13B.6 STATISTICAL DESIGN OF EXPERIMENTS IN NUMERICAL WEATHER PREDICTION: SOME EMERGING RESULTS

Jeffrey A. Smith*, Richard S. Penc, John W. Raby
US Army Research Laboratory, White Sands Missile Range, New Mexico

1 INTRODUCTION

A typical motivation in many Atmospheric Sciences analyses, especially those involving modeling, is to trace an adjustment in some set of parameters to an effect on some output metric of interest subject to some set of conditions. One problem facing such analyses is how does the analyst tease out what could be a potentially small effect from what are, arguably, larger effects due to conditions such as the location of the modeling domain or the day(s) over which the model run is executed. We attempt to address these questions through the use a technique called "statistical design of experiments." Though in wide use elsewhere, there is little evidence in the literature of statistical design of experiments use within the Atmospheric Science community.

In section 2, we present some background on the basic elements of statistical design of experiments (DoE). Section 3 follows with a description of our problem, and with section 4 how we employed DoE to address that problem. In section 5 we define the model, domains and cases over which we conduct our analyses. In section 6, we outline some of our current results, and in section 7, we present our initial conclusions. We conclude the abstract by addressing a question about how DoE methods compare to Stein and Alpert (1993) in section 8.

2 STATISTICAL DOE AND NUMERICAL WEATHER PREDICTION

DoE, as a collection of methods, emerged from the early work of Fisher (1925, 1935) and his colleagues (e.g., Yates 1937) in their attempts to systematize the study of fertilizer and other treatments in support of the agricultural sciences at the Rothamsted Experimental Station in England. Building on the work of Fisher and his colleagues, Box et al. (1978); Montgomery (2013) among many others applied DoE methods to many problem areas such as those in industrial process optimization and control. Beginning with McKay et al. (1979) and the work of Sacks et al. (1989a) and (1989b) researchers further extended DoE methods for use with numerical simulations involving continuous factors for application to uncertainty quantification. More recently researchers such as Santner et al. (2003), Kleijnen et al. (2005) and Kleijnen (2008), have applied experimental design to the study of complex simulation codes where the factors may be a mix of continuous, categorical and/or discrete valued. In almost a 100 years of research, experimental design has evolved into a robust and comprehensive collection of methodologies that allow rigorous experimentation in many complex systems far removed from Fisher's initial application.

Despite the evolution of these techniques in other fields, there is little direct evidence to suggest that researchers have applied these techniques to study numerical weather prediction (NWP). Absent any direct evidence, Berci et al. (2014); Rahimi et al. (2014); and Zhu et al. (2015) have applied experimental design methods to computational fluid dynamics codes as part of an engineering development process. By recognizing that computational fluid dynamics codes share many of the same complexities exhibited by NWP, we suggest that experimental design may prove useful in forecasting the weather via mathematical models as well as the analysis of the attendant models.

3 PROBLEM DESCRIPTION

Unlike a typical operational use of NWP that may provide a forecast over a fixed region and which may be tuned over time to perform well, NWP for tactical forces will likely have neither of those luxuries; time and personnel with specialized training in NWP are premiums on a tactical battlefield. This fact often precludes detailed error analyses of why forecasts were "off" and how they may be improved. Thus we ask the question:

* *Corresponding author address:* Jeffrey A. Smith, US ARL, WSMR, NM, 88002, e-mail: jeffrey.a.smith1.civ@mail.mil

"how can we provide the warfighter with a NWP capability that is both 'robust' and capable of providing a 'good' forecast with minimal user intervention?" Although answering this question completely is beyond the scope of this extended abstract, we suggest that DoE can provide the means to answer this question.

Consider a forecast as a map from an observed to some future atmospheric state. Mathematically, we say: $f: x \to y,$ where f can be a NWP code such as WRF-ARW (Skamarock et al. 2008), x represents initialization conditions and y a forecast. Because NWP codes often have many user selectable features such as parameterization schemes, nesting ratios, integration times, etc. it is useful to think of f as a set of functions where each $f_i \in f, i = 1, 2, \ldots, N$ is a specific configuration of the NWP code. Note, N in this case, though countable, is so large as to preclude search using brute force methods.

With these definitions we define our problem as one of finding $f_{(i)} \in f$ that, for purposes of this paper, minimizes the difference between the forecast and observed values at a point in time. Note: the parenthesis around the subscript indicates our object of search; a specific, but as yet unknown configuration satisfying our criteria.

4 APPLYING DOE

Though the DoE methods are general in nature we restrict our attention exclusively to WRF-ARW (henceforth WRF) so that our application of these methods are clear. Figure 1 shows interaction of the macro level processes modeled within WRF. It is the complex interaction between parameterized processes operating on the initial conditions that produce a forecast value at a given point in space. Thus, to attempt answering the question raised in section 3 requires that we first create a meta-model of the error. To do so, will strategically sample f using a 'few' runs in order to identify those parameterized processes that are dominant for a set of domains. Our goal here is to create a "screening experiment" wherein we experiment with many factors with a goal of reducing them to a few that are statistically driving the error.

To create our design, we employ a linear program to create a "design matrix" which we define as a $m \; x \; n$ matrix of run configurations wherein each row of that matrix is a $f_i \in f$ as we have described above. The specific method we've adapted, based on Vieira et al. (2011; 2013), supports continuous, discrete, and categorical factors. For our design, we treat all the factors as categorical factors (Smith and Penc 2015, 2016; Smith et al. 2017, 2018). By using a linear program, we can create an experimental design that is both "nearly orthogonal" which allows us to separate the effects of the various factors in analysis, and "balanced" which results in statistics that are not overly skewed. In addition, linear programs allow us to allocate specific arrangements to one of a desired set of domains as well as a particular day. In this manner, we create what are called *blocks* that allow us to account for experimental condition that we cannot control yet are important considerations in the analysis.

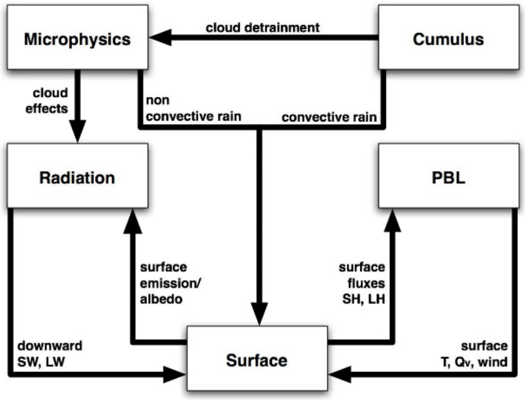

Figure 1: Direct interactions of parameterization schemes in WRF (Dudhia 2015)

The statistical model we are considering is given by

$$\delta_{ijkl} = D_i + C_j + T_{ijk} + \varepsilon_{ijkl} \qquad (1)$$

where δ is the model bias $\delta_{ijkl} = F_{ijkl} - O_{ijkl}$, with F_l representing the model forecast at some station in the domain, and O_l the observed value at the same station, D models the effect of the domain on the error, C models the effect of the day which serves as a proxy for the large scale synoptic situation, T the specific configuration of the model used to produce the data, and ε any residual error. The subscripts i, j and k index the domain, case, and treatment effects while l indexes the matched forecast – observation pair data element produced by MET Point-Stat (National Center for Atmospheric Research 2016) for a given domain and case combination.

We present the resulting experiment in the next section.

5 THE EXPERIMENT

The model employed uses WRF (Skamarock et al. 2008) initialized with 0.5° GFS (NOAA 2018b) forecast data providing gridded background fields with raw observations analyzed onto the background fields; 1/12° (~9 km) RTG high resolution SST (NOAA2018a); and 1 km NOHRSC SNODAS (NOAA 2018c) snow data when available and GFS snow data elsewhere. We conducted data assimilation using a 6-h pre-forecast with observation nudging (12-18 UTC). Observation nudging during data assimilation uses TAMDAR (AirDat 2018) aircraft data and various MADIS (NOAA 2018d) datasets [standard surface observations, mesonet surface observations, maritime surface observations, profiler data, rawinsondes, and ACARS (aircraft) data (Mamrosh 1997)] (Dumais and Reen 2013; Dumais et al. 2015). The model top was set at 10mb for all runs.

As equation (1) captures, there are two elements which complicate this design: 1), the region modeled, and 2) the large scale synoptic conditions driving our forecast. While these elements must be considered in our analysis they are, in the statistical sense, nuisance factors by which we mean that the 'signal' sought is buried within what is likely to be the larger signal contributed by the regional and synoptic pictures. To account for these factors, we consider two domains with the grid arrangement defined in Table 1 and depicted in Figure 2. The synoptic features of the atmosphere are considered using the case days as given in Table 2. Our rationale in choosing these particular domains was twofold: 1) both domains are densely populated with various stations which should provide good observational comparisons, and 2) the distance between the domains provides some potential insight into the effects of latitude on the forecast as well as an approximately similar synoptic weather situation we expect to allow us to treat domain and case day as statistically independent.

Table 1: Grid configuration

Scale (km)	Points	Dimension (km)
9	175 × 175	1566 × 1566
3	242 × 241	720 × 720
1	127 × 127	126 × 126

Table 3 arrays the parameterization schemes we use to create our design, and Table 4 is the resulting design matrix. Examine Table 3 and note that the two columns noted as 'blocks' correspond to the terms D and C while the six columns marked as 'parameterization schemes' model T in (1).

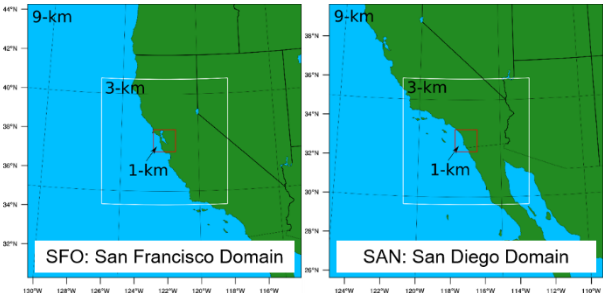

Figure 2: Model grid centers. The left domain is centered at San Francisco Airport and the right domain centered at San Diego Airport.

As a check on our design, we convert each factor and level to a numerical form, i.e., for the domain "SAN" write 1, and "SFO" write 2, etc. Once complete, we can find some indication of how well our design fared in terms of orthogonality and balance by examining the correlation matrix. This matrix is given in Figure 3. Figure 3 leverages the symmetry property of the correlation matrix to provide the pairwise correlation coefficient in the upper triangle, and in the lower triangle, a dot where we have made a run exhibiting the particular set of characteristics described by the row and column labels. Note, a given run will result in multiple dots depending as the pair changes. Along the main diagonal we have used a shape to give us a rough sense of balance in the design.

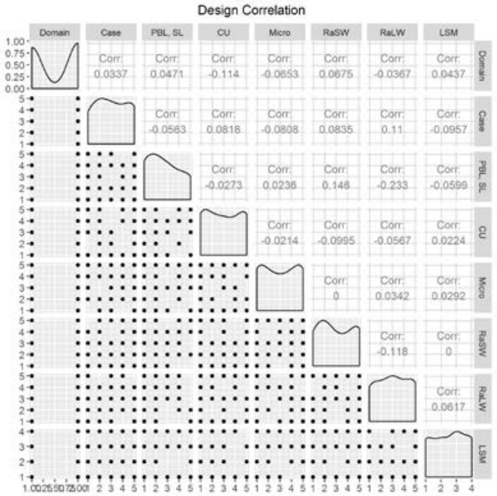

Figure 3: Assessing the strength of our design through the correlation matrix.

Table 2: The synoptic description for each domain by particular case.

		Domain	
Case[1]	Dates (2012)	San Francisco (SFO)	San Diego (SAN)
1	Feb. 07–08	An upper level trough with associated frontal system moved onshore which led to widespread precipitation in the region that diminished mid-period.	Surface front / upper level trough moved onshore, which led to widespread precipitation in the region.
2	Feb. 09–10	Quiescent weather dominated the region with an upper level ridge remaining centered over central California	Quiescent weather was in place with an upper level ridge centered over central California at 12 UTC.
3	Feb. 16–17	An upper level ridge located over northern California in combination with a surface high pressure area centered over the Rocky Mountains east of the domain produced quiescent weather in the region.	An upper-level low located near the California/Arizona border with Mexico at 12 UTC brought precipitation to that portion of the domain. This pattern moved south and east over the course of the day.
4	Mar. 01–02	A weak shortwave upper level trough with associated cold front resulted in considerable cloudiness and light precipitation over the region until after mid-period when conditions stabilized following frontal passage.	A weak shortwave trough resulted in precipitation in northern California at the beginning of the period that spread to Nevada, then moved southward and decreased in coverage.
5	Mar. 05–06	Weak surface pressure gradients at the surface and broad zonal flow aloft slowly gave way to stronger synoptic forcing in advance of a cold front that approached the region near the end of the period bringing increased cloudiness, but very limited precipitation.	Widespread high-level cloudiness due to weak upper-level low pressure but very limited precipitation.

[1]: All case studies are 24 hours in length, running from 12 UTC to 12 UTC on the days listed with forecasts made on the hour.

Examining the correlation coefficients in Figure 3 reveals that in most cases the correlation is less than 10% suggesting the design is very nearly orthogonal, with only the RaLW (long wave radiation scheme) being somewhat strongly correlated with the PBL_SL (boundary layer coupled with the surface layer) factor. Vieira et al. (2011) and Vieira et al. (2013) cite Bathke (2004) and assert that correlations less than about 20% typically can be analyzed as if the factors were orthogonal. Given that only RaLW and PBL_SL factors marginally depart from this standard, we will proceed as if these two factors were indeed orthogonal. Although the design is not in perfect balance, a reasonable degree of balance does exist which suggests that no one factor setting will dominate the analysis.

6 EMERGING RESULTS

Model runs based on the design as described in the previous section were configured and executed on the Army's High Performance Computer name "excalibur". Of the designed 40 runs, approximately 50% of those runs "crashed" at the outset, a condition we attribute to configurations that have not been tested before or are just incompatible with current WRF code. Accordingly, we adapted the design and recovered approximately 8 runs for a total of 28 runs for which we present emerging results. These runs were post processed using MET Point-Stat (National Center for Atmospheric Research 2016) to produce matched pair files for each forecast hour. All files were combined using the R environment (R Core Team 2017) and the tidyverse approach to data (Wickham and Grolemund 2016; Wickham 2017). We augmented the resulting data frame with the conditions based on our design and computed bias values for each matched pair.

Table 3: Parameterization schemes employed in the design by WRF namelist entry[1]

Planetary Bound. Layer, Surface (PBL, SL)		Short Wave (RaSW)	
1, 1	YSU with revised MM5	1	Dudhia
2, 2	MYJ with ETA	2	Goddard
5, 1	MYNN2 with revised MM5	4	RRTMG
7, 7	ACM2 with Pleim-Xu	7	FLG
11, 1	Shin-Hong with revised MM5	99	GFDL
Cumulus (CU)[2]		Long Wave (RaLW)	
1	KainFritsch (KF)	1	RRTM
2	Betts-Miller-Janjic (BMJ)	4	RRTMG
6	Tiedke	5	New Goddard
16	New Tiedke	7	FLG[3]
93	Grell-Devenyi	99	GFDL
Microphysics (Micro)		Land Surface Model (LSM)	
2	Lin (Purdue)	1	5 layer Thermal Diffusion
4	WSM5	2	NOAH
5	ETA (Ferrier)	3	RUC operational
7	Goddard	5	CLMv4
8	Thompson		

1: For specific references to the physics schemes along with translations of the acronyms please refer to Skamarock et al. (2008)

2: Cumulus scheme applied to the outer domain only (Not considered in this extended abstract)

3: Every run with the FLG long wave radiation scheme failed, but not every failed run used the FLG scheme We are investigating replacement schemes for these failed points in order to come closer to our desired 40 runs.

We compared, the forecasts at two times: 21Z and 00Z for each domain and case combination as well as for the parameterization classes depicted in Figure 1 save the cumulus scheme which was ignored for this study. We calculated the mean bias for each domain and case day as a function of the boundary layer, microphysical and land surface as well as the short wave and long wave radiation parameterization schemes. Results for a typical comparison case are given in Figure 6 for the Z2 temperature values.

In Figure 6, we note distinct variation as a function of domain and case days which is to be expected. However, one also notes variation due both to the physics parameterization schemes and time. We have yet to carry out the detailed analysis, but visual examination of these changes, as well as the number of matched pairs (≈ 200) for each point on the figure, suggests that these changes, as mean comparisons, are likely to be statistically significant. Furthermore, the regularity in Figure 6 suggests a reasonable expectation that we will be able to estimate an effect due to the domain and case day factors. Should this expectation prove fruitful, we will be able to remove these large scale effects from the data and focus on the treatment effects [the T in (1)] and potentially partition that effect into direct contributions of the physical parameterization schemes, and thus the parameterization classes given in Figure 1.

7 DISCUSSION AND CONCLUSIONS

One common criticism that is often raised in a discussion on design of experiments is how can one show value? Penc et al. (2018a, b) studied the variation of Bias and RMSE exclusively as a function of a range of PBL schemes that included those found in Table 3. In Figure 4 we show the mean dew point bias for a portion of that study based solely on the PBL schemes found in Table 3, and in Figure 5 we show a Box plot of the bias values at 00 UTC. From Figure 4 we see little evidence that the various schemes perform statistically different at 00 UTC, but from the Box plot we see that there may be a difference in the means (location).

Table 4: The design matrix arrived at, and used in this study.

	BLOCKS		PARAMETERIZATION SCHEMES					
CASE[*][†]	**DOMAIN**[‡]	**CASE DAY**	**BL_PBL**[1]	**SF_SFCLAY**[1]	**MP**	**RA_LW**[2]	**RA_SW**	**SF_SURFACE**[3]
CASE24	SAN	02/07	MYNN2	revised MM5	ETA (Ferrier)	NG	RRTMG	CLMv4
CASE3	SAN	02/07	Shin-Hong	revised MM5	Goddard	RRTM	Goddard	NOAH
CASE27	SAN	02/09	MYNN2	revised MM5	Lin (Purdue)	NG	Goddard	RUC operational
CASE6	SAN	02/09	Shin-Hong	revised MM5	Lin (Purdue)	RRTMG	GFDL	NOAH
CASE31	SAN	02/09	MYJ	ETA	Thompson	RRTM	Dudhia	5 layer
CASE17	SAN	02/09	MYNN2	revised MM5	Lin (Purdue)	NG	Dudhia	RUC operational
CASE22	SAN	02/16	MYJ	ETA	WSM5	GFDL	Goddard	NOAH
CASE23	SAN	02/16	ACM2	Pleim-Xu	Thompson	NG	GFDL	5 layer
CASE16	SAN	02/16	YSU	revised MM5	Goddard	GFDL	GFDL	CLMv4
CASE28	SAN	03/01	MYJ	ETA	Thompson	NG	Goddard	RUC operational
CASE34	SAN	03/01	YSU	revised MM5	Thompson	GFDL	FLG	NOAH
CASE21	SAN	03/01	MYJ	ETA	WSM5	GFDL	Dudhia	NOAH
CASE5	SAN	03/05	Shin-Hong	revised MM5	ETA (Ferrier)	RRTM	Goddard	CLMv4
CASE35-2	SAN	03/05	YSU	revised MM5	Lin (Purdue)	RRTM	GFDL	CLMv4
CASE4	SFO	02/07	MYNN2	revised MM5	Thompson	NG	RRTMG	RUC operational
CASE38	SFO	02/07	YSU	revised MM5	WSM5	RRTMG	Goddard	RUC operational
CASE12	SFO	02/07	MYJ	ETA	Thompson	RRTMG	Dudhia	RUC operational
CASE40-1	SFO	02/09	ACM2	Pleim-Xu	Goddard	RRTMG	FLG	RUC operational
CASE2-1	SFO	02/09	MYNN2	revised MM5	Lin (Purdue)	NG	Dudhia	CLMv4
CASE9	SFO	02/16	YSU	revised MM5	Lin (Purdue)	RRTM	GFDL	RUC operational
CASE7	SFO	02/16	Shin-Hong	revised MM5	Goddard	RRTMG	GFDL	RUC operational
CASE20	SFO	02/16	MYJ	ETA	Lin (Purdue)	GFDL	RRTMG	5 layer
CASE26	SFO	03/01	MYNN2	revised MM5	Thompson	GFDL	Goddard	CLMv4
CASE14	SFO	03/01	ACM2	Pleim-Xu	Goddard	GFDL	RRTMG	RUC operational
CASE15	SFO	03/01	ACM2	Pleim-Xu	WSM5	RRTMG	GFDL	5 layer
CASE8	SFO	03/05	YSU	revised MM5	Thompson	RRTM	Dudhia	5 layer
CASE13	SFO	03/05	YSU	revised MM5	Goddard	NG	Dudhia	NOAH
CASE39	SFO	03/05	MYNN2	revised MM5	ETA (Ferrier)	NG	FLG	RUC operational

Note: All model runs were executed using WRF 3.8.1 on the excalibur super computer.

[1] Although BL_PBL and SF_SFCLAY are reported as separate factors, they were treated as a single factor for the design. Thus, 'MYNN2' and 'revised MM5' are considered as one level of a single factor.

[2] NG is New Goddard.

[3] 5 layer is 5 layer Thermal Diffusion.

[*] Runs are not in design order

[†] Numbers after dashes in case name, e.g., case35-2, indicate that run was repeated after a after a change in parameterization scheme. The specific number indicates the number of times that a change was made.

[‡] SAN: Domain Center near San Diego Airport. SFO: Domain Center near San Francisco Airport, see Figure 2.

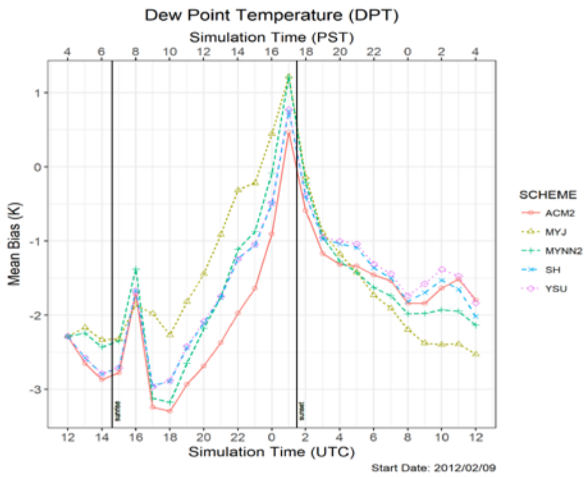

Figure 4: Mean Dew Point Temperature Bias (K) for a subset of data from Penc et al. (2018a, b)

Using R language (R Core Team 2017) we compared the mean biases pairwise, the results of which are given Table 5. This test reveals that the MYJ PBL scheme behaves differently in statistically significant manner than the remaining schemes. This observation holds true save only for the MYNN2 scheme at 00 UTC. When conducted at different times, other results are obtained.

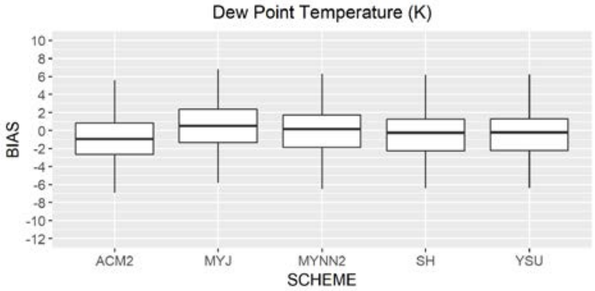

Figure 5: Box plot of Dew Point Temperature Bias (K) at 00 UTC for a subset of data from Penc et al. (2018a, b)

Penc et al. (2018a, b) commented on the relative lack of variation among the various boundary layer schemes and a simple calculation suggests that only 3% to 4% of total model variance is attributable solely to the variation in PBL scheme. This calculation is of η_G^2 as defined by Bakeman (2005) and implemented in the lsr package (Navarro 2015) for the R statistical analysis environment.

Table 5: Tukey's Honest Significant Difference Test conducted at 0.01 significance level.

		Confidence Range		**Adjusted**
Comparison	**Estimate**	**Low**	**High**	**p value**
MYJ-ACM2[1]	1.347748	0.523185	2.172312	1.18E-06
MYNN2-ACM2	0.817729	-0.00683	1.642293	0.010937
SH-ACM2	0.40435	-0.42021	1.228913	0.497443
YSU-ACM2	0.419775	-0.40479	1.244339	0.458493
MYNN2-MYJ	-0.53002	-1.35458	0.294545	0.221756
SH-MYJ[1]	-0.9434	-1.76796	-0.11883	0.001859
YSU-MYJ[1]	-0.92797	-1.75254	-0.10341	0.002344
SH-MYNN2	-0.41338	-1.23794	0.411184	0.474543
YSU-MYNN2	-0.39795	-1.22252	0.426609	0.513797
YSU-SH	0.015425	-0.80914	0.839989	0.999997

[1] Statistically significant effect at 0.01 (Also noted by *italics*).

Our emerging results suggest that we can estimate an effect due to domain, as well as one due to the synoptic condition (using case day as a proxy for the synoptic condition). In doing so, we expect to be able examine the contribution of the parameterization schemes as the sole remaining source of variation in the data. Consequently, we expect we will be able to shed numerical insight into the various flows between schemes described in Figure 1. This is an advance in our ability investigate how NWP perform under various conditions, which consequently allows us to undertake a more macro view of the model for verification purposes. The effect of all of this implies that for a study such as Penc et al. (2018a, b), we can peer more deeply into the model and potentially diagnose more substantial effects with DoE methods at a modest increase in model runs. This ability provides an answer to the question we raised at the beginning of this section.

8 FACTOR SEPARATION VS DOE

Smith et al. (2018) discussed some preliminary consequences of this work at the Annual AMS meeting. During the discussion, Ligia Bernardet, Research Scientist with the Cooperative Institute for Research in Environmental Sciences (CIRES) and NOAA Earth System Research Laboratory, asked “How does this work differ from that of Stein and Alpert (1993) on factor separation?” The answer to that question comes in two parts, the methods used to: 1) probe the model with the desired variations to produce the sample data, and

2) form the analysis based on that same sample data.

Addressing the first part, we note no difference between Stein and Alpert and DoE when sampling the model. Though Stein and Alpert did not state as such, the approach they employed to sample their model is what is called a 2^k Full Factorial design (e.g., Montgomery 2013) in the DoE literature. In this notation, the term 2^k means k factors are considered at 2 levels, a method requiring, at a minimum, a full 2^k runs to be performed which is the same for Stein and Alpert for the same conditions.

Considering the second part, we note that the typical analysis methods employed when analyzing design experiments, and those employed by Stein and Alpert are wholly different. Stein and Alpert base their method on identifying a function from the data using a method based on a Taylor series expansion whereas for DoE, one typically uses an analysis of variance method to identify a generalized linear model that minimizes the residual model error in the squared sense. Initial steps in investigating this question suggest that when “noise” or “measurement error” is not present, the method of Stein and Alpert does identify the function when we test against toy problems; however, this appears not to be the case when noise or error is included as is the case when one investigates model bias error when the observations compared against the forecast come with some unknown zero mean Gaussian error.

To sum up our comment regarding Stein and Alpert, we note that although the 2^k Full Factorial designs are considered the “gold standard” in DoE, they are just one in an array of designs available in the statistical literature. In many cases these other designs can shed insight into particular analytical questions by trading off, say, the inability to identify some higher order interaction effects for a decrease in the computational expense of producing the data. This fact alone suggests to us that there is merit in considering a role for design of experiments in the Atmospheric Science. Finally, one of us (Smith) along with a summer student, will continue to investigate this second question, and expect to present the results of this study at a future time.

9 ACKNOWLEDGEMENTS

We gratefully recognize the assistance of Dr. Brian Reen and Mr. Bob Dumais who generously contributed their expertise to the construction and execution of the models over our domains, as well as the assistance provided by Ms. Leelinda Dawson and Ms. Norma Rivera who helped automate various parts of the data reduction/post processing efforts.

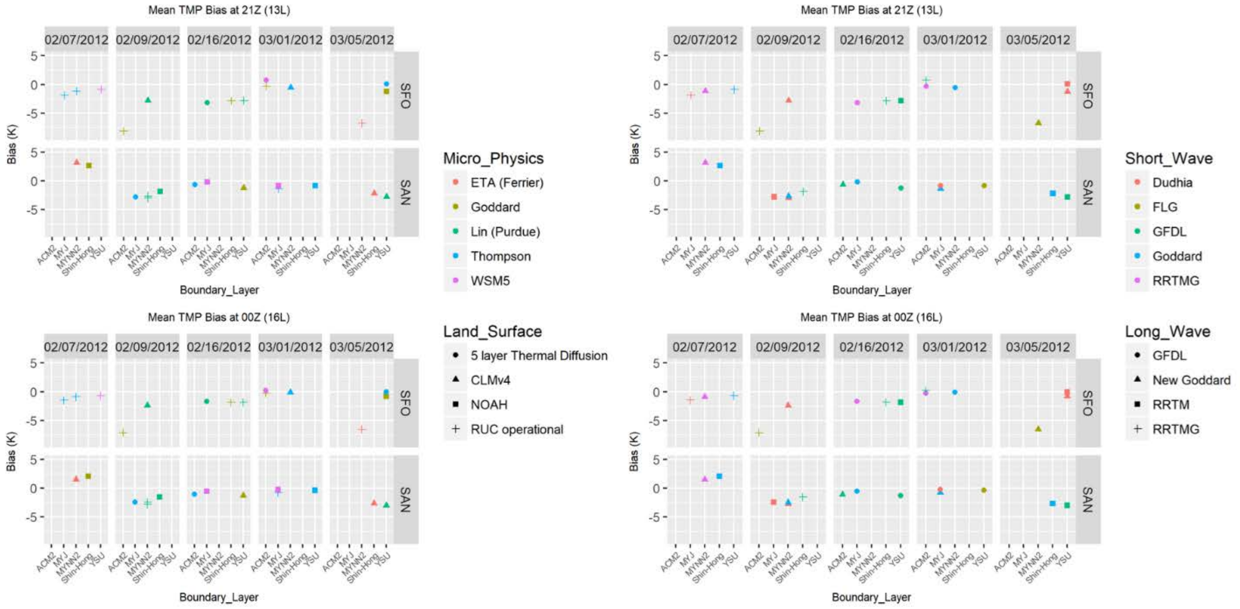

Figure 6: Mean Temperature Bias (K) values at the Z2 level as a function of the Boundary Layer, Micro Physics, Land Surface, Short Wave Radiation and Long Wave Radiation parameterization schemes for a midafternoon (21Z) and transition time (00Z) period for each of two domains and five case days. All plots are scaled and arranged the same so horizontal and vertical comparisons are possible.

10 REFERENCES

AirDat, cited 2018: Airdat Real-Time Tamdar Weather Data and Products. [Available online at http://www.airdat.com/.]

Bakeman, R., 2005: Recommended Effect Size Statistics for Repeated Measures Designs. *Behavior Research Methods*, **37,** 379-384. doi: 10.3758/bf03192707.

Bathke, A., 2004: The Anova F Test Can Still Be Used in Some Balanced Designs with Unequal Variances and Nonnormal Data. *Journal of Statistical Planning and Inference*, **126,** 413-422.

Berci, M., V. V. Toropov, R. W. Hewson, and P. H. Gaskell, 2014: Multidisciplinary Multifidelity Optimisation of a Flexible Wing Aerofoil with Reference to a Small Uav. *Structural and Multidisciplinary Optimization*, **50,** 683-699. doi: 10.1007/s00158-014-1066-2.

Box, G. E. P., W. G. Hunter, and J. S. Hunter, 1978: *Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building.* Wiley.

Dudhia, J., 2015: Overview of WRF Physics. *2015 Basic WRF Tutorial*, Boulder, CO, National Center for Atmospheric Research.

Dumais, R. E., and B. P. Reen, 2013: Data Assimilation Techniques for Rapidly Relocatable Weather Research and Forecasting Modeling. Final Report ARL-TN-0546.

Dumais, R. E., Jr., B. P. Reen, J. A. Smith, D. I. Knapp, and H. Cai, 2015: Developing a WRF-Based Mixed Variational and Nudging Data Assimilation Scheme for the Us Army Convection-Scale Nowcasting System. *95'th Annual AMS Meeting, 19th Conference on Integrated Observing and Assimilation Systems for the Atmosphere, Oceans, and Land Surface (IOAS-AOLS)* Phoenix, AZ, Paper 5.2.

Fisher, R. A., 1925: *Statistical Methods for Research Workers.* Oliver and Boyd.

——, 1935: *The Design of Experiments.* Oliver and Boyd.

Kleijnen, J. P. C., 2008: *Design and Analysis of Simulation Experiments.* Springer.

Kleijnen, J. P. C., S. M. Sanchez, T. W. Lucas, and T. M. Cioppa, 2005: A User's Guide to the Brave New World of Designing Simulation Experiments. *INFORMS Journal on Computing*, **17,** 263-289.

Mamrosh, R. D., 1997: The Use of High-Frequency ACARS Soundings in Forecasting Convective Storms. *Weather and Forecasting Conference*, Phoenix, AZ, American Meteorological Society.

McKay, M. D., R. J. Beckman, and W. J. Conover, 1979: A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*, **21,** 239-245. doi: 10.2307/1268522.

Montgomery, D. C., 2013: *Design and Analysis of Experiments.* 8th ed. Wiley.

National Center for Atmospheric Research, 2016: Model Evaluation Tools Version 5.2 (METV5.2). User's Guide 5.2.

National Oceanic and Atmospheric Administration (NOAA), cited 2018: Real-Time Global Sea Surface Temperature (RTG_SST). [Available online at http://www.nco.ncep.noaa.gov/pmb/products/sst/.]

——, cited 2018: Global Forecast System (GFS). [Available online at https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs.]

——, cited 2018: National Operational Hydrologic Remote Sensing Center (NOHRSC), Snow Data Assimilation System (SNODAS). [Available online at https://www.nohrsc.noaa.gov/.]

——, cited 2018: Meteorological Assimilation Data Ingest System (MADIS). [Available online at http://madis.noaa.gov/.]

Navarro, D. J., cited 2018: Learning Statistics with R: A Tutorial for Psychology Students and Other Beginners. (Version 0.5). [Available online at https://CRAN.R-project.org/package=lsr.]

Penc, R. S., J. A. Smith, J. W. Raby, B. P. Reen, and R. E. Dumais, Jr., 2018a: Intercomparison of Seven Planetary Layer/Surface Layer Physics Schemes over Complex Terrain for Battlefield Situational Awareness Applications. *25th Conference on Numerical Weather Prediction, Joint with 29th Conference on Weather and Forecasting*, Denver, CO, American Meteorological Society, Paper 12B.5.

Penc, R. S., J. A. Smith, J. W. Raby, R. E. Dumais, Jr., B. P. Reen, and L. P. Dawson, 2018b: Intercomparison of 7 Planetary Boundary-Layer/Surface-Layer Physics Schemes over Complex Terrain for Battlefield Situational Awareness. Technical Report ARL-TR-8353.

R Core Team: R: A Language and Environment for Statistical Computing. [Available online at https://www.R-project.org/.]

Rahimi, A., T. Tavakoli, and S. Zahiri, 2014: Computational Fluid Dynamics (Cfd) Modeling of Gaseous Pollutants Dispersion in Low Wind Speed Condition: Isfahan Refinery, a Case Study. *Petroleum Science and Technology*, **32,** 1318-1326. doi: 10.1080/10916466.2011.653701.

Sacks, J., S. B. Schiller, and W. J. Welch, 1989a: Designs for Computer Experiments. *Technometrics*, **31,** 41-47. doi: 10.2307/1270363.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn, 1989b: Design and Analysis of Computer Experiments (Includes Comments and Rejoinder). *Statistical Science*, **4,** 409-435. doi: 10.1214/ss/1177012413.

Santner, T. J., B. J. Williams, and W. I. Notz, 2003: *The Design and Analysis of Computer Experiments.* Springer-Verlag.

Skamarock, W. C., and Coauthors, 2008: A Description of the Advanced Research WRF Version 3. NCAR Technical Note NCAR/TN-475+STR.

Smith, J. A., and R. S. Penc, 2015: A Design of Experiments Approach to Evaluating Parameterization Schemes for Numerical Weather Prediction. *Conference on Applied Statistics in Defense*, George Mason University, Fairfax VA, ASA Section on Statistics in Defense and National Security.

——, 2016: A Design of Experiments Approach to Evaluating Parameterization Schemes for Numerical Weather Prediction: Problem Definition and Proposed Solution Approach. *Joint Statistical Meetings Proceedings, Section on Statistics in Defense and National Security, Conference on Applied Statistics in Defense 2015***,** 4183-4192.

Smith, J. A., R. Penc, and J. W. Raby, 2017: Is There a Role for Statistical Design of Experiments in Numerical Weather Prediction? *97th Annual AMS Meeting, Joint with 28th Conference on Weather Analysis and Forecasting / 24th Conference on Numerical Weather Prediction*, Seattle, WA, Poster 616.

——, 2018: Statistical Design of Experiments in Numerical Weather Prediction: Emerging Results. *98th Annual AMS Meeting, Joint with 25th Conference on Probability and Statistics*, Austin, TX, American Meteorological Society, Paper 6.1.

Stein, U., and P. Alpert, 1993: Factor Separation in Numerical Simulations. *Journal of the Atmospheric Sciences*, **50,** 2107-2115. doi: 10.1175/1520-0469(1993)050<2107:FSINS>2.0.CO;2.

Vieira, H., Jr., S. Sanchez, K. H. Kienitz, and M. C. N. Belderrain, 2011: Generating and Improving Orthogonal Designs by Using Mixed Integer Programming. *European Journal of Operational Research*, **215,** 629-638. doi: 10.1016/j.ejor.2011.07.005.

Vieira, H., Jr., S. M. Sanchez, K. H. Kienitz, and M. C. N. Belderrain, 2013: Efficient, Nearly Orthogonal-and-Balanced, Mixed Designs: An Effective Way to Conduct Trade-Off Analyses Via Simulation. *Journal of Simulation*, **7,** 264-275. doi: 10.1057/jos.2013.14.

Wickham, H., cited 2018: Tidyverse: Easily Install and Load the 'Tidyverse'. R Package Version 1.2.1. [Available online at https://CRAN.R-project.org/package=tidyverse.]

Wickham, H., and G. Grolemund, 2016: *R for Data Science : Import, Tidy, Transform, Visualize, and Model Data.* O'Reilly Media.

Yates, F., 1937: The Design and Analysis of Factorial Experiments. Technical Communication No. 35.

Zhu, B., X. Wang, L. Tan, D. Zhou, Y. Zhao, and S. Cao, 2015: Optimization Design of a Reversible Pump-Turbine Runner with High Efficiency and Stability. *Renewable Energy*, **81,** 366-376. doi: 10.1016/j.renene.2015.03.050.