Modelling Long-Term Persistence in Hydrological Time Series

Mark Andrew Thyer B.E. (Civil) (Hons. I)

A thesis submitted for the degree of Doctor of Philosophy

at The University of Newcastle

December 2000

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

Mark A. Thyer

“If we knew what we were doing
it would not be called research”

Albert Einstein

Acknowledgements

The completion of a PhD thesis represents a major undertaking and it would not have been possible without the support and advice from my family, friends and colleagues.

It was been an absolute pleasure to have Associate Professor George Kuczera as my supervisor. His enthusiasm and wisdom are a continual source of inspiration. Countless times I have walked into George's office with a seemingly insurmountable problem and then emerged (usually hours later) fully motivated to tackle this thesis.

I would like to gratefully acknowledge the financial support provided by the industry sponsors of this project, Sydney Catchment Authority and Hunter Water Corporation. Personal thanks go to Mark Powell from SCA for answering my many questions and to Janice Lough from AIMS for providing the Burdekin data.

Special thanks go to my friends, Craig Gardiner - my SIB brewing partner, Kerry McIntosh, Greg Hancock - the Dad of the postgrads, Jeff (and Wendy) Walker for finding some desk space at NASA and Emily Slatter for providing me with a home during the throes of writing up this thesis. The times I have shared with these good people have made my stay in Newcastle very memorable.

Thank you to the postgrads and academics, Andrew Frost, Lyndon Bell, Scott Wooldridge, Heber Sugo and Dr Stewart Franks for helpful advice and fun distractions. Thank you also to the support staff of my Department, they do an excellent job.

Jodie Berkefeld deserves a special mention. She was my friend, my housemate, my girlfriend and now my friend again. I think she deserves a PhD for putting up with me over the last 3 years and 9 months. Thank you for everything, Jodie.

I cannot express my appreciation for the continual love and support provided by my family. To my parents, Alan and Tricia, and my sister, Jo-Anne, thank you for always believing in me, I am forever grateful.

And finally, thank you to all my friends. You make my life what it is, an enjoyable journey.

Table of Contents

Chapter 4 - Hydrological Data

Chapter 5 - Calibration Procedure for Single Site HSM Model

Chapter 6 - Developing an Interpretative Framework for HSM Model using Synthetic Data

Chapter 7 - Single Site HSM Model Calibration Results

Chapter 10 - Calibration Procedure for Multi-site HSM Model

Chapter 11 - Multi-site HSM Model Calibration Results

Chapter 12 - Conclusions

Appendix A - Estimation of Hurst Coefficient

Appendix B - Filling-In and Correction of Hydrological Data

Appendix C - Markov Chain Monte Carlo Methods: An Overview

Appendix D - Sampling Distributions for the Gibbs Sampler

Appendix E - Original Single Site Prior Specification

Appendix F - Verification of Single Site HSM Model Calibration Procedure

Appendix G - Comparison of Priors to Posteriors for Single Site HSM Model

Abstract

The hidden state Markov (HSM) model is introduced as a new conceptual framework for modelling long-term persistence in hydrological time series. Unlike the stochastic models currently used, the conceptual basis of the HSM model can be related to the physical processes that influence long-term hydrological time series in the Australian climatic regime. A Bayesian approach was used for model calibration. This enabled rigourous evaluation of parameter uncertainty, which proved crucial for the interpretation of the results. Applying the single site HSM model to rainfall data from selected Australian capital cities provided some revealing insights. In eastern Australia, where there is a significant influence from the tropical Pacific weather systems, the results showed a weak wet and medium dry state persistence was likely to exist. In southern Australia the results were inconclusive. However, they suggested a weak wet and strong dry persistence structure may exist, possibly due to the infrequent incursion of tropical weather systems in southern Australia. This led to the postulate that the tropical weather systems are the primary cause of two-state long-term persistence. The single and multi-site HSM model results for the Warragamba catchment rainfall data supported this hypothesis. A strong two-state persistence structure was likely to exist in the rainfall regime of this important water supply catchment. In contrast, the single and multi-site results for the Williams River catchment rainfall data were inconsistent. This illustrates further work is required to understand the application of the HSM model. Comparisons with the lag-one autoregressive [AR(1)] model showed that it was not able to reproduce the same long-term persistence as the HSM model. However, with record lengths typical of real data the difference between the two approaches was not statistically significant. Nevertheless, it was concluded that the HSM model provides a conceptually richer framework than the AR(1) model.

Notation

Following will be a description of the notation commonly used in this thesis. Unless otherwise stated the convention is to refer to scalars using lowercase italic nonbold type, vectors using lowercase nonitalic bold type and matrices using uppercase nonitalic bold type.

Probability Notation

?	general term for model parameters
y	general term for the observed data
y_t	observed data scalar at time t
$\mathbf{y_t}$	observed data vector at time t
Y_N	general term for the set of observed scalar data, $Y_N = \{y_1, \ldots, y_n\}$
$\mathbf{Y_N}$	general term for the set of observed vector data, $\mathbf{Y_N} = \{\mathbf{y_1}, \ldots, \mathbf{y_n}\}$
$p(?, y)$	joint probability density of model parameters θ and observed data y
$p(?\|y)$	conditional probability density of model parameters θ conditioned on the observed data y, also referred to as the posterior probability density
$p(?)$	probability density of the model parameters θ, also known as the prior probability density
$p(y\|?)$	conditional probability density of observed data y conditioned on the model parameters θ, also known as the likelihood function
$p(\mathrm{y}^{rep}\|Y_N)$	posterior predictive distribution of the replicated data y^{rep} simulated given the model parameters conditioned on the observed data

Probability Distribution Notation

$U(0,1)$	uniform probability distribution with limits 0 and 1
$N(\mu, s^2)$	univariate Gaussian distribution with scalar mean μ and variance s^2 (σ is standard deviation)

$N_r(\boldsymbol{\mu}, \mathbf{S})$	multivariate Gaussian distribution with r dimensions, mean vector $\boldsymbol{\mu}$ of length r and $r \times r$ symmetric positive definite covariance matrix $\mathbf{S}$, the inverse of the covariance matrix, $\mathbf{S}^{-1}$ is referred to as a precision matrix
$Inv\text{–}?^2(?,s^2)$	scaled inverse-chi-square distribution with degrees of freedom ? and scale s^2
$Beta(\alpha,\beta)$	Beta distributions with parameters α and β
$Gamma(\alpha,\beta)$	Gamma distribution with shape α and inverse scale β
$W_r(\nu, \mathbf{W})$	Wishart distribution with dimension r, degrees of freedom ν and $r \times r$ symmetric positive definite scale matrix $\mathbf{W}$

Prior and Posterior Parameter Notation

In general, the prior parameters are denoted by the subscript o and the posterior parameters are by the subscript n.

μ_0, τ_0^2	prior mean and variance for state mean
μ_n, τ_n^2	posterior mean and variance for state mean
ν_0, σ_0^2	prior degrees of freedom and scale for state variance
ν_n, σ_n^2	posterior degrees of freedom and scale for state variance
$\boldsymbol{\mu}_\mathbf{0}, \kappa_0$	prior mean and number of prior measurements on $\mathbf{S}$ scale for state mean vector
$\boldsymbol{\mu}_\mathbf{n}, \kappa_n$	posterior mean and number of posterior measurements on $\mathbf{S}$ scale for state mean vector
$\nu_0, \mathbf{W_0}$	prior degrees of freedom and prior scale matrix for the state precision matrix, $\mathbf{W_0}$ can also be thought of as the prior precision matrix
$\nu_n, \mathbf{W_n}$	posterior degrees of freedom and posterior precision matrix for the state precision matrix

Hidden State Markov Model Notation

$\mathbf{P}$	Markovian state transition probability matrix
p_{ij}	state transition probability that represents the probability of moving from state i to state j where $i, j \in W, D$
W	wet state
D	dry state
s_t	hidden state (wet or dry) at time t
S_N	time series of hidden states, $S_N = \{s_1, \ldots, s_n\}$
Y^W	observed data that has been classified in the wet state
Y^D	observed data that has been classified in the dry state

Two State Persistence Structure Notation

$E(SRT_k)$	expected state residence time in state k, calculated from $1/p_{TRANS}$, where p_{TRANS} is probability of transition out of state k
SPS_k	strength of the persistence structure in state k Defined as either weak (W), medium (M), strong (S), or very strong (VS) using the expected state residence time, refer to Table 6.1
$WADSI$	wet and dry separation index, calculated from $(\mu_W - \mu_D)/\sqrt{\sigma_W^2 + \sigma_D^2}$
TSP	two state persistence structure notation, for the single site it is defined as $[SPS_W, SPS_D, WADSI]$ whereas for the multi-site there is a vector of $WADSI$ values, one for each site

ARMA model notation

μ	scalar mean of time series
ϕ_j	autoregressive parameter at lag j
α_j	moving average parameter at lag j
ε_t	error term, an uncorrelated Gaussian random variable, $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$

Markov Chain Monte Carlo Derivation Notation

This notation is primarily used in Appendix C. The usual convention of denoting vectors as nonitalic bold type is not used in Appendix C.

x^i	a sample x, possibly scalar, generally a vector, generated by a MCMC method at iteration i
$\pi(\cdot)$	"target" probability density
$\pi^*(\cdot)$	target probability distribution
$P(A\|x)$	Markov chain transition probability kernel, that represents the probability of making a transition from a point x to a point in the region defined by A
$q(y\|x)$	MCMC candidate generating density, represents the probability of generating a sample y given the current state of the process x
$p(y\|x)$	function that gives the probability of a transition from x to y
$r(x)$	probability that chain remains at x, $r(x) = 1 - \int p(y\|x)\, dy$
$\alpha(y\|x)$	probability of move from x to y
$\delta(x)$	Dirac delta function, with property that $\int_{-\infty}^{\infty} \delta(x)\, dx = 1$, with $\delta(x) \neq 0$ when $x \in A$, for some region A and $\delta(x) = 0$ when $x \notin A$
b	number of iterations required to "warm-up" the MCMC chain before convergence is achieved