{f}_{\text{live}}$

M&S – Modeling and Simulation

How should statistical testing incorporate statistical surrogates?

H_0: Sim=Live

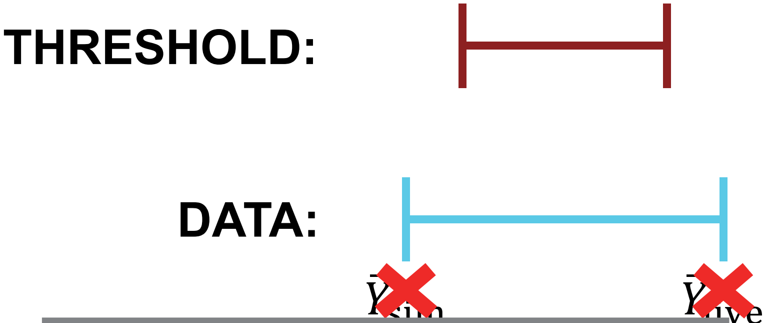
H_A: H_0 is false

$$W = \xcancel{\frac{(\bar{Y}_{\text{sim}} - \bar{Y}_{\text{live}})^2}{\text{Var}(\bar{Y}_{\text{sim}} - \bar{Y}_{\text{live}})}}$$

$$W = \xcancel{(\hat{\beta}_{\text{sim}} - \hat{\beta}_{\text{live}})^{\text{T}} V^{-1} (\hat{\beta}_{\text{sim}} -}$$

$$W = \left\| \hat{f}_{\text{sim}} - \hat{f}_{\text{live}} \right\|^2_{V^{-1}}$$

Reject if $W > C_\alpha$.



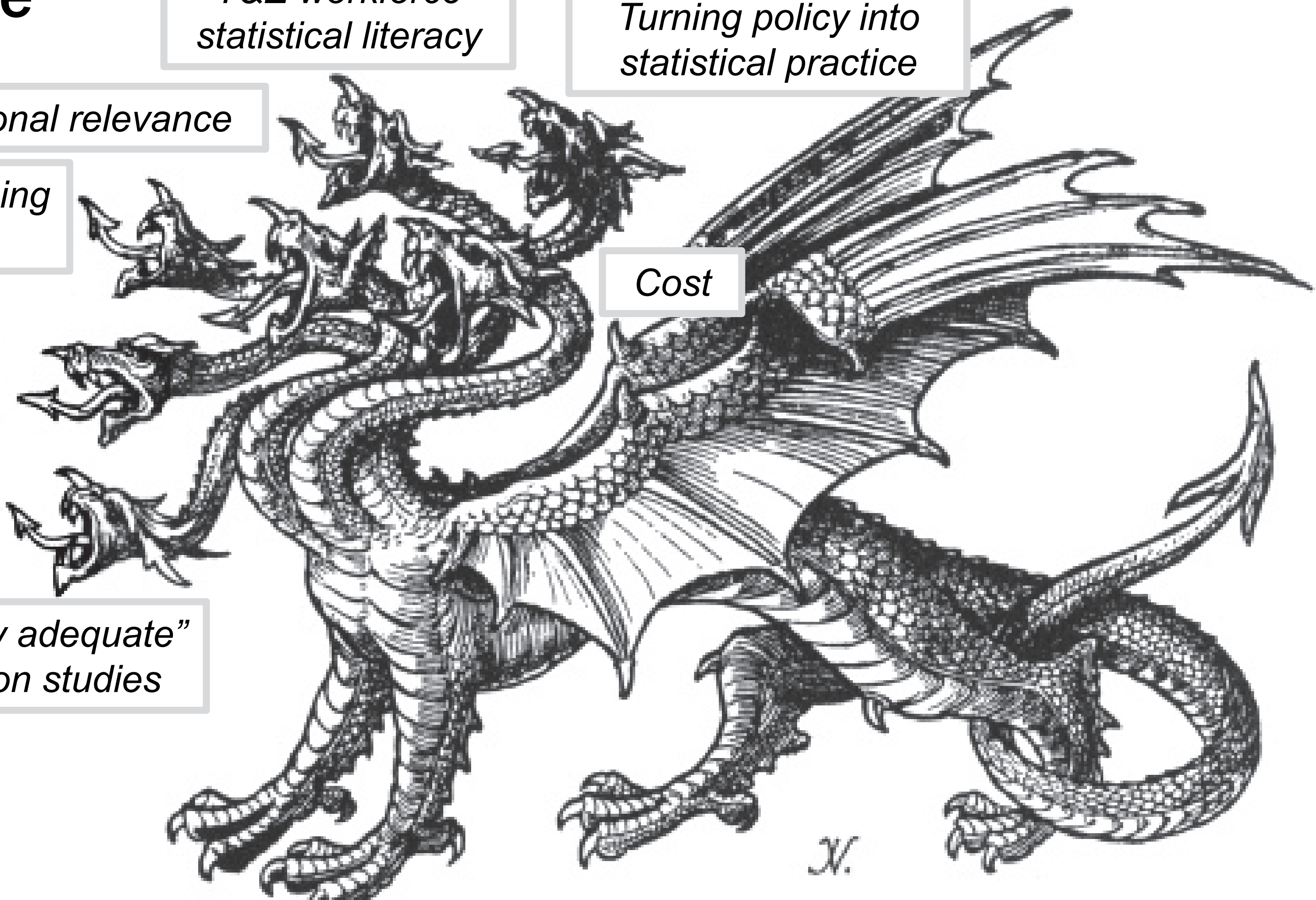
Should these be

- Regression model coefficients?
- Response functions (e.g., GAM smooth)?
- Average response at select factor combinations?

M&S – Modeling and Simulation

How will we overcome the challenges of modeling and simulation validation?

M&S validation faces many challenges that must be overcome



T&E workforce statistical literacy

Turning policy into statistical practice

Operational relevance

Continually changing M&S systems

Cost

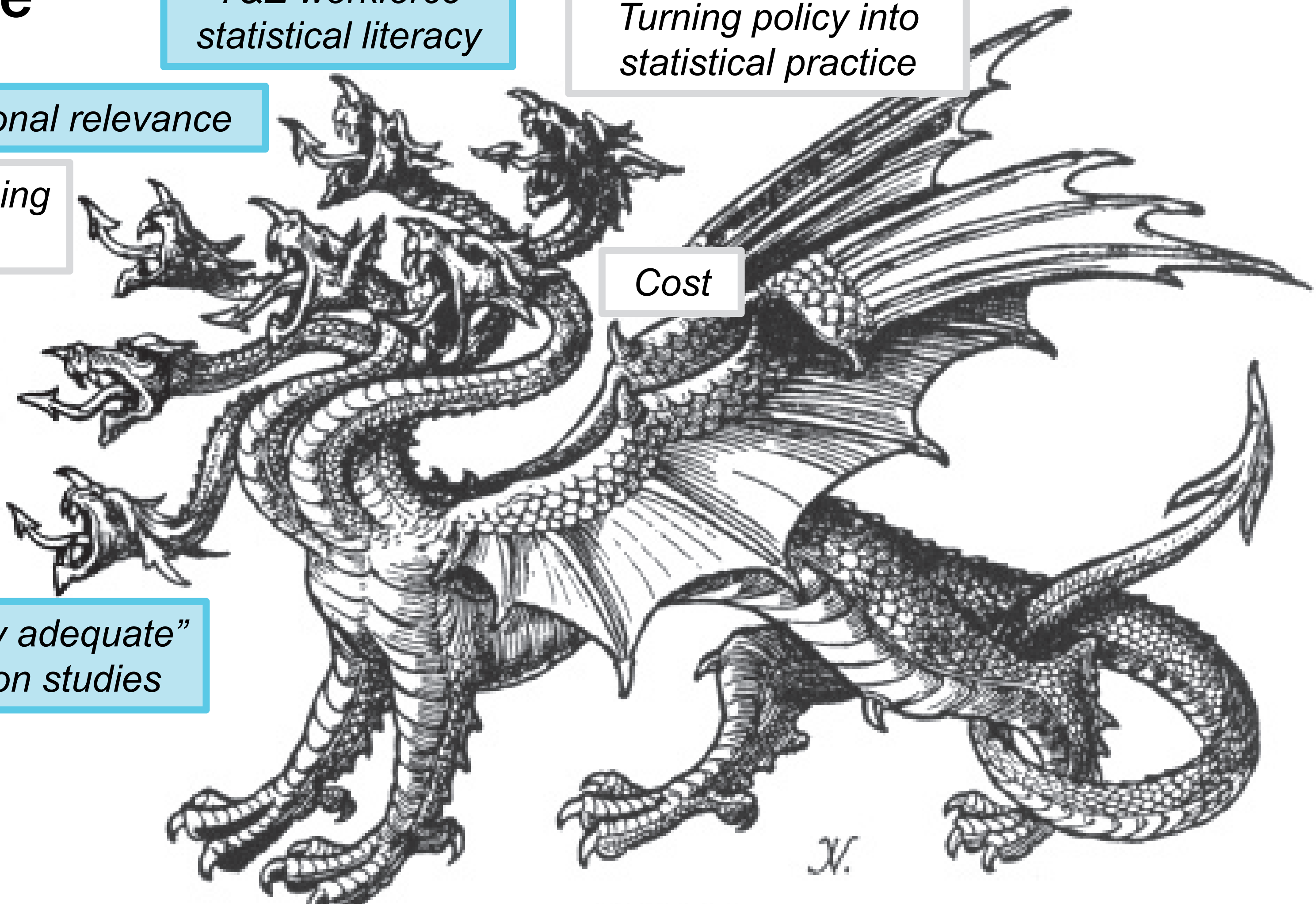
Small real world samples

"Minimally adequate" simulation studies

The Hydra.

M&S – Modeling and Simulation; T&E – Test and Evaluation

M&S validation faces many challenges that must be overcome



T&E workforce statistical literacy

Turning policy into statistical practice

Operational relevance

Continually changing M&S systems

Cost

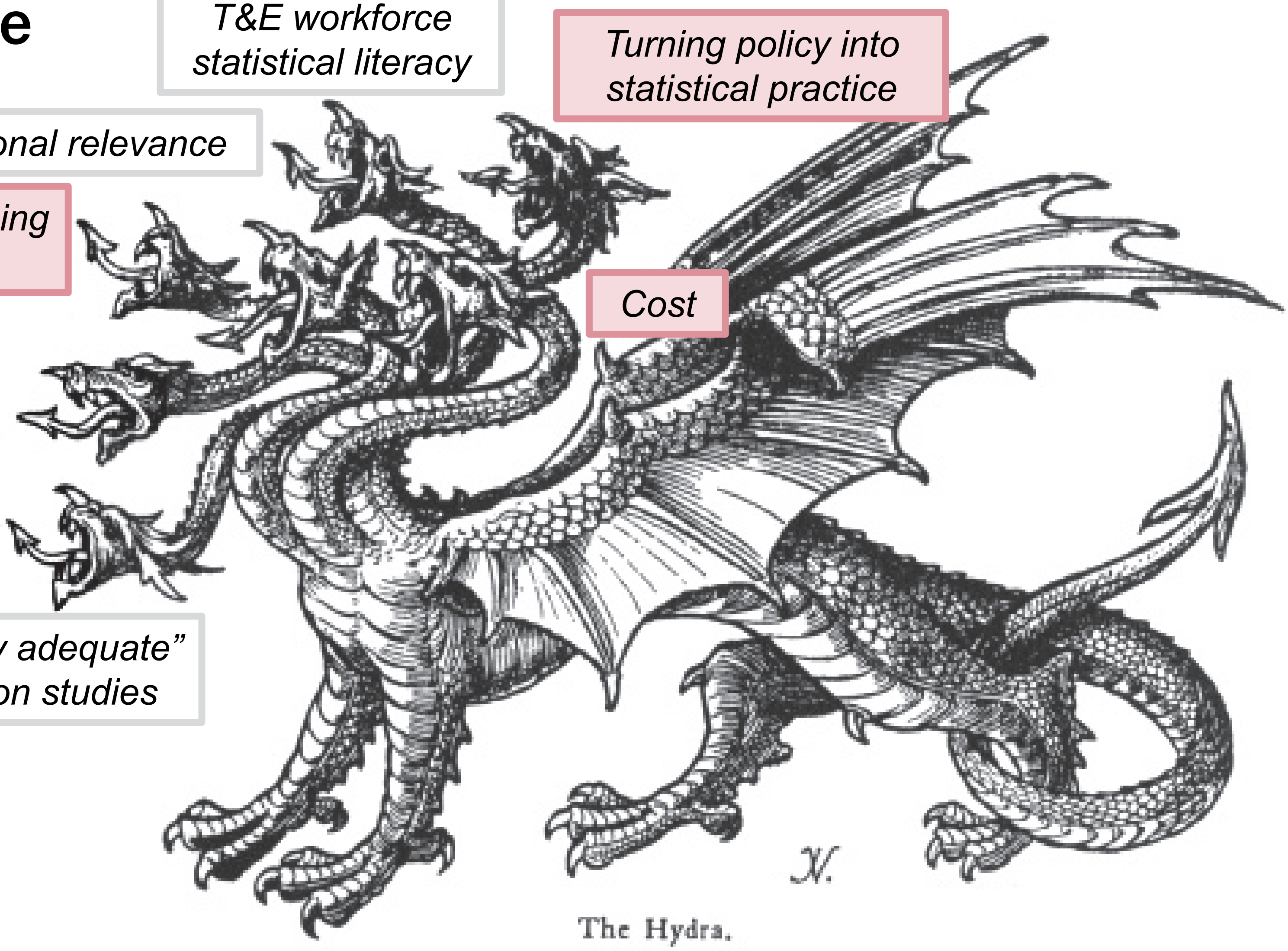
Small real world samples

"Minimally adequate" simulation studies

The Hydra.

M&S – Modeling and Simulation; T&E – Test and Evaluation

M&S validation faces many challenges that must be overcome



T&E workforce statistical literacy

Turning policy into statistical practice

Operational relevance

Continually changing M&S systems

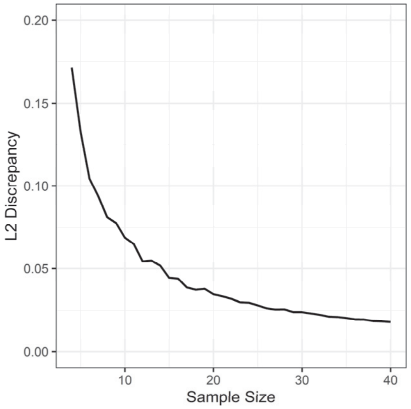
Cost

Small real world samples

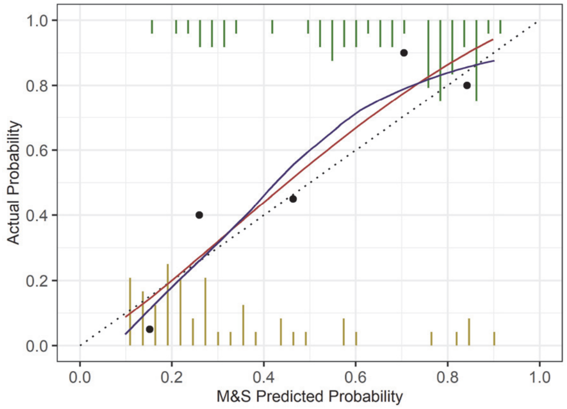
“Minimally adequate” simulation studies

M&S – Modeling and Simulation; T&E – Test and Evaluation

We want statisticians to address…



… how many observations to collect from M&S



… how to best handle binary responses

$$W = \left\|\hat{f}_{\text{sim}} - \hat{f}_{\text{live}}\right\|_{V^{-1}}^{2}$$

… how to make statistical decisions with surrogate models

M&S – Modeling and Simulation

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*
08-2023

2. REPORT TYPE
OED Draft

3. DATES COVERED *(From - To)*

4. TITLE AND SUBTITLE
Space-Filling Experimental Design and Surrogate Models for U.S. Department of Defense Modeling and Simulation Evaluation

5a. CONTRACT NUMBER
HQ0034-19-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Miller, Curtis, G.

5d. PROJECT NUMBER
BD-09-2299

5e. TASK NUMBER
2299(90)

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Defense Analyses
730 East Glebe Road
Alexandria, Virginia 22305

8. PERFORMING ORGANIZATION REPORT NUMBER
1038122

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Director, Operational Test and Evaluation
1700 Defense Pentagon
Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)
DOT&E

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
This OED Draft has not been approved by the sponsor for distribution and release. Reproduction or use of this material is not authorized without prior permission from the responsible IDA Division Director.

13. SUPPLEMENTARY NOTES
Project Leader: John T. Haman

14. ABSTRACT
The U.S. Department of Defense uses modeling and simulation (M&S) for test and evaluation of systems acquired by the Services. The Director, Operational Test and Evaluation (DOT&E), who provides oversight of operational testing, needs "to have the same understanding of and confidence in the data obtained from M&S as... any other data," specifically requiring that design of experiments (DOE) methodologies be used when generating M&S output to explore conditions in which system will be employed, and statistical surrogates be estimated to characterize M&S predictions. Current policy does not recommend specific design and analysis methods, and there remains a gap in the defense community's statistical practice for M&S that needs to be filled. This presentation recommends how DOT&E policy can be fully operationalized by the test community. Specifically, we advocate for the use of space-filling designs (SFDs) to collect M&S data and for statistical emulators, Gaussian processes (GPs), or generalized

15. SUBJECT TERMS
Space Filling Designs; metamodel; Modeling and Simulation (M&S) Validation; Director of Operational Test and Evaluation (DOT&E); Statistics; American Statistical Association (ASA)

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			John T. Haman
Unclassified	Unclassified	Unclassified	SAR	48	**19b. TELEPHONE NUMBER** *(Include area code)* (703) 845-2132