Peter C. Young

The Art and Craft of

Data-Based Mechanistic Modelling, Forecasting and Control

with detailed tutorial chapters on how it is applied successfully to examples of current topical interest: the Global Climate; the COVID Epidemic & Investment-Unemployment Dynamics in the USA; and a large tutorial appendix on modelling Pollution and Flow in Rivers

March 24, 2026

Springer Nature

Dedication

To my wife Wendy Anne, without whose continual support, help and encouragement this book and my previous publications since 1963 would never have been produced. The phrase ‘Art and Craft’ in the title of the book alludes to the analogy we have observed between the art and craft involved in her production of art textiles and the art and craft required in the research and development associated with the production of Data-Based Mechanistic models.

A Ruskin Lace embroidered *étui* by Wendy Anne Young.

Acknowledgements

Naturally, all of the work reported in the book has been influenced by the many colleagues with whom I have been associated for over half a century. Most of these are listed and thanked in my earlier 2011 book[1] and I would like to thank them all again. However, I am still stimulated by regular communications with my old and valued friends, to whom I would like to express a particular gratitude at this time (in alphabetical order): Professor Emeritus Geoff Allen; Professor Emeritus Keith Beven (Lancaster); Professor Emeritus Robert Fildes (Lancaster); Professor Emeritus Antonio Garcia-Ferrer (UAM, Madrid, Spain); Professor Hugues Garnier (University Lorraine, France); Professor Diego Pedregal (Universidad de Castilla-La Mancha, Spain); Professor Vic Solo, University of NSW Australia; Professor James Taylor (Lancaster); Reader Emeritus, Dr. Granville Tunnicliffe-Wilson (Lancaster); Dr. Wlodek Tych (Lancaster) and Professor Liuping Wang (RMIT, Melbourne). I am particularly grateful to Geoff Allen, who provided a lot of the background information and data used in chapter one, as well as continual and very helpful comments on the preparation of the chapter. Finally, another great influence on my life was my oldest friend Professor Emeritus Robert Spear (Berkeley), who sadly passed away in 2025. He and I communicated with each other continually from 1965 until 2025 and I will miss both him and our interaction on matters of mutual interest.

[1] see reference Young (2011) later in book

Preface

Although I evolved the concept of *Data-Based Mechanistic* (DBM) modelling over a number of years during the 1970s, I did not coin the phrase until 1993, then elaborating upon its meaning in subsequent years. It is a phrase that has current resonance following the much more recent use of the term 'data-driven' modelling. It is important to note, however, that the inherently stochastic DBM models are quite different to the purely black-box, data-driven models used in areas such as 'machine learning' and 'artificial intelligence'. These utilize various modelling, estimation and pattern recognition techniques, such as neural-network models, many of which do not quantify the uncertainty in their outputs and inferences.

In complete contrast to these heavily parameterized 'black-box' models, normally demanding large amounts of data and excessive computer resources, the stochastic DBM models are obtained using statistically efficient, optimal methods of model identification and parameter estimation, most often based on relatively small amounts of well selected data and requiring only modest, widely available computer resources. In the present book, for instance, these are an Apple iMac computer and computational routines from the 'CAPTAIN' Toolbox (see later in the Introductory section of the book) that run in the Matlab™ numeric computing environment (which also contains other tools that are useful in DBM modelling). And these DBM models are normally characterized by relatively few parameters that have an obvious, physically meaningful ('mechanistic') interpretation, in areas such as science, engineering, ecology, medicine and socio-economics.

It is important to note that DBM modelling is a general approach to the modelling of dynamic systems and the main aim of the book is to describe how this is applied in practice. Consequently, although Matlab is used as the numeric computing environment in the book, any other similar digital utilities, such as the the popular 'R' programming language, would be alternatives. Of course this would require computational tools similar to those in CAPTAIN and Matlab that are used in the book to be programmed by the user. Whatever computing environment is used, however, the 'art and craft' of the DBM modelling, i.e. the 'how-to-do-it', would remain the same. And the primary purpose of this book is to explain this in full detail.

The book starts with a short discussion that considers how the philosophical basis of DBM modelling fits within the context of previous scientific philosophy, from Isaac Newton to Karl Popper and Thomas Khun in the 20th Century. Then the ten main steps that constitute this DBM modelling approach are itemized and the associated algorithmic tools, all available in the free CAPTAIN Toolbox for use in the Maltlab™ software environment, are introduced. The main chapters in the book then follow, demonstrating in considerable tutorial detail how to apply these tools and understand the 'art and craft' of DBM modelling based on real time-series data.

The applications considered in this detailed manner are selected for two reasons. First and foremost, because they present problems in data-based modelling from real, monitored time series data that need to be handled carefully in order to obtain a well-identified dynamic model; a model with statistically well-estimated parameters that can be interpreted in appropriate and meaningful mechanistic terms. Consequently, these applications are very well-suited for tutorial purposes, demonstrating the care that must be taken in applying DBM modelling tools, or indeed any alternative modelling tools, when the data are problematic in some ways.

All of these tutorial examples involve the purely data-based statistical identification and estimation of physically meaningful differential equation models, based on time-series data that are freely available from the internet. Consequently, the resulting DBM models provide a mainly objective interpretation of these data and, unlike many approaches to scientific modelling that follow a 'hypothetico-deductive' form, they do not depend upon prior hypotheses about the nature of model; hypotheses that may prejudice the size and complexity of the resulting model. In other words, the differential equation model structure and the parameters that characterize this structure, all derive directly from the data alone. Whether this model is similar to that derived from prior hypotheses is considered only *after* this, when the DBM model may then sometimes be modified to a 'hypothetico-inductive' form, if this seems useful.

Fortunately the data used for these tutorial purposes in the chapters of the book are also interesting in a wider sense: they relate to four main areas of great topical interest to humanity: the global climate, the COVID pandemic, the macro-economics of the USA and finally, within a very large tutorial appendix, the hydrological systems that affect us all. The first three have been chosen because the results obtained via the DBM modelling carry another wider message and, lest this is lost in the tutorial detail, I feel this message should be considered briefly in this Preface.

The first two chapters deal with the modelling of globally-averaged warming in the World between 1850 to the present. The first chapter demonstrates how low dynamic order DBM differential equation models are able to closely emulate the *globally-averaged behaviour* of the extremely large *Atmosphere and Ocean General Circulation Models* (AOGCMs) that are so prominent in research on global-warming. In doing so, it reveals limitations in these models that appear to affect their predictive abilities, as well as questioning why the AOGCMs produced by different research centres have such different static and dynamic properties. This, in turn, leads to increased uncertainty about the future behaviour of global warming.

The second chapter focusses on the stochastic DBM modelling of the relationship between the *Total Radiative Forcing* (TRF) coming from various sources, including the much-debated effects of human activity, on the changes in globally-averaged surface temperature, as quantified in the *Global Temperature Anomaly* (GTA). Despite its low order, the DBM model is able to predict the various 'leveling' episodes that have occurred over the last 150 years, including the most recent one just after the turn of the Millenium. The better explanation of GTA is due to the presence of a quasi-periodic component in the model output that, speculatively but with good modelling support, could be linked with heat-exchange between the atmosphere and ocean. And it is this enhancement of the forecasting ability that allows the models to predict the short to medium term global warming better than the AOGCMs, which do not generate such globally-averaged behaviour.

The uncertainty associated with the COVID pandemic is well-known and the third chapter in the book confirms this, particularly in relation to the UK where poor governmental management led to increased COVID-related deaths, as well as gross waste and mishandling of public finances. One aspect of the reporting that may have led to uncertainty in the eyes of the general public was the 'basic reproductive number' that was promoted as a good measure of monitoring the progress of the epidemic and whether or not it was being kept under control. The chapter addresses this by suggesting an additional quick and straightforward approach to monitoring the progress of such epidemics based on the efficiently estimated rate of change in the recorded deaths (or indeed any other epidemic-related measure of its progress). This utilizes the powerful *Dynamic Harmonic Regression* (DHR) algorithm in the CAPTAIN Toolbox, which includes estimation and removal of the large weekly cycle in this series and so is robust to its effects.

The chapter goes on to show how, once again, DBM model-based forecasting can be helpful and it is utilized to produce quite reasonable multi-day forecasts of deaths caused by COVID infection. Finally, it is concluded that the most popular nonlinear *Susceptible, Infectious, Removed* (SIR) differential equation model developed and used by epidemiologists has good explanatory ability and that further research would be required to find a DBM model-based alternative. On the other hand, the SIR model requires data on the SIR components that are not easily accessible outside the epidemiological community. Consequently, the analytical procedures used in DBM modelling are applied to modify the SIR model so it can be estimated from the more easily available daily data on COVID deaths. This modified, 'hypothetico-inductive' model is then used to forecast the UK deaths due to COVID at various stages in the epidemic and highlight what activities, or lack thereof, appear to be associated with its dynamic behaviour.

Chapter 4 in the book considers an important aspect of the macro-economy in the USA from 1945 to the present. In particular, it identifies a low order DBM differential equation model that shows how the American unemployment percentage was affected by the changes in Government and Private capital investment over this period. It reveals how the model had to be modified around 1980, presumably after the neoliberal policies, such as those introduced into the USA by Ronald Reagan and in the UK by Margaret Thatcher, started to affect the economic systems of the Western World. The worst effects of these neoliberal policies, involving the deregulation of the private economy and privatization of public enterprises, really became apparent after the start of the new Millennium and led to the great crash of 2008.

After 2008, the model provides a quite good explanation of the data until, in 2015, a breakdown occurs, with the measured unemployment rate falling and the model's prediction showing little change, mainly because there is little recorded change in either government or private investment between 2015 and 2020 (mean values 17.5% and 17.7% of GDP, respectively). I hope that these DBM modelling results may help to shed further light on to what was happening to the US economy over this important period of time which Robert M. Solow, in a *New Republic* article (see `https://newrepublic.com/article/117429/capital-twenty-first-century-thomas-piketty-reviewed`), referred to as the "rich-get-richer dynamic", presumably harking back to the well known phrase "The rich get richer and the poor get poorer" attributed to the UK poet Percy Bysshe Shelley.

So, to conclude, I hope that this preface, *while emphasizing the predominant tutorial nature of this book*, also makes clearer the underlying messages that are revealed by the results of DBM modelling. In particular, I hope that:

- The first two chapters, through their climate-related tutorial examples, will help to throw some light on the limitations of the conventional climate models and reveal the value of the very much simpler DBM modelling in extracting, explaining and forecasting the dominant dynamics of the climate *at the global level*;
- The value of the simple yet effective monitoring approach suggested in the COVID pandemic chapter will be accepted by epidemiologists as an additional and easily available tool in their management of such epidemics; and that the other contributions on forecasting and modification of the SIR model may influence their thinking in these regards;
- And finally, if nothing else, the DBM modelling of unemployment in the USA will help to reveal to readers how and when this economy, that is so important to the other economies of the Western World, changed over the 21st Century. I am sure that any macro-economists, who happen to come across this book, will be able to explain very much better than I do what happened over the Trump Presidency, when the the DBM model breaks down and, despite very little change in public or private investment, unemployment continued to fall.

But I will be pleased if just a few of these hopes turn to reality; and that in learning from the tutorial content what is meant by the 'art and craft' of DBM modelling, which is the primary aim of the book, readers will also join with me in reflecting on what the topical examples used in this tutorial context may reveal about the nature of our uncertain World after 25 years of the 21st Century.

Peter Colin Young, M.A., Ph.D. (Cantab), Wh.F. Lancaster, U.K.
Professor Emeritus, Lancaster University January 2025

Contents

Introduction: Philosophy and Methodology 1
References 22

Part I DBM Emulation Modelling of Large Dynamic Simulation Models

Introduction to Part I 29

1 DBM Emulation Modelling 31
1.1 Introduction 31
1.2 DBM Emulation Model Identification and Estimation 32
1.2.1 Dominant Modes 35
1.2.2 Conventional Model Reduction and DBM Emulation 38
1.3 Initial Comments on AOGCM Computer models 40
1.4 Conventional Climate Emulation Modelling 43
1.4.1 The Climate Data 43
1.4.2 The MAGICC Model 45
1.4.3 DBM Emulation 47
1.4.4 Third Order DBM Model Emulation of MAGICC 49
1.4.5 The Choice of Input Signal 51
1.4.6 Stochastic Simulation and the Quantification of Uncertainty . . 53
1.4.7 Conclusions and Implications 57
1.4.8 DBM Emulation in a Wider Dynamic Systems Context 58
1.5 AOGCM Emulation Data 60
1.6 Conventional Regression-Based Emulation of AOGCMs 63
1.7 DBM Model Emulation of AOGCMs 70
1.7.1 CESM1044X: Third Order Model Emulation 71
1.7.1.1 Is there a drift in the piControl series? 80
1.7.1.2 How does the model perform when used with real data? 82
1.7.1.3 A problem that requires correction 83
1.7.2 CESM1044X: Fifth Order Model Emulation 85
1.7.3 Evaluating Estimation Convergence: Recursive Estimation 88

1.7.4 Final Comments on the CESM1044X Emulation Modelling Results 90
1.8 Other DBM Emulation Modelling Results 91
1.8.1 Emulation of the GFDLCM31pct2X Data Set 92
1.8.2 The Input Nonlinearity Exposed by the HadCM3L Data Set . . . 95
1.8.3 Scenario Simulation Analysis 100
1.9 Final Discussion and Conclusions 101
References 104

Part II The DBM Modelling of Dynamic Systems from Real Data

Introduction to Part II 111

2 DBM Modelling, Forecasting and Control illustrated by its application to globally-averaged climate data 113
2.1 Introduction 113
2.2 CO_2 Emissions and Atmospheric $pCO2$ 115
2.2.1 Model Identification and Estimation 117
2.3 Total Radiative Forcing and Surface Temperature 122
2.3.1 Global Climate Parameters 123
2.3.2 Initial Model Structure Identification 123
2.3.3 Final Model Identification and Estimation 130
2.3.4 Monte Carlo Analysis and Uncertainty 139
2.3.5 DHR Analysis and Uncertainty in the TRF 141
2.3.6 Two Alternatives to Simulink-Based Optimization 143
2.4 Mechanistic Interpretation of the DBM Model 146
2.4.1 The Main model: State $x_1(t)$ 146
2.4.2 The Main Model: Oscillatory State $x_2(t)$ 148
2.4.3 An Interesting Practical Analogy: Heat Exchange in a Test Chamber 151
2.4.4 The Complete DBM Model 153
2.5 Adding a Model for AMO 154
2.5.1 Higher Input Uncertainty 159
2.5.2 Is the AMO State $x_3(t)$ Part of the Model? 160
2.6 An Alternative Two-Input Model 161
2.6.1 Mechanistic Interpretation 167
2.7 Fractional Order Models 169
2.7.1 The Hypothetico-Deductive Model of Procyk et al. (2022) 170
2.7.2 Fractional Order DBM Models 173
2.7.3 A Single Component Fractional Model 178
2.7.4 Bayesian Estimation of the Equivalent Integer Order Model . . . 184
2.7.5 Conclusions 188
2.8 Scenario Simulation 189
2.9 Forecasting 189
2.9.1 Forecasting with a Kalman Filter 195

2.10 DBM Models and Emission Management Exploiting Control System Design 199
2.10.1 Implementation Using Simulink 210
2.11 Summary and Conclusions 214
References 223

3 DBM Monitoring, Forecasting and Modelling illustrated by its application to COVID-19 Pandemic Data 229
3.1 Introduction 229
3.2 The Rate-of-Change (RC) Metric and Its Estimation 232
3.2.1 The RC Metric 233
3.2.2 Analysis of the UK COVID-19 Data 234
3.2.3 The Situation in Other Countries 239
3.3 Modelling and Forecasting UK COVID-19 Deaths 240
3.4 Conventional Epidemiological Modelling 246
3.4.1 Modified SIR Model Updating and Adaptive Forecasting 254
3.5 Can an Alternative DBM Model of the UK Epidemic be Developed? . 259
3.5.1 A Simple SDP-Based Model of UK COVID-19 Deaths 259
3.5.2 More Complicated SDP-Based Models 263
3.6 Conclusions 265
References 267

4 DBM Modelling, Forecasting and Control Illustrated by its application to USA Investment-Unemployment data . . 269
4.1 Introduction 269
4.1.1 Previous Studies 276
4.2 A Continuous-Time Relativistic DBM Model 278
4.3 DBM Model Identification and Estimation 284
4.4 What happens after 2014? 289
4.5 Can the Model be Interpreted in Macro-economic Terms? 292
4.6 Could DBM Model-Based Control System Design Assist Macroeconomic Management? 294
4.6.1 Modelling Inflation Rate and Including This in the Design 300
4.7 Concluding Comments 304
References 311

Tutorial Appendix with Hydrological Examples 315
A.1 Mathematical and Statistical Preliminaries 317
A.1.1 Linear Transfer Function Models 317
A.1.1.1 First order TF model 317
A.1.1.2 Second order TF model 318
A.1.1.3 Laplace Transform 320
A.1.2 Linear Stochastic DBM Models 326
A.1.3 Model Structure Identification 329
A.2 DBM Modelling of Solute Transport in Rivers 333
A.2.1 The River Conder Tracer Data 333

A.2.2 Standard hypotheses Prior to DBM Modelling 335
A.2.3 The Simplest Hypothetico-Inductive Model 337
A.2.4 Inductive DBM Modelling 342
A.2.4.1 Initial model identification and estimation 343
A.2.4.2 A mechanistically unacceptable third order model . . . 345
A.2.4.3 Second order model estimation 347
A.2.5 Uncertainty and Monte Carlo Analysis 352
A.2.6 Mechanistic Interpretation of the Model 355
A.2.7 More Speculative Models 358
A.2.7.1 A speculative model 358
A.2.7.2 A combined ADZ-SDADZ model 360
A.2.7.3 Mechanistic Interpretation of the combined ADZ-SDADZ model 365
A.2.8 Extension to Other Flow Rates 367
A.2.9 Conclusions 368
A.3 Nonlinear DBM Models 368
A.3.1 TVP Models 369
A.3.2 SDP Models 370
A.4 Nonlinear Modelling and Forecasting River Flow 372
A.4.1 Standard Hypotheses Prior to DBM Modelling 373
A.4.2 Inductive DBM Modelling 375
A.4.3 TVP Model Estimation 376
A.4.4 SDP Model Estimation 379
A.4.5 Stochastic Validation 386
A.4.6 Mechanistic Justification 388
A.4.7 Handling Initial Conditions 389
A.4.8 Practical Utility: Flow Forecasting 391
A.4.8.1 Using a constant parameter model 391
A.4.8.2 Adaptive forecasting 396
A.4.9 Evaluation of a Standard Method for Decomposing Measured Flow 404
A.4.10 Kalman Filter Forecasting and Smoothing 407
A.4.11 Rainfall-Flow Nonlinear Modelling: Final Comments 409
A.4.12 A Final Hydrological Example using the New rivcbjdd Tool in CAPTAIN 410
A.5 Dynamic Harmonic Regression (DHR) 414
A.5.1 A Trend or Not a Trend? 420
A.6 The Advantages of Continuous-Time Modelling 423
A.7 Matlab Coding and CAPTAIN Tools 427
A.7.1 Coding for Estimation of the ADZ-SDADZ Model 427
A.7.2 Final Canning Model Estimation 430
A.8 The CAPTAIN 'help' listings 431
A.8.1 The rivcbjid Tool in CAPTAIN 432
A.8.2 The rivcbj Tool in CAPTAIN 433
A.8.3 The srivcarma Tool in CAPTAIN 434
A.8.4 The rivcbjfdid and rivcbjfd Tools in CAPTAIN 435

A.8.5 The New rivcbjdd Tool in CAPTAIN 436
A.8.6 The kalmanfis Tool and the sdadz Model 437
A.8.7 The dhr and dhropt Tools in CAPTAIN 438
References 439

Basic Mathematical and Statistical Background 445
B.1 Matrix Algebra 445
B.1.1 Matrices 445
B.1.2 Vectors 446
B.1.3 Matrix Addition (or Subtraction) 446
B.1.4 Matrix or Vector Transpose 446
B.1.5 Matrix Multiplication 447
B.1.6 Determinant of a Matrix 448
B.1.7 Partitioned Matrices 449
B.1.8 Inverse of a Matrix 450
B.1.9 Quadratic Forms 451
B.1.10 Positive Definite or Semi-Definite Matrices 451
B.1.11 The Rank of a Matrix 451
B.1.12 Differentiation of Vectors and Matrices 451
B.1.13 Instrumental Variable Estimation 454
B.1.14 Cholesky Decomposition 455
B.1.15 Singular Value Decomposition (SVD) 455
B.2 Statistics and Probability 456
B.2.1 Discrete Random Variables 456
B.2.2 Law of large Numbers 457
B.2.3 Discrete Random Vectors 458
B.2.4 Conditional Probabilities 458
B.2.5 Continuous Random Variables and Vectors 459
B.2.6 The Normal or Gaussian Density Function 460
B.2.7 Properties of Estimators 461
B.2.8 The Likelihood Function and Maximum Likelihood Estimation 461
B.2.9 The Cramer-Rao Lower Bound 463
B.2.10 Maximum Likelihood Estimators: the Vector Case 463
B.3 Time Series 464
B.3.1 Gauss-Markov Random Processes 466
B.3.2 The State Space Model of a Linear, Discrete-time, Stochastic Dynamic System 467
B.3.3 The Discrete-Time Transfer Function Model 468
B.3.4 Continuous-Time, Stochastic Dynamic Models 472
B.3.5 Hybrid Stochastic Dynamic Models 473
B.3.6 Multivariable (Multi-Input, Multi-Output) TF Models 475
B.3.6.1 Conversion between continuous and discrete-time . . . 475
B.3.6.2 Frequency domain representations of TF models 477
B.3.7 Differentiation of a TF with Respect to a Given Parameter 480
B.3.8 A Simple Introduction to Monte Carlo Simulation 481
B.3.8.1 The generation of random realizations 481

References 484