



Título del Trabajo Fin de Máster:

**Evaluación de la Incertidumbre de los  
Parámetros y su efecto sobre la  
Incertidumbre Predictiva de un Modelo  
Hidrológico Distribuido, mediante un  
algoritmo *Markov Chain Monte Carlo* de  
tipo Adaptativo.**

Intensificación:

**RECURSOS HÍDRICOS**

Autor:

***HERNÁNDEZ LÓPEZ, MARIO R.***

Director:

***DR. FRANCÉS GARCÍA, FÉLIX***

Fecha: **Noviembre 2012**





Título del Trabajo Fin de Máster:

**Evaluación de la Incertidumbre de los Parámetros y su efecto sobre la Incertidumbre Predictiva de un Modelo Hidrológico Distribuido, mediante un algoritmo *Markov Chain Monte Carlo* de tipo Adaptativo.**

Autor: **HERNÁNDEZ LÓPEZ, MARIO R.**

<b>Tipo</b>	A <input type="checkbox"/> B <input checked="" type="checkbox"/>	<b>Lugar de Realización</b>	<b>VALENCIA</b>
<b>Director</b>	<b>DR. FÉLIX FRANCÉS GARCÍA</b>	<b>Fecha de Lectura</b>	<b>Noviembre, 2012</b>
<b>Codirector1</b>			
<b>Codirector2</b>			
<b>Tutor</b>			

**Resumen:**

Uno de los principios básicos de la Directiva Marco del Agua (DMA, 2000/60/CE) es "contribuir a la consecución de los objetivos de preservar, proteger y mejorar la calidad del medio ambiente con el uso prudente y racional de los recursos naturales, basándose en el principio de precaución". En la medida en que el principio de precaución tiene como objetivo proteger a los humanos y al medio ambiente contra los riesgos inciertos, por medio de medidas preventivas, dicha tarea no puede llevarse a cabo sin la incorporación de la evaluación de la incertidumbre en el proceso de toma de decisiones. La evaluación de la incertidumbre de los modelos de simulación es por lo tanto, importante, cuando los modelos se utilizan para apoyar las decisiones sobre la gestión del agua.

Los modelos hidrológicos suelen contener parámetros que no pueden ser determinados, mediante el empleo de medidas directas de los mismos en campo. El proceso de calibración del modelo, permite el ajuste de esos parámetros dentro de unos rangos razonables, mediante el mejor ajuste posible entre los resultados de ciertas variables del modelo, y las observaciones que sobre ellas se haya realizado. Sin embargo, los valores calibrados pueden verse afectados por correlaciones entre los parámetros y por otras características estadísticas de los residuos. Esta circunstancia conduce a la llamada Equifinalidad de las soluciones. Por otra parte, en modelos hidrológicos complejos son varias las fuentes de incertidumbre, y esta incertidumbre acaba propagándose a los resultados del modelo. Los datos de entrada del modelo, su estructura y el valor de los parámetros son las principales fuentes de incertidumbre.

Entonces, las predicciones de un modelo deberían ser presentadas con un intervalo de confianza, en lugar de como un valor "cierto".

El objetivo de esta tesina es explorar algunas técnicas de análisis de sensibilidad e incertidumbre sobre los parámetros, prestando especial atención a los algoritmos MCMC. Así mismo se amplían las capacidades de análisis del modelo conceptual de simulación hidrológica TETIS mediante la integración de dicho modelo hidrológico (programado en Fortran) con el paquete de R denominado FME, Flexible Modeling Environment.

El trabajo se completa, con la aplicación de la herramienta desarrollada a un caso práctico: El análisis de incertidumbre de los parámetros del modelo hidrológico de la cuenca vertiente al embalse de La Baells (Alto Llobregat). Seguidamente se realiza un análisis de sensibilidad global del modelo, cuyo resultado nos muestra la incertidumbre que los parámetros transmiten a la variable respuesta Caudal. Así mismo se estima mediante un modelo de error, la parte de incertidumbre de la variable respuesta, que no es debida a los parámetros.



**Abstract:**

One of the basic principles of the Water Framework Directive (WFD, 2000/60/EC) is "to contribute to the achievement of the objectives of preserving, protecting and improving the quality of the environment with prudent and rational use of natural resources, based on the precautionary principle". Since the precautionary principle is to protect humans and the environment against uncertain risks, through preventive measures, this task can not be accomplished without incorporating the assessment of uncertainty in the process decision making. The uncertainty evaluation of simulation models is therefore important, when models are used to support decisions on water management.

Hydrological models often contain parameters that can not be determined, by using direct field measures. The model calibration process allows adjustment of these parameters within a reasonable range, using the best possible fit between the simulated variables through the model and the observations. However, the calibrated values are affected by correlations between parameters and other statistical characteristics of the residuals. This circumstance leads to the widely known Equifinality. Moreover, in complex hydrological models there are several sources of uncertainty, and this uncertainty eventually spread to the model results. The model inputs, data, structure and value of the parameters are the main sources of uncertainty.

Therefore, the predictions of a model should be presented along with a confidence interval, rather than as a "true" value.

The objective of this thesis is to explore some techniques of sensitivity and uncertainty analysis on parameters, with particular attention to the MCMC algorithms. The aim is also to extend the analysis capabilities of the hydrological conceptual model TETIS by coupling the hydrological model (programmed in Fortran) with the R package called FME, Flexible Modeling Environment. The work is completed with the application of the developed tool to a case study: Uncertainty analysis of hydrological model parameters of the watershed to the reservoir of La Baells (Alto Llobregat). This followed by a global sensitivity analysis of the model, shows the parameters uncertainty transmitted to the response variable, that is the discharge flow. The uncertainty of the simulated variable, which is not due to the parameter, is estimated using an error model.

**Resum:**

Un dels principis bàsics de la Directiva Marco de l'Aigua (DMA, 2000/60/CE) és "contribuir a la consecució dels objectius de preservar, protegir i millorar la qualitat del medi ambient amb l'ús prudent i racional dels recursos naturals, basant-se en el principi de precaució". En la mesura que el principi de precaució té com a objectiu protegir als humans i al medi ambient contra els riscos incerts, per mitjà de mesures preventives, la dita tasca no pot dur-se a terme sense la incorporació de l'avaluació de la incertesa en el procés de presa de decisions. L'avaluació de la incertesa dels models de simulació és per tant, important, quan els models s'utilitzen per a recolzar les decisions sobre la gestió de l'aigua.

Els models hidrològics solen contindre paràmetres que no poden ser determinats, per mitjà de la presa de mesures directes dels mateixos en camp. El procés de calibratge del model, permet l'ajust d'eixos paràmetres dins d'uns rangs raonables, per mitjà del millor ajust possible entre els resultats de certes variables del model, i les observacions que sobre elles s'haja realitzat. No obstant això, els valors calibrats poden veure's afectats per correlacions entre els paràmetres i per altres característiques estadístiques dels residus. Esta circumstància condueix a la denominada Equifinalitat de les solucions. D'altra banda, en models hidrològics complexos són diverses les fonts d'incertesa, i esta incertesa acaba propagant-se als resultats del model. Les dades d'entrada del model, la seua estructura i el valor dels paràmetres són les principals fonts d'incertesa. Llavors, les prediccions d'un model haurien de ser presentades amb amb un interval de confiança, en compte de com un valor "cert". L'objectiu d'esta tesina és explorar algunes tècniques d'anàlisi de sensibilitat i incertesa sobre els paràmetres, prestant especial



atenció als algoritmes MCMC. Així mateix s'amplien les capacitats d'anàlisi del model conceptual de simulació hidrològica TETIS per mitjà de la integració del dit model hidrològic (programat en Fortran) amb el paquet de R denominat FME, Flexible Modeling Environment. El treball es completa, amb l'aplicació de la ferramenta desenvolupada a un cas pràctic: L'anàlisi d'incertesa dels paràmetres del model hidrològic de la conca vessant a l'embassament de La Baelis (Alt Llobregat) . A continuació es realitza una anàlisi de sensibilitat global del model, el resultat del qual, mostra la incertesa que els paràmetres transmeten a la variable resposta Cabal. Així mateix s'estima per mitjà d'un model d'error, la part d'incertesa de la variable resposta, que no és deguda als paràmetres.

**Palabras clave:** *Modelación hidrológica distribuida/ TETIS/ Incertidumbre/ Estadística Bayesiana/ Métodos Markov Chain Monte Carlo.*

**Key words:** *Distributed hydrological modeling/ TETIS/ Uncertainty/ Bayesian Statistics/ Markov Chain Monte Carlo Methods.*

**Paraules Claus:** *Modelació hidrològica distribuïda/ TETIS/ Incertesa/ Estadística Bayesiana/ Mètodes Markov Chain Monte Carlo.*



*A la memoria de Mario, mi Padre, mi guía en el pasado.*

*A Noelia, por... tanto.*

*A Sarah, nuestro futuro.*

*“Así como una jornada bien empleada produce un dulce sueño, una vida bien usada causa una dulce muerte.”*

Leonardo Da Vinci (1452-1519). Pintor, escultor e inventor.

*“No me siento obligado a creer que un dios que nos ha dotado de inteligencia, sentido común y raciocinio, tuviera como objetivo privarnos de su uso.”*

Galileo Galilei (1564-1642). Astrónomo, filósofo, matemático y físico.





## AGRADECIMIENTOS

El trabajo presentado en este estudio ha sido desarrollado dentro del *Grupo de Investigación de Modelación Hidrológica y Ambiental- GIMHA* ([http://lluvia.dihma.upv.es/index\\_es.html](http://lluvia.dihma.upv.es/index_es.html)), conformado por investigadores de la Universidad Politécnica de Valencia.

Todo mi agradecimiento al Dr. Félix Francés García, director del GIMHA, que una vez más me dio su confianza para incorporarme a su equipo de trabajo y que en los inicios de esta andadura, contribuyó de forma decisiva con sus ideas e intuición, a que el que escribe, iniciase un trabajo de estas características.

Igualmente muestro mi aprecio en estas líneas, a los compañeros del grupo de investigación GIMHA, con los que he compartido horas de trabajo bajo las luces de neón. Juan Camilo, Alicia García, Giamba Bussi, Ismael Orozco, Joanna, Cata, Jesús, Marta, Chiara, Miguel, Joaquín,...a todos ellos y a los que me olvido, gracias. Con ellos he compartido muchos y necesarios cafés, durante los que hemos mantenido conversaciones, algunas triviales que alegraban el espíritu, y otras no tan triviales que alimentaban el ansia por seguir aprendiendo de todos ellos, los conocimientos que atesoran, haciendo del grupo un lugar excepcional para explotar la Creatividad y el Ingenio.

A César Elis por el ánimo y apoyo recibidos.

A Noelia, mi mujer, mi apoyo y mi brújula en el camino, durante estos 12 años. Gracias por estar a mi lado. Cuánto he de agradecerte que así haya sido.

He de agradecer también la amabilidad que han mostrado los funcionarios de las instituciones que han prestado su información. A la Agencia Estatal de Meteorología- AEMET, al Instituto Geológico y Minero de España- IGME y a la Agencia Catalana del Agua- ACA, por la disponibilidad de sus datos, los cuales han permitido llevar a cabo este trabajo.

Al proyecto SCARCE: *Assessing and Predicting Effects on Water Quantity and Quality in Iberian rivers caused by Global Change* (ref.: CSD2009-00065), por ser una de las bases esenciales de esta investigación.





## TABLA DE CONTENIDOS

<b>1.- INTRODUCCIÓN .....</b>	<b>1</b>
1.1.- MOTIVACIÓN Y OBJETIVOS.....	1
1.2.- JUSTIFICACIÓN DEL TRABAJO .....	2
1.3.- ESTRUCTURA DEL DOCUMENTO .....	3
<b>2.- MODELACIÓN HIDROLÓGICA E INCERTIDUMBRE.....</b>	<b>5</b>
2.1.- INTRODUCCIÓN.....	5
2.2.- MODELOS HIDROLÓGICOS .....	6
2.3.- SOBRE LA IDENTIFICABILIDAD .....	10
2.4.- SOBRE LA INCERTIDUMBRE.....	13
2.4.1.- Generalidades.....	13
2.4.2.- Fuentes de Incertidumbre.....	15
2.5.- EVALUACIÓN DE MODELOS.....	18
<b>3.- BASES METODOLÓGICAS .....</b>	<b>25</b>
3.1.- ESTADÍSTICA BAYESIANA COMPUTACIONAL.....	25
3.1.1.- Introducción.....	25
3.1.2.- Formulación del Teorema de Bayes.....	28
3.1.3.- Obtención de la posterior .....	29
3.1.4.- Estimación Máximo-Verosímil y Estimación Bayesiana.....	30
3.2.- CADENAS DE MARKOV .....	35
3.2.1.- Introducción.....	35
3.2.2.- Cadenas de Markov invariantes en el tiempo y con espacio de estados finito.....	35
3.2.3.- Distribución de Equilibrio y Ecuación de Estado Estacionario.....	36
3.2.4.- Clasificación de estados de una cadena de Markov .....	37
3.2.5.- Ergodicidad de las cadenas de Markov .....	38
3.2.6.- Muestreo de una cadena de Markov.....	39
3.2.7.- Cadenas de Markov Reversibles en el tiempo. Condición de Balance Detallado.....	40
3.2.8.- Cadenas de Markov con espacio de estados continuo .....	41
3.3.- MUESTREO DE LA POSTERIOR. EL ALGORITMO MH .....	42
3.3.1.- Introducción.....	42
3.3.2.- Ecuación de Estado Estacionario en un espacio de parámetros continuo.....	43
3.3.3.- Reversibilidad y ecuación de balance .....	44
3.3.4.- El algoritmo Metropolis-Hastings.....	45
3.4.- INFERENCIA ESTADÍSTICA A PARTIR DE MUESTRAS MCMC.....	49
3.4.1.- Introducción.....	49
3.4.2.- Propiedades de mezclado de la cadena.....	50
3.4.3.- Obtención de una muestra aproximadamente aleatoria .....	51
3.4.4.- Testeo de la cadena de Markov.....	55
3.5.- UN ALGORITMO MCMC ADAPTATIVO: DRAM.....	56
3.5.1.- La adaptación como mejora al algoritmo MH.....	56
3.5.2.- Delayed Rejection Adaptive Metropolis (DRAM).....	57
3.5.3.- Implementación de DRAM.....	58



<b>4.- APLICACIÓN A UNA CUENCA DE ALTA MONTAÑA: LA BAEELS, ALTO LLOBREGAT (BARCELONA).....</b>	<b>62</b>
4.1.- HERRAMIENTAS EMPLEADAS EN EL ANÁLISIS .....	62
4.1.1.- El modelo hidrológico TETIS.....	62
4.1.2.- El paquete estadístico FME.....	72
4.2.- IMPLEMENTACIÓN DEL MODELO HIDROLÓGICO .....	73
4.2.1.- Cuenca hidrológica .....	74
4.2.2.- Clima.....	76
4.2.3.- Precipitación y temperatura .....	77
4.2.4.- Estimación de parámetros espaciales .....	78
4.3.- APLICACIÓN PRÁCTICA .....	80
4.3.1.- Planteamiento del problema .....	80
4.3.2.- Objetivos de la aplicación práctica .....	82
4.3.3.- Análisis de Sensibilidad Local.....	84
4.3.4.- Incertidumbre de los parámetros .....	88
4.3.5.- Efecto de la incertidumbre paramétrica sobre la predicción del modelo .....	100
4.3.6.- Sensibilidad Global del modelo a los parámetros.....	103
4.3.7.- Incertidumbre total en la predicción del modelo.....	107
4.3.8.- Verificación del modelo de error.....	109
<b>5.- CONCLUSIONES.....</b>	<b>112</b>
5.1.- SOBRE LA METODOLOGÍA .....	113
5.1.1.- Debilidades .....	113
5.1.2.- Fortalezas .....	113
5.2.- SOBRE LOS RESULTADOS .....	114
<b>6.- APORTES DEL TRABAJO Y FUTURAS LINEAS DE INVESTIGACIÓN .....</b>	<b>116</b>
6.1.- APORTES DEL TRABAJO .....	116
6.2.- LINEAS FUTURAS .....	116
<b>7.- REFERENCIAS BIBLIOGRÁFICAS .....</b>	<b>119</b>



## LISTA DE FIGURAS

FIGURA 2.1. CLASIFICACIÓN GENERAL DE MODELOS HIDROLÓGICOS.....	9
FIGURA 3.1. SUPERFICIE DE DENSIDAD DE OBSERVACIONES (FUENTE: BOLSTAD 2010) .....	31
FIGURA 3.2. SUPERFICIE DE DENSIDAD DE OBSERVACIONES CON LA FUNCIÓN DE VEROSIMILITUD (FUENTE: BOLSTAD 2010) .....	31
FIGURA 3.3. DISTRIBUCIÓN CONJUNTA DE LOS PARÁMETROS Y LAS OBSERVACIONES. EN EL PLANO LATERAL SE DIBUJA LA PRIOR DE LOS PARÁMETROS (FUENTE: BOLSTAD 2010) .....	32
FIGURA 3.4. DISTRIBUCIÓN POSTERIOR DE LOS PARÁMETROS (FUENTE: BOLSTAD 2010) .....	33
FIGURA 3.5. CLASIFICACIÓN DE ESTADOS DE UNA CADENA DE MARKOV (FUENTE: BOLSTAD 2010) .....	38
FIGURA 3.6. CONCEPTO DE ENSEMBLE DE UN PROCESO ALEATORIO (FUENTE: BRAS AND RODRÍGUEZ-ITURBE 1985) .....	39
FIGURA 3.7. RESUMEN DE LAS 6 PRIMERAS EXTRACCIONES DE UNA CADENA RANDOM WALK, Y ASPECTO DE LA TRAZA DE LAS 1000 SIMULACIONES (FUENTE: BOLSTAD 2010).....	49
FIGURA 3.8. RESUMEN DE LAS 6 PRIMERAS EXTRACCIONES DE UNA CADENA CON CANDIDATA INDEPENDIENTE Y ASPECTO DE LA TRAZA DE LAS 1000 SIMULACIONES (FUENTE: BOLSTAD 2010).....	49
FIGURA 3.9. MEZCLADO DE LA CADENA BUENO (IZQ) Y MALO (DER). (FUENTE: WEB).....	51
FIGURA 4.1 ESQUEMA CONCEPTUAL DE TANQUES A NIVEL DE CELDA DEL MODELO TETIS .....	63
FIGURA 4.2. MOVIMIENTO HORIZONTAL PROPUESTO POR EL MODELO TETIS .....	66
FIGURA 4.3. ELEMENTOS DEL SISTEMA, ÁREAS UMBRALES Y COMPONENTES DE LA ESCORRENTÍA (FUENTE: MONTOYA 2008) .....	67
FIGURA 4.4. PERFIL LONGITUDINAL DEL RÍO LLOBREGAT (FUENTE: WEB) .....	75
FIGURA 4.5. LOCALIZACIÓN DE LA CUENCA DE LA BAELLS. ....	76
FIGURA 4.6: SEGUIMIENTO DE LOS PARÁMETROS. SENSIBILIDAD LOCAL POR MONTE CARLO.....	86
FIGURA 4.7. TRAZAS DE LA SIMULACIÓN MCMC .....	92
FIGURA 4.8. TESTEO DE ALEATORIEDAD DE LA MUESTRA MEDIANTE LA INSPECCIÓN DE LOS CORRELOGRAMAS.....	94
FIGURA 4.9. POSTERIOR DE LOS PARÁMETROS DEL MODELO LA BAELLS.....	99
FIGURA 4.10. INCERTIDUMBRE DE LOS PARÁMETROS DE LA BAELLS. ....	100
FIGURA 4.11. INCERTIDUMBRE PREDICTIVA DEBIDA A LOS PARÁMETROS. ....	101
FIGURA 4.12. INCERTIDUMBRE PREDICTIVA DEBIDA A LOS PARÁMETROS. DETALLE ENTRE LOS DÍAS 100 Y 300 DE SIMULACIÓN. ....	102
FIGURA 4.13. INCERTIDUMBRE PREDICTIVA DEBIDA A LOS PARÁMETROS. DETALLE ENTRE LOS DÍAS 130 Y 160 DE SIMULACIÓN. ....	102
FIGURA 4.14. SENSIBILIDAD GLOBAL A UN PARÁMETRO.....	103
FIGURA 4.15. EFECTO RELATIVO DE CADA PARÁMETRO SOBRE LA INCERTIDUMBRE PREDICTIVA. ....	105
FIGURA 4.16. COMPARACIÓN ENTRE LA INCERTIDUMBRE DE LOS PARÁMETROS Y SU EFECTO SOBRE LA INCERTIDUMBRE PREDICTIVA.....	105
FIGURA 4.17. PROYECCIÓN DE LA INCERTIDUMBRE DEL PARÁMETRO SOBRE LA INCERTIDUMBRE PREDICTIVA (CASO 1). ....	106
FIGURA 4.18. PROYECCIÓN DE LA INCERTIDUMBRE DEL PARÁMETRO SOBRE LA INCERTIDUMBRE PREDICTIVA (CASO 2). ....	106
FIGURA 4.19. INCERTIDUMBRE TOTAL DE LA PREDICCIÓN. ....	108
FIGURA 4.20. INCERTIDUMBRE TOTAL DE LA PREDICCIÓN. DETALLE ENTRE LOS DÍAS 100 Y 300 DE SIMULACIÓN.....	108
FIGURA 4.21. COMPARACIÓN ENTRE LAS DISTRIBUCIONES TEÓRICA Y REAL DE LOS RESIDUOS.....	109
FIGURA 4.22. DIAGRAMA DE DISPERSIÓN DE LOS RESIDUOS EN EL QUE SE APRECIA HETEROCEDASTICIDAD. ....	110
FIGURA 4.23. FUNCIÓN DE AUTOCORRELACIÓN DE LOS RESIDUOS. ....	111



## LISTA DE TABLAS

TABLA 4.1. INTERVALOS DE VARIACIÓN DE LOS COEFICIENTES Y EXPONENTES GEOMORFOLÓGICOS.....	68
TABLA 4.2. FACTORES CORRECTORES Y RESPECTIVOS PARÁMETROS AJUSTADOS. ....	70
TABLA 4.3. INFORMACIÓN CARTOGRÁFICA UTILIZADA PARA LA ESTIMACIÓN DE PARÁMETROS.....	79
TABLA 4.4 EQUIVALENCIA EN LA NOMENCLATURA DE LOS PARÁMETROS.....	88
TABLA 4.5. PARÁMETROS ORDENADOS DE MAYOR A MENOR DESV. TÍPICA.....	99
TABLA 4.6. COEFICIENTES DE NASH DE LOS DISTINTOS CÁLCULOS.....	104
TABLA 4.7. VARIACIÓN DE LOS COEFICIENTES DE NASH QUE PROVOCA LA FIJACIÓN DE CADA PARÁMETRO.....	104

## 1.-INTRODUCCIÓN

### 1.1.- MOTIVACIÓN Y OBJETIVOS

Uno de los principios básicos de la Directiva Marco del Agua (DMA, 2000/60/CE) es "contribuir a la consecución de los objetivos de preservar, proteger y mejorar la calidad del medio ambiente con el uso prudente y racional de los recursos naturales, basándose en el principio de precaución". En la medida en que el principio de precaución tiene como objetivo proteger a los humanos y al medio ambiente contra los riesgos inciertos, por medio de medidas preventivas, dicha tarea no puede llevarse a cabo sin la incorporación de la evaluación de la incertidumbre en el proceso de toma de decisiones.

La evaluación de la incertidumbre de los modelos de simulación es por lo tanto, importante, cuando los modelos se utilizan para apoyar las decisiones sobre la gestión del agua (Refsgaard et al. 2007).

Los modelos hidrológicos suelen contener parámetros que no pueden ser determinados, mediante el empleo de medidas directas de los mismos en campo (Beven and Freer 2001). El proceso de calibración del modelo, permite el ajuste de esos parámetros dentro de unos rangos razonables, mediante el mejor ajuste posible entre los resultados de ciertas variables del modelo, y las observaciones que sobre ellas se haya realizado. Sin embargo, los valores calibrados pueden verse afectados por correlaciones entre los parámetros y por otras características estadísticas de los residuos. Esta circunstancia conduce a la llamada Equifinalidad de las soluciones (Beven and Binley 1992; Duan and Sorooshian 1992)

Por otra parte, en modelos hidrológicos complejos son varias las fuentes de incertidumbre, y esta incertidumbre acaba propagándose a los resultados del modelo. Los datos de entrada del modelo, su estructura y el valor de los parámetros son las principales fuentes de incertidumbre. Entonces, las



predicciones de un modelo deberían ser presentadas con un intervalo de confianza, en lugar de cómo un valor “cierto”.

El objetivo de esta tesina es explorar algunas técnicas de análisis de sensibilidad e incertidumbre sobre los parámetros, prestando especial atención a los algoritmos MCMC. Así mismo se amplían las capacidades de análisis del modelo conceptual de simulación hidrológica **TETIS** (Francés et al. 2007) mediante la integración de dicho modelo hidrológico (programado en Fortran) con el paquete de **R** denominado **FME**, *Flexible Modeling Environment* (Soetaert and Petzoldt 2010; Soetaert 2010a; b).

FME incluye herramientas de calibración, análisis de sensibilidad, identificabilidad de parámetros, así como el algoritmo MCMC denominado **DRAM**, *Delayed Rejection Adaptive Metropolis* (Haario et al. 2006), para el análisis de incertidumbre de los parámetros.

Para abordar el acople de las dos herramientas se han desarrollado subrutinas de código en lenguaje R y en lenguaje ©**Matlab**, las cuales permiten la comunicación secuencial entre los entornos Tetis y FME. Al conjunto de herramientas de modelación se le ha denominado **RLAB-TETIS-FME**.

El trabajo se completa, con la aplicación de la herramienta desarrollada a un caso práctico: El análisis de los parámetros del modelo hidrológico de la cuenca vertiente al embalse de La Baells (Alto Llobregat).

## 1.2.- JUSTIFICACIÓN DEL TRABAJO

De las distintas aproximaciones existentes para abordar el tema de la incertidumbre de los parámetros, los métodos basados en simulación de Monte Carlo se han hecho más populares que otros métodos tradicionales, como consecuencia de sus fortalezas y de la mejora de capacidad computacional experimentada en los últimos años. Entre sus fortalezas destaca la capacidad de poder tratar la no linealidad y la interdependencia de parámetros en modelos hidrológicos complejos. Entre estos métodos, destacan los del tipo *Markov Chain Monte Carlo*, como una de las técnicas numéricas más





importantes para la extracción de una muestra de la distribución posterior. MCMC está siendo ampliamente usado en modelación hidrológica para cuantificar la incertidumbre de los parámetros, (He et al. 2011; George Kuczera and Parent 1998; Li et al. 2010; Vrugt et al. 2003; Vrugt, Braak, Diks, et al. 2009; Vrugt, Braak, and Gupta 2009), entre otros.

El interés de aplicar estas metodologías junto con el modelo conceptual hidrológico distribuido TETIS, desarrollado en el Departamento de Ingeniería Hidráulica y Medio Ambiente de la Universitat Politècnica de València, se explica con los siguientes argumentos:

- Contribuir con investigaciones científicas enfocadas al uso de modelos hidrológicos distribuidos.
- Posibilitar el empleo conjunto de Tetis con una herramienta de estadística bayesiana formal, que permita comprobar las posibilidades de este marco de trabajo para la evaluación de la incertidumbre. Una de las ventajas de las aproximaciones bayesianas formales respecto a otros marcos de evaluación de la incertidumbre, es que mediante ellas es posible discernir el efecto de las distintas fuentes de la incertidumbre, lo cual es la clave que permitirá mejorar el conocimiento hidrológico (Vrugt et al. 2008).
- Analizar mediante una técnica MCMC, una de las posibles fuentes de incertidumbre (los parámetros) de un modelo hidrológico, delimitando su importancia sobre la respuesta del modelo y sobre su incertidumbre.

### 1.3.- ESTRUCTURA DEL DOCUMENTO

El documento consta de 7 capítulos. A continuación se realiza una breve descripción de los mismos.

En el **Capítulo 2** se realiza una introducción a la línea de investigación de la tesina, mediante una revisión bibliográfica sobre la perspectiva desde la que en modelación hidrológica, se han visto problemas como la falta de identificabilidad en la estimación de parámetros así como el problema de la incertidumbre de los procesos modelados y la identificación de sus fuentes. Así



mismo se revisan los posibles métodos con los que cuenta la evaluación de modelos hidrológicos.

En el **Capítulo 3** se explican las bases conceptuales relacionadas con la aproximación formal bayesiana al cálculo de incertidumbres. Se comienza introduciendo los conceptos básicos de la estadística bayesiana computacional. En el siguiente apartado se explican de forma concisa las bases sobre las que se asienta la teoría de cadenas de Markov. Establecidos los principios teóricos básicos, en el apartado 3.3 se introducen los conceptos que permiten comprender los fundamentos de los métodos *Markov Chain Monte Carlo* y más concretamente el algoritmo de Metrópolis para el muestreo de la posterior. En el apartado 3.4 se dan las claves para obtener una muestra aleatoria de la distribución posterior de los parámetros, a partir de los elementos de una cadena de Markov obtenida mediante un algoritmo MCMC. Y finalmente en el apartado 3.5 se describen las características del algoritmo MCMC de tipo adaptativo, que se ha empleado en la aplicación práctica del trabajo.

El **Capítulo 4** presenta una aplicación práctica del algoritmo MCMC adaptativo denominado DRAM, para la evaluación de la incertidumbre de los parámetros. El análisis MCMC se realiza con el paquete FME de R. El análisis de la hidrología se efectúa con el modelo hidrológico distribuido TETIS, sobre una subcuenca de alta montaña ubicada en la cuenca del Llobregat.

Finalmente el **Capítulo 5** expone las conclusiones del trabajo de investigación y en el **Capítulo 6** se proponen las posibles líneas de trabajo futuro. En el **Capítulo 7** se presentan las referencias a la bibliografía consultada.



## 2.-MODELACIÓN HIDROLÓGICA E INCERTIDUMBRE

### 2.1.- INTRODUCCIÓN

Una labor importante en ciencia es entender y predecir la naturaleza mediante la modelación. El análisis estadístico estudia las incertidumbres en el proceso de inferencia estadística, mediante razonamientos probabilísticos (Laine 2008).

En el tratamiento de la incertidumbre, se asume que las cantidades desconocidas pueden describirse mediante distribuciones estadísticas. Esas cantidades desconocidas pueden ser los llamados datos *Forcing* (p.e. variables meteorológicas), las variables de control o de estado (p.e. escorrentía directa, infiltración,..), los parámetros del modelo, las predicciones, etc.

Cuando se está interesado en el valor de los parámetros, el proceso de inferencia se denomina Estimación de Parámetros, o Calibración o el Problema Inverso.

Si se contempla un modelo de error aditivo cualquiera, se puede escribir un modelo, en términos simples, como (Moradkhani and Sorooshian 2008):

$$y_{obs} = y_{mod} + \xi \quad \text{Ec. 2.1}$$

donde:

- $y_{obs}$ , representa las observaciones de una variable cuyo comportamiento medio esperado, es descrito por el modelo
- $y_{mod}$ , representa al modelo, que es función de ciertas variables de control o de estado y de unas cantidades desconocidas, los parámetros ( $\theta$ ). Sobre los parámetros puede conocerse alguna información a priori, la cual podría describirse en forma de distribución estadística.



- $\xi$  representa al error, Residuo o grado de incertidumbre que presentan los resultados del modelo, respecto a las observaciones, también llamado **Incertidumbre Total de la Predicción**.

Dicho error se describe mediante su distribución de probabilidad y está conformado por la suma de:

- Todos los efectos no considerados por el modelo (o considerados de forma muy simplificada) y que sí forman parte significativamente de los procesos subyacentes a la variable simulada.
- Los errores observacionales.

Para un mejor entendimiento de los procesos modelados, será importante desenmarañar el origen de los errores que el término de error cuantifica de forma agregada. Conociendo las llamadas **Fuentes de Incertidumbre**, y cuantificando sus respectivas incertidumbres, se podrá mejorar el proceso de modelación y por tanto el conocimiento científico sobre el área de investigación.

## 2.2.- MODELOS HIDROLÓGICOS

Se puede definir como Modelo Ambiental, al proceso de aplicar sobre unos datos, las ideas preconcebidas sobre el funcionamiento del sistema, con el propósito de realizar una predicción del comportamiento del mismo, sobre la cual se pueda basar una decisión (Gourley and Vieux 2006).

Para evaluar la respuesta hidrológica de un sistema, es indispensable la elaboración de algún tipo de esquema, por medio del cual se pueda representar en forma simplificada, un sistema real; en otras palabras: un modelo. El mismo podrá ser utilizado para reconstruir eventos pasados y predecir los futuros (Puricelli 2003)

Los modelos en cualquier ámbito de la ciencia se pueden clasificar según diferentes criterios (Beck 1987; Gourley and Vieux 2006):



1. En función del tratamiento que se hace sobre inputs y parámetros como funciones del espacio y del tiempo.
2. Según la extensión de los principios físicos que forman parte de la estructura del modelo.
3. Por la forma de presentar los resultados, determinista o estocástica.

Según el primer criterio los modelos pueden ser:

- A. Agregados. Los parámetros y los inputs adoptan el valor de un promedio sobre la cuenca. De esta manera, no se tiene en cuenta la variabilidad espacial y/o temporal.
- B. Distribuidos. Emplean mapas con la distribución espacial de parámetros e inputs.
- C. Semi-Distribuidos. Son un híbrido entre las dos aproximaciones previas. Tienen en cuenta en diferentes grados la variabilidad espacial de parámetros e inputs.

Con el transcurrir del tiempo aumenta la controversia y el interés respecto a las **ventajas de la modelación distribuida**. La problemática en la aplicación de modelos distribuidos está enfocada a definir la escala de representación espacial y temporal de los fenómenos hidrológicos de interés, teniendo en cuenta los datos existentes o los posibles de recopilar.

Son puntos clave a considerar en los modelos distribuidos:

- La representación de la heterogeneidad topográfica y climática.
- La variación entre la escala espacio-tiempo de las observaciones y la utilizada en el modelo.
- La definición de la escala de análisis.
- El elevado número de parámetros, no siempre con significado físico.



La falta de sensibilidad de un modelo ante una variación en su estructura o ante la variación de alguno/s de sus parámetros, se denomina falta de **Identificabilidad**. Esta característica es propia de modelos sobreparametrizados, y en función de ella, se establece **el segundo criterio** con el que se pueden definir los siguientes tres tipos de modelos (Beck 1987):

- A. **Modelos de Clase I.** Son conocidos como modelos físicamente basados, modelos mecanicistas o modelos de *caja blanca*. En ellos, la estructura (ecuaciones, parámetros y variables de estado) contiene todas las leyes físicas conocidas, que rigen la dinámica del sistema. Dados unos Inputs al modelo, los resultados son una solución directa de los principios físicos. Estos modelos requieren una fuerte parametrización, lo cual conduce a una falta de identificabilidad que conlleva a su vez, una gran incertidumbre estructural. Por otro lado los modelos físicamente basados tienen la flexibilidad de poder predecir estados futuros del modelo, que aún no han sido observados. Ello no evita que voces críticas argumenten que la extrapolación de resultados, incluso con modelos mecanicistas, sea algunas veces más un ejercicio de profecía que de predicción científica (Beven 1987, 1989, 1993).
- B. **Modelos de Clase III.** Se basan únicamente en el comportamiento pasado del sistema, es decir en las observaciones de Inputs y Outputs, no empleando por tanto ningún tipo de leyes o principios físicos. Estos modelos se conocen como modelos *Data Driven*, *Caja Negra*, o modelos estadísticos. Las predicciones en este tipo de modelos vienen limitadas por la “memoria” de las observaciones del comportamiento del sistema en el pasado, es decir, no son capaces de predecir comportamientos que no hayan sido ya observados en el pasado.
- C. **Modelos de Clase II.** Son una solución intermedia entre los modelos de Clase I y los de Clase III. En estos modelos se aplican un conjunto de conceptos o leyes básicas, con calado científico menos profundo que en los físicamente basados. Se conocen como modelos Conceptuales o de *Caja Gris*.

Requieren necesariamente de un proceso de calibración de parámetros que permite el ajuste entre las variables de estado simuladas, y sus correspondientes observaciones. Los modelos Conceptuales permiten, al igual que los de Clase I, la predicción de estados futuros, nunca observados previamente (extrapolación). El número de conceptos, o procesos simulados, es menor que en el caso de modelos físicamente basados, así también el número de parámetros. Por ello, estos modelos son más Identificables, tienen menos incertidumbre estructural y son más flexibles, por lo que son preferibles en modelación ambiental.

**El tercer criterio** para distinguir modelos, es la forma en que se presentan los resultados. Los modelos **Determinísticos** proporcionan un único vector de resultados, dadas unas condiciones iniciales y de contorno. Este tipo de resultados es útil cuando los Inputs, los parámetros y el sistema físico modelado son conocidos de forma precisa. Sin embargo, en simulación ambiental, es rara tanta certeza, siendo más probable la existencia de incertidumbre asociada a los Inputs, a los parámetros, a la estructura del modelo y a las observaciones de las variables de estado. Toda esta incertidumbre puede ser incluida en las variables de predicción y presentadas en un marco **estocástico o probabilístico**.

Como resumen de la clasificación de modelos hidrológicos, se presenta la Figura 2.1.



Figura 2.1. Clasificación General de Modelos Hidrológicos

(Fuente: Apuntes Modelación Hidrológica y Distribuida, Francés (2009))



Según este marco de análisis, en el presente trabajo se aplicará un modelo hidrológico distribuido de tipo conceptual llamado TETIS (Francés et al. 2007), desarrollado en el Departamento de Ingeniería Hidráulica y Medio Ambiente de la Universitat Politècnica de València. La aplicación práctica de este trabajo se desarrolla en Simulación Continua. Si bien el modelo TETIS está concebido como un modelo Determinista, mediante las herramientas desarrolladas en el presente trabajo, puede ser empleado como modelo de tipo Estocástico.

### 2.3.- SOBRE LA IDENTIFICABILIDAD

Según Brun and Reichert (2001), con frecuencia, los modelos complejos de simulación ambiental están sobre-parametrizados en relación al conjunto de observaciones disponibles (estadísticamente se recomienda un mínimo de entre 20 a 30 observaciones independientes, por parámetro a estimar). Esto provoca que los parámetros del modelo resulten pobremente identificables o no identificables.

Se dice que el modelo es Identificable si es teóricamente posible conocer el verdadero valor de sus parámetros después de obtener un número infinito de resultados del mismo, para unos datos dados. Matemáticamente, esto equivale a decir que las diferentes combinaciones de valores de los parámetros deben generar diferentes distribuciones de probabilidad de las variables observables. En términos estadísticos, la Identificabilidad es una propiedad que un modelo debe satisfacer para que la inferencia (deducción de propiedades estadísticas poblacionales a partir de estadísticos muestrales) sea posible.

Un conjunto de parámetros del modelo se dice que es identificable, si todos los parámetros del mismo, pueden ser estimados de forma única a partir de los valores observados y simulados de una o más variables del modelo. Se cumple que si un conjunto de parámetros presenta colinealidad (dependencia lineal) entre ellos, no será identificable. Sin embargo la colinealidad es solo una condición suficiente aunque no necesaria, para que los parámetros sean no-identificables.





Los problemas de identificabilidad de parámetros, sobrevienen principalmente en modelos sobre-parametrizados, siendo los modelos físicamente basados el mayor exponente de la problemática, la cual ha sido ampliamente tratada en la literatura (Beck 1987). La principal razón de la pobre identificabilidad de la que padecen los modelos mecanicistas, según Beck, es que “Lo que nos gustaría conocer sobre la descripción interna del sistema [...] es de un orden muy superior a lo que podemos observar sobre la descripción externa del mismo”.

Según Yapo and H. Gupta (1996), las posibles razones por las que los parámetros del modelo pueden ser difícilmente identificables son:

- La ya mencionada colinealidad o interdependencia entre parámetros, que produce la fuerte interacción entre ellos. Dicha interacción es la causante de que se compensen las variaciones en el espacio de parámetros.
- No estacionariedad de los parámetros. Las medidas de localización y precisión de los parámetros se correlacionan con diferentes características de la cuenca o de los datos, que no se consideran durante la calibración.
- Los datos no informan. Los datos no incluyen las condiciones hidrológicas necesarias para identificar correctamente los parámetros.
- No adecuación de criterio. La función objetivo no extrae adecuadamente la información contenida en los datos.
- Insensibilidad. Variaciones en los valores de los parámetros no afectan el resultado del modelo, de manera significativa.

El análisis de Identificabilidad de un modelo se compone de 2 partes (Brun and Reichert 2001):

- El problema de la selección de la estructura del modelo

- El problema de la identificación de parámetros

Asumiendo que el modelo recoge en lenguaje matemático todo el conocimiento científico sobre el sistema analizado, queda el problema de la identificabilidad de los parámetros. Cuando se trabaja con modelos ambientales complejos, no es razonable esperar poder identificar los parámetros “verdaderos” del modelo. Más bien se espera poder encontrar (mediante un proceso de Estimación de Parámetros) unos valores de parámetros “razonables” que además describan los datos observados de la manera más adecuada.

En modelos no lineales, puede darse la circunstancia de encontrar varios conjuntos de parámetros que cumplan con esos dos requerimientos. La explicación a esta no unicidad de la solución del problema es la que sigue: El efecto sobre los resultados del modelo, de cambiar el valor a un parámetro, es cancelado por el cambio de valor en otro parámetro del modelo. Esta circunstancia lleva implícita una fuerte dependencia entre parámetros (colinealidad), y se dice entonces que los parámetros son pobremente identificables. El objetivo del análisis de identificabilidad es proporcionar alguna información que ayude a intuir, de entre todos los posibles, el valor adecuado de los parámetros (Brun and Reichert 2001), descartando así mismo los parámetros que no son útiles al modelo (según los datos con los que se cuenta) y eliminando del proceso de estimación aquellos parámetros poco identificables, debiendo asignarles un valor prefijado. Dicha pre-asignación de valores a ciertos parámetros será una posible fuente de sesgo para la solución obtenida en la estimación del resto de parámetros.

La principal herramienta para la evaluación de la identificabilidad de los parámetros ha sido el **Análisis de Sensibilidad**. Este análisis puede ser realizado **localmente**, actuando sobre un solo parámetro de forma simultánea, bien en la cercanía de un punto dado del espacio de parámetros o bien sobre una región amplia de valores razonables del espacio de parámetros. Otra aproximación es el análisis de sensibilidad **global**, de forma simultánea sobre todos los parámetros y sobre el espacio de parámetros completo, muy



adecuada para cuantificar la influencia media de los parámetros con sus interacciones, sobre el resultado del modelo (Brun and Reichert 2001).

Una aproximación prometedora, que parece superar los problemas de identificabilidad de parámetros, es el uso de los **Métodos Bayesianos**, que pueden incorporar información a priori sobre los parámetros y no requieren que los parámetros sean identificables (Omlin and Reichert 1999; Reichert and Omlin 1997). Estos métodos presentan un alto coste computacional, por lo que hasta hace poco han estado restringidos a un uso en modelos con pocos parámetros y tiempos cortos de simulación.

## 2.4.- SOBRE LA INCERTIDUMBRE

### 2.4.1.- GENERALIDADES

Incertidumbre es un término usado en formas sutilmente diferentes en una serie de campos tan distintos como la filosofía, estadística, economía, la ingeniería y la ciencia. Se aplica a predicciones de eventos futuros, a medidas físicas ya realizadas, o a lo desconocido. No hay una definición universal aplicable al concepto de incertidumbre, en el campo de la modelación ambiental, de la cual forma parte la modelación hidrológica. Según Refsgaard et al. (2007), la incertidumbre puede ser vista como una propiedad de la información empleada para tomar decisiones, o como una manifestación de la confianza que se tiene en dicha información.

Son diversas las taxonomías que sobre la incertidumbre se encuentran en la literatura. En Matott et al. (2009b) se propone una sencilla clasificación, separando a la incertidumbre en reducible e irreducible. La incertidumbre reducible puede, en teoría, ser disminuida mediante mejoras en la adquisición de los datos de entrada al modelo, en el refinamiento de la estructura del mismo, o con calibraciones más prolongadas. La incertidumbre irreducible, como la asociada con la predicción de futuros resultados, no puede ser disminuida, debido a que no se conoce el estado futuro de todos los factores y



fuerzas reflejadas en el modelo. Resumiendo, la incertidumbre puede ser clasificada en:

- Reducible: Derivada de unos datos con error, o del conocimiento erróneo de los procesos simulados.
- Irreducible: La que se produce por la variabilidad inherente a un proceso.

Desde la perspectiva del análisis de modelos y de la toma de decisiones, la incertidumbre total de una determinada cantidad o valor, puede a su vez ser caracterizada de cuatro maneras diferentes (Cullen and Frey 1999), a saber:

- Puramente irreducible: La cantidad es variable, pero su población asociada, se ha muestreado completa y correctamente (sin error).
- Parcialmente irreducible o parcialmente reducible: La cantidad varía, pero además su población asociada se ha muestreado de manera parcial y/o con error.
- Puramente reducible: La cantidad no varía, pero ha sido muestreada con error.
- Cierta: El valor no varía y se ha muestreado de forma correcta (sin error).

Con frecuencia la incertidumbre es pensada en términos de exactitud y precisión. Exactitud es el grado en el que una medida, o resultado de un modelo, reflejan el “verdadero” valor o naturaleza de los parámetros o procesos subyacentes. Se entiende por precisión al grado en el que un proceso puede alcanzar repetidamente el mismo resultado, dados los mismos inputs. Se puede decir que la exactitud es la capacidad de reproducir (en promedio) el valor “real” de un proceso o parámetro, mientras que la precisión es la magnitud de la dispersión (varianza) de los resultados obtenidos en distintas ejecuciones para la obtención del resultado.



#### 2.4.2.- FUENTES DE INCERTIDUMBRE

La incertidumbre de un modelo procede de distintas fuentes. En la bibliografía se encuentran numerosas clasificaciones de las fuentes de incertidumbre, las cuales no siempre pueden ser conciliadas (Matott et al. 2009b), por ejemplo:

- Beck (1987): Incertidumbre del estado inicial del sistema, de los parámetros, de los Inputs, y de las variables de salida del modelo.
- Refsgaard et al. (2007): Incertidumbre sobre el contexto del problema, de los Inputs, de los Parámetros, Estructural (del modelo), e incertidumbre técnica.
- Brendecke (2012): Incertidumbre del modelo (estructura y parámetros) e incertidumbre de los datos (Inputs y observación de algunos Outputs). La incertidumbre de un modelo puede variar según los escenarios considerados, o según la localización espacial considerada. Es probable que el grado de incertidumbre se incremente con el nivel de detalle de las predicciones

Por su parte, Matott et al. (2009b), intentan subsanar las inconsistencias que en la literatura científica se encuentran sobre el concepto de incertidumbre. Para ello proponen la introducción de un glosario de términos (Matott et al. 2009a) así como una taxonomía simplificada, que en resumen consiste en una incertidumbre cuantificable debida al modelo y a las entradas al modelo, la cual queda cubierta bajo el paraguas de una incertidumbre cualitativa provocada por la acción del modelador. Así las fuentes de incertidumbre en modelación ambiental pueden identificarse en:

##### 1. El Modelador

Es el responsable de confeccionar los inputs del modelo para poderlos introducir al modelo, y simular para obtener las salidas o resultados. Esta incertidumbre puede ser medida mediante el estudio comparativo (contrastando los resultados de múltiples modeladores independientes).



## 2. Los Inputs del Modelo

Se consideran inputs del modelo a los datos de entrada, los datos observados de las variables respuesta y los parámetros del modelo.

**Los datos de entrada (*Forcing Data*)**, hacen referencia tanto a las funciones o métodos de generación de los datos de entrada (por ejemplo el método de interpolación de la lluvia), como a los propios datos de las fuentes (temperatura, precipitación,...) y sumideros (pérdidas) del modelo así como a las condiciones iniciales y condiciones de contorno del mismo.

Los datos de las variables respuesta, representan a **las observaciones** (medidas) efectuadas en sitios específicos, sobre las variables respuesta del modelo. Estas mediciones pueden ser comparadas con los resultados simulados por el modelo.

Tanto los inputs del modelo como las observaciones, son tratados desde el punto de vista de la incertidumbre como Irreducibles o Parcialmente Irreducibles/Reducibles. La utilidad (o el contenido de información útil) del conjunto de los Inputs y los datos de observaciones, dependerá del grado en el que los inputs ejercen influencia sobre los resultados del modelo, es decir de la sensibilidad del modelo a dichos inputs.

**Los parámetros** del modelo son similares a los inputs del modelo, a excepción de los parámetros que se obtienen mediante un proceso de estimación de parámetros (calibración del modelo). En estos casos, cuantificar la incertidumbre del parámetro requiere considerar tanto la incertidumbre del modelo como de los datos (inputs y observaciones). La incertidumbre de los parámetros se considera como Puramente Reducible, es decir, los parámetros calibrados son constantes para una localización dada, si bien sus valores son desconocidos (incierto).

La no-unicidad (Equifinalidad) o falta de identificabilidad de las soluciones en la calibración, es considerada un riesgo endémico de los modelos ambientales, debido a la sobre-parametrización de los mismos (Beven 2006).



El empleo de objetivos múltiples de calibración ha reducido aunque no eliminado, el riesgo de las soluciones equifinales.

Si además la incertidumbre debida a la estructura del modelo no es tenida en consideración de manera separada o explícita, el proceso de calibración puede combinar los errores en la estructura con la incertidumbre de los parámetros, para poder alcanzar un ajuste a los datos históricos aceptable que haga parecer que el modelo funciona bien, aunque no por las razones adecuadas (Beven 2006).

La variabilidad sistemática de los parámetros, es un importante indicador del error del modelo que no suele ser tratado de forma explícita por los modeladores (Beck 1987; Kuczera et al. 2006) y que produce que las distintas fuentes de incertidumbre (*forcing data*, observaciones, parámetros y modelo) puedan quedar confundidas en una incertidumbre híbrida.

### 3. El Modelo

La incertidumbre del modelo refleja su incapacidad, incluso cuando es provisto con Inputs perfectos (es decir ciertos o Puramente Irreducibles), para generar resultados simulados indistinguibles de las observaciones del mundo real. Esta incertidumbre puede a su vez ser dividida en:

- Estructural. Relativa a las hipótesis científicas y ecuaciones sobre las que se fundamenta el modelo.
- De Resolución. Consecuencia de la discretización espacio-temporal empleada en los cálculos, las especificaciones de contorno, y la dependencia de escala de los procesos simulados.
- De Resolución Muestral. Hace referencia a la posible falta de correspondencia entre los Inputs-Outputs del modelo y las observaciones (muestras de la realidad). Por ejemplo, valores asignados a Inputs volumétricos/areales, son extrapolados a partir de datos puntuales. Igualmente los Outputs volumétricos/areales del modelo son comparados con observaciones de datos puntuales. Y similares



consideraciones podrían aplicarse a las variables en el dominio temporal.

- De Implementación del Código y/o errores numéricos.

Todas ellas son desconocidas y difíciles de cuantificar. Entonces, la identificación de la incertidumbre del modelo se realiza mediante la cuantificación del resto de fuentes de error, siendo la incertidumbre del modelo un componente del error total que no es explícitamente tratado, es decir, un componente de los “residuos” (Gourley and Vieux 2006). La incertidumbre relacionada con **la estructura del modelo es rara vez evaluada**. Es costoso y difícil comprobar si el conjunto de modelos plausibles ha sido completamente explorado (Refsgaard et al. 2007). Por ello en general esta dimensión de la incertidumbre se ignora en favor de optimizar una estructura ya elegida del modelo.

## 2.5.- EVALUACIÓN DE MODELOS

Existen distintas tecnologías de evaluación de modelos con las complejidades inherentes a su integración con los modelos numéricos. Esta integración modelos-métodos de evaluación es todavía un reto (Beven 2007), si bien proporcionará sólidos cimientos a los procedimientos de evaluación ambiental.

En (Matott et al. 2009b) se propone una clasificación de métodos de evaluación de modelos, que distingue **9 categorías, entre las que cabe que exista cierto grado de solapamiento**. Por ejemplo, el Análisis de Identificabilidad contiene elementos tanto de la Estimación de Parámetros como del Análisis de Sensibilidad. De igual manera el Análisis Multimodelo incorpora la Estimación de Parámetros. Y las Redes Bayesianas son análogas a la Estimación de Parámetros simultáneamente con el Análisis de Incertidumbre.



## 1. Análisis de Datos

Se refiere a los procedimientos analíticos, estadísticos y gráficos para evaluar y resumir los datos Input y los resultados del modelo. Incluye el cribado de los datos y la parametrización de funciones de las funciones de distribución de los Inputs, de las Observaciones y de los resultados del modelo. Especial atención se presta al análisis de datos geoespaciales y de las series temporales, debido a la presencia de estructuras de correlación dependientes de la escala considerada.

## 2. Verificación del Modelo

Consiste en la elaboración y ejecución de procedimientos, que establecen que **el código** del modelo resuelve correctamente el conjunto de formulaciones matemáticas que representan el sistema de interés.

## 3. Validación del Modelo

Es el procedimiento mediante el cual se examina de forma rigurosa, la capacidad del modelo para replicar datos o comportamientos previamente observados en el mundo real.

## 4. Análisis de Identificabilidad

Persigue poner de manifiesto deficiencias en los datos o sugerir mejoras en la estructura del modelo. Alternativamente puede advertir de qué parámetros del modelo no pueden ser determinados adecuadamente a causa de una insuficiente cantidad o diversidad de datos de observaciones.

## 5. Estimación de Parámetros (EP)

Técnicas que cuantifican los inciertos parámetros del modelo, en base a la comparación de los resultados del modelo con los datos de observaciones del mundo real. Las técnicas de estimación de parámetros se pueden agrupar en dos grandes grupos: Aproximaciones de Solución Simple y Aproximaciones de Solución Múltiple (conjuntos de parámetros).



## A. Aproximaciones de Solución Simple

Abordan la EP como un problema de optimización cuya única solución es el “mejor ajuste” de parámetros del modelo. Gran variedad de algoritmos de búsqueda han sido desarrollados para este proceso, que se conoce con el nombre de **Calibración Automática** del modelo. El conjunto de algoritmos podría agruparse en 3 categorías generales en función del método de búsqueda del óptimo: Técnicas Locales, Globales e Híbridas.

Mediante las aproximaciones de Solución Simple se obtienen estimaciones puntuales de los parámetros las cuales no van acompañadas de ninguna información relativa su grado de confiabilidad. Algunos códigos que abordan la EP, intentan suplir esta carencia proporcionando junto a la propia calibración, ciertos resultados de tratamiento estadístico post-calibración como son: Intervalos de Confianza (modelo lineal), Coeficientes de Correlación entre parámetros y Análisis de Sensibilidad de los parámetros.

## B. Aproximaciones de Solución Múltiple

Se pueden subdividir en dos categorías: Técnicas del tipo *Importance Sampling* y técnicas **Markov Chain Monte Carlo (MCMC)**. Ambas aproximaciones permiten obtener la distribución del valor de cada parámetro, en lugar de una estimación puntual.

*Importance Sampling* busca la identificación de una familia de parámetros “aceptable” o behavioral. Los conjuntos de parámetros muestreados son clasificados en dos grupos: *behavioral* y *non-behavioral*, conforme a un umbral de aceptación para alguna función objetivo. El proceso de descarte de los conjuntos *non-behavioral* se denomina *Rejection Sampling*. Después del descarte, la distribución de probabilidad de cada parámetro es estimada mediante una combinación ponderada de los conjuntos de parámetros *behavioral*. En la comunidad de modelación ambiental es popular el método *GLUE (Generalized Likelihood Uncertainty Engine)* (Beven and Binley 1992).



Por su parte, los métodos *Markov Chain Monte Carlo* (MCMC), incorporan algunos aspectos de los métodos *Importance Sampling* y *Rejection Sampling* en un procedimiento robusto desde el punto de vista matemático, que permite la evaluación de la distribución de probabilidad condicional (condicional dadas las observaciones) de cada parámetro. El muestreador MCMC evoluciona las distribuciones a priori de los parámetros, que deben ser suministradas al algoritmo sin tener en cuenta los datos observados de las variables respuesta, es decir, son distribuciones que deben recoger el conocimiento a priori que se tenga del valor del parámetro. El resultado de esa evolución de las distribuciones a priori es la distribución a posteriori del parámetro, condicionada a los datos observados.

Independientemente de la técnica de estimación empleada, hay consideraciones adicionales que deben ser tenidas en cuenta en el proceso de EP:

- La selección de la función objetivo
- La optimización de una o múltiples funciones objetivo de forma simultánea
- La incorporación de información a priori de los parámetros
- La regularización de parámetros

## 6. Análisis de Incertidumbre (AI)

Los métodos AI propagan las fuentes de incertidumbre a través del modelo, con el fin de obtener los momentos estadísticos o las distribuciones de probabilidad de las variables respuesta del modelo. Dos son los grupos de métodos principales de AI: Métodos de Aproximación y Métodos de Muestreo.

### A. Métodos de Aproximación

Caracterizan la incertidumbre de los Outputs del modelo mediante la propagación de uno o más momentos (media, varianza, asimetría y kurtosis) de las distribuciones de los Inputs del modelo, a través del sistema modelado.



## B. Métodos de Muestreo

Caracterizan las distribuciones de los Outputs del modelo mediante la propagación de un intensivo muestreo aleatorio de la distribución de los Inputs del modelo. Son métodos computacionalmente intensivos por lo que son varias las técnicas que se emplean para reducir el tiempo real (*wall time*) de este tipo de análisis. Algunas pasan por abordar la sustitución de modelos complejos, o parte de ellos, con aproximaciones más “económicas” (por ejemplo Redes Neuronales). La computación paralela es otra de las soluciones para disminuir los tiempos de proceso. Estos métodos incluyen los métodos de Monte Carlo (MC), métodos de Muestreo Estratificado, métodos *Importance Sampling* o una combinación de ellos.

- a. **Los métodos MC** extraen muestras aleatorias insesgadas desde una distribución preestablecida. El error estadístico del muestreo se estima mediante la teoría estadística clásica; dado un determinado nivel de exactitud y precisión del muestreo, el tamaño de la muestra requerida queda determinado sin ambigüedades.
- b. **El Muestreo Estratificado** (Hiper cubo Latino), divide la distribución de un Input en intervalos, de manera que se extraen muestras de cada intervalo. Para un muestreo eficiente, los intervalos suelen presentar la misma probabilidad de ocurrencia.
- c. **Los métodos *Importance Sampling*** son empleados para generar muestras intencionadamente sesgadas, asegurando que alguna clase particular de comportamiento es analizada (por ejemplo eventos poco probables). Así mismo, *Importance Sampling* es útil cuando la distribución del Input es desconocida o difícil de muestrear, y es entonces “envuelta” por una distribución alternativa fácilmente muestreable. En estos casos se muestrea la distribución alternativa y la distribución del Input es inferida mediante la técnica *Importance Weighting*, en la cual los pesos muestrales reflejan la probabilidad de que la muestra de la distribución alternativa pueda también proceder de

la distribución de interés del Input. Como se vio anteriormente, *Importance Sampling* es también una técnica muy útil en la estimación de parámetros.

## 7. Análisis de Sensibilidad (AS)

El AS estudia el grado en el que los resultados del modelo se ven influenciados por cambios en los Inputs del modelo incluyendo en la categoría de Inputs, a los parámetros del mismo. Así estos métodos ayudan a identificar puntos críticos del modelo, por falta de conocimiento del sistema o insuficiencia de datos observados. Los métodos AS pueden ser clasificados en métodos de *Screening*, análisis de Sensibilidad Local y análisis de Sensibilidad Global.

### A. Seguimiento (*Screening*)

Son métodos simplistas pero eficientes que permiten ordenar los Inputs/parámetros según su importancia relativa en el modelo, generalmente sin tener en consideración las posibles interacciones. Su utilidad es grande sobretodo en modelos altamente parametrizados, en los que con frecuencia, son unos pocos Inputs/parámetros los que condicionan fuertemente los resultados del modelo. Así, un proceso de *Screening* o seguimiento previo, ayuda a guiar una posterior aplicación de métodos más rigurosos.

### B. Sensibilidad Local

Los métodos locales emplean la información del gradiente, para cuantificar la sensibilidad del modelo alrededor de un valor específico del Input/parámetro. Estos métodos son el núcleo de muchos métodos de estimación de parámetros mediante aproximación por Solución Simple y Búsqueda Local.

### C. Sensibilidad Global

Evalúan la sensibilidad del modelo a los Inputs/parámetros, sobre el rango completo del espacio de valores del Input/parámetro, en lugar de hacerlo alrededor de un valor específico. Pertenecen a este grupo métodos basados en



análisis de regresión, descomposición de la varianza (método *FAST*, método *SOBOL*) y métodos *Parameter Bounding*.

## 8. Análisis Multimodelo

Consiste en la evaluación de un problema específico con un enunciado concreto, mediante distintos modelos posibles, los cuales consideran procesos alternativos, emplean distintos códigos o definen condiciones de contorno alternativas. Seguidamente se asignan puntuaciones al rendimiento de cada modelo, de manera que estos puedan ser ordenados y se les pueda asignar un peso mayor o menor peso según su puntuación. Dicho peso será tenido en cuenta a la hora de realizar predicciones sobre el problema analizado.

## 9. Redes Bayesianas

Las Redes Bayesianas son modelos gráficos probabilísticos que combinan distribuciones a priori con conocimiento general (de uno o más modelos) y datos específicos locales, para obtener un conjunto de distribuciones actualizado (a posteriori). En teoría, las Redes Bayesianas pueden abordar de forma simultánea la incertidumbre en los Inputs y en las Observaciones, las distribuciones de los parámetros reducibles e irreducibles y los errores en el código del modelo, en su estructura y los debidos a su resolución (Aguilera et al. 2011; Clark and Gelfand 2006; Clark 2004; Uusitalo 2007; Zhang et al. 2011). El desarrollo de una Red Bayesiana requiere:

- Definir un grafo acíclico direccionado que especifica una red de dependencias de probabilidad condicional.
- Definir para todos los nodos del grafo (fuentes de incertidumbre) una distribución Prior.
- Definir la función de verosimilitud (*likelihood*) y la estrategia de muestreo (por ejemplo MCMC) para inducir las distribuciones posteriores a partir de las distribuciones a priori.

## 3.-BASES METODOLÓGICAS

### 3.1.- ESTADÍSTICA BAYESIANA COMPUTACIONAL

#### 3.1.1.- INTRODUCCIÓN

En los primeros años del siglo XXI se ha incrementado notablemente el empleo de métodos bayesianos en la práctica de la estadística aplicada. El desarrollo de la estadística bayesiana computacional es la clave de este cambio en la estadística. Los estadísticos han conocido desde hace tiempo las claras ventajas que la aproximación bayesiana ofrece sobre la aproximación frecuentista.

Sin embargo, las soluciones bayesianas, fáciles en la teoría, se vuelven complicadas de llevar a la práctica. La **estadística bayesiana computacional** está cambiando esto gracias al empleo de algoritmos que permiten **muestrear la desconocida distribución posterior y usar esas muestras como base para la inferencia**.

En la **estadística clásica o frecuentista**, el parámetro es considerado un valor fijo pero desconocido. Así, la evaluación del parámetro se basa en una función de densidad de probabilidad (en adelante FDP) en el espacio muestral. El espacio muestral es el conjunto de todos los valores observados posibles. La probabilidad se interpreta como una frecuencia relativa en el largo plazo, sobre todos los valores en el espacio muestral, dado el desconocido parámetro.

En la **estadística bayesiana** todo se basa en el conocido teorema descubierto por el reverendo Thomas Bayes. A diferencia de la estadística frecuentista, la evaluación del parámetro se basa en su distribución posterior, situada en el espacio de parámetros. El Teorema de Bayes es una muy inteligente actualización de la fórmula de probabilidad condicional, mediante la cual se pueden actualizar las probabilidades de eventos no observados, dados



otros eventos relacionados que sí se han observado. Como **ideas básicas de la aproximación bayesiana** se tiene que:

- En estadística bayesiana los parámetros son considerados como variables aleatorias. Esto contrasta con la aproximación frecuentista, en la cual, los parámetros se consideran constantes fijas y desconocidas.
- El Teorema de Bayes es un algoritmo de actualización que nos proporciona la distribución de probabilidad posterior de los parámetros (en adelante, “la posterior”), mediante la combinación de dos fuentes de información sobre dichos parámetros:
  - La distribución de probabilidad a priori de los parámetros (en adelante “la prior”). Proporciona el peso relativo de nuestra certeza sobre el valor de los parámetros, antes de las observaciones.
  - Los datos observados. Mediante la función de verosimilitud de los parámetros, dadas las observaciones, se considera el peso relativo que dichas observaciones aportan al valor de los parámetros.

Por tanto, el Teorema de Bayes resume en la posterior, todo el conocimiento que tenemos sobre los parámetros, y no solo el originado por las observaciones. En una interpretación directa de la inferencia bayesiana clásica, la información “Prior” o a priori, representa la honesta opinión que el investigador tiene sobre lo desconocido (Laine 2008).

En la práctica, evaluar la distribución “Prior” de, por ejemplo, un parámetro, puede verse como un proceso iterativo semejante a la construcción y mejora de un modelo. Así la distribución “Prior” no solo se incorpora al modelo cuando construimos el modelo. De hecho la “Posterior” obtenida en la simulación actual se empleará como “Prior”, cuando simulemos para las nuevas observaciones que se vayan añadiendo al modelo.

En muchos ámbitos científicos, la idea de una componente subjetiva de la probabilidad parece resultar poco apropiada. Sin embargo la información a priori, basada en experiencia empírica previa, propia o basada en referencias,





no tiene porqué resultar más subjetiva que la consideración de un P-Value (umbral de significación estadística) mayor o menor, en las clásicas pruebas de hipótesis de la estadística frecuentista. En última instancia el modelador que hace uso de la inferencia Bayesiana, ante una falta de criterio o conocimiento previo, podría optar por el empleo de “Priors” no-informativas. Así mismo, también hay quien cuestiona si es apropiado o incluso posible, asignar distribuciones estadísticas a los valores de los parámetros de un modelo, cuando en el mundo real esos valores no son variables aleatorias. Ante este cuestionamiento, hay que decir que si bien los modelos pretenden describir la realidad, este **proceso de inferencia estadística se realiza sobre los modelos, no sobre la propia realidad.**

Se acaba de hacer mención indirecta al “conflicto” existente en el ámbito del cálculo estadístico, entre los defensores de la escuela que ha dominado la estadística del Siglo XX (frecuentista) y los que abogan por el empleo del “nuevo pensamiento” Bayesiano, los cuales han hecho resurgir esta forma de hacer estadística principalmente durante el último cuarto del siglo pasado. El motivo fundamental que ha potenciado esta corriente, es el rápido y continuo incremento experimentado en la potencia de cálculo de los ordenadores, que permite abordar problemas que hasta ahora se quedaban en su mera descripción teórica. Para que el lector pueda hacerse una idea de lo que defienden ambas visiones, y concluir por sí mismo, son buenas referencias (Clark and Gelfand 2006; Clark 2004; Dennis 1996; S. Ríos and D. Ríos 2001; Silva 2001). El autor de este trabajo se queda con la siguiente reflexión sobre el controvertido tema. El verdadero test de un modelo es su habilidad para predecir las observaciones, viejas y nuevas (Laine 2008). La inferencia bayesiana no es más que una herramienta que persigue ese objetivo ajustándose al paradigma de Popper (**Falsacionismo o Racionalismo Crítico**), en Popper (1957): *Las observaciones refutarán tanto un mal modelo como unas mal seleccionadas “Priors”.*

### 3.1.2.- FORMULACIÓN DEL TEOREMA DE BAYES

Con el propósito de introducir parte de la terminología y nomenclatura que se empleará en sucesivos apartados del trabajo, se escribe el Teorema de Bayes de la siguiente forma:

$$p(\underline{\theta}|y^{obs}) = \frac{p(y^{obs}|\underline{\theta}) \times p(\underline{\theta})}{p(y^{obs})} \quad \text{Ec. 3.1}$$

donde:

1.  $\underline{\theta}$  es el vector de parámetros. Hipótesis Nula inferida antes de que la nueva evidencia ( $y^{obs}$ ) estuviese disponible.
2.  $p(\underline{\theta})$  es la distribución marginal a priori de  $\underline{\theta}$ .
3.  $p(y^{obs}|\underline{\theta})$  es la probabilidad condicional de que se cumpla la evidencia, dado que la hipótesis  $\underline{\theta}$  es cierta. Su expresión analítica es equivalente a la Función de Verosimilitud de los parámetros  $L(\underline{\theta})$ .
4.  $p(y^{obs})$  es la probabilidad marginal de  $y^{obs}$ . Probabilidad de observar la evidencia, bajo todas las hipótesis de  $\underline{\theta}$ , mutuamente excluyentes. El cálculo de esta constante de escalamiento se realiza según la Ec. 3.2.

$$p(y^{obs}) = \int_{\Theta} p(y^{obs}|\underline{\theta}) \times p(\underline{\theta}) d\underline{\theta} \quad \text{Ec. 3.2}$$

Siendo  $p(y^{obs}|\underline{\theta}) \times p(\underline{\theta})$  la distribución de probabilidad conjunta de las observaciones y los parámetros  $p(y^{obs}, \underline{\theta})$ . Este valor normaliza el numerador de la Ec. 3.1 convirtiéndolo en una correcta función de densidad de probabilidad. El cálculo de esta integral convierte al cálculo computacional bayesiano en un reto a resolver mediante algoritmos especiales.

5.  $p(\underline{\theta}|y^{obs})$  es la probabilidad a posteriori (tras considerar la evidencia) de  $\underline{\theta}$ .

6. El impacto que la evidencia tiene sobre la creencia de la hipótesis de  $\underline{\theta}$  viene representado por el factor  $p(y^{obs} | \underline{\theta}) / p(y^{obs})$ . Este factor altera la creencia a priori en la hipótesis de  $\underline{\theta}$ .

### 3.1.3.- OBTENCIÓN DE LA POSTERIOR

Encontrar la posterior es fácil en la teoría pero no lo es tanto en la práctica. El Teorema de Bayes expresado en su forma no escalada se expresa según la siguiente ecuación:

$$g(\underline{\theta} | \underline{y}) \propto g(\underline{\theta}) \cdot f(\underline{y} | \underline{\theta}) \quad \text{Ec. 3.3}$$

Es decir, la posterior es proporcional a prior veces la función de verosimilitud. La posterior no escalada presenta la misma forma que la verdadera posterior.

Para poder realizar inferencia sobre los parámetros se necesita la posterior, la cual es obtenida mediante el escalamiento de la posterior no escalada, de forma que su integral tome valor la unidad.

La constante de escala tiene la expresión de la Ec. 3.2. Esta integral p-dimensional ( $p=n^{\circ}$  de parámetros) presenta solución analítica solamente cuando la FDP de las observaciones procede de una familia exponencial, y la prior procede a su vez de una familia que es conjugada de la FDP de las observaciones.

En el resto de ocasiones, la integral debe ser abordada numéricamente, siendo incluso así, difícil de obtener en el caso de espacios paramétricos de elevada dimensión. Por tanto, si bien es fácil la obtención de la posterior no escalada, el problema surge al querer obtener la verdadera posterior.

Los **métodos bayesianos computacionales** se basan en el desarrollo de algoritmos que permiten extraer muestras de la posterior, incluso cuando solo conocemos su versión no escalada.

De forma general hay dos tipos de algoritmos que permiten esto:

A. **Métodos de Monte Carlo directos:** en ellos se extrae una muestra aleatoria de una FDP fácilmente muestreable. Dicha muestra se altera mediante la aceptación de solamente algunos de los valores, en la muestra final. Los valores aceptados serán una muestra aleatoria de la posterior. Son métodos ineficientes (bajo ratio de aceptación) en espacios paramétricos de dimensión mayor que uno. *Acceptance-Rejection-Sampling* y *Sampling-Importance-Resampling* son métodos de este tipo.

B. **Métodos Markov Chain Monte Carlo** (en adelante MCMC): configuran una cadena de Markov que tiene como distribución de equilibrio a la posterior. Simulando la cadena un número suficiente de iteraciones se extrae una muestra de la posterior. Metropolis-Hastings y el muestreador de Gibbs son los dos principales algoritmos de este tipo de métodos.

La aproximación computacional a la estadística bayesiana no sigue la vía de integrar el factor de escala numéricamente. En su lugar emplea la técnica de Monte Carlo para la extracción de muestras aleatorias, las cuales aproximarán la verdadera posterior cuando el tamaño de la muestra sea suficientemente grande. La ventaja de esta forma de atacar el problema, es no estar limitado al empleo de modelos cuya formulación deba ser tratable analítica o numéricamente, en perjuicio de su buena aproximación a la realidad subyacente.

#### 3.1.4.- ESTIMACIÓN MÁXIMO-VEROSÍMIL Y ESTIMACIÓN BAYESIANA

Se define como Universo Inferencial, al producto cartesiano del espacio de parámetros (en adelante EP) y el espacio muestral (en adelante EM). El EP presenta dimensión  $P$  (número de parámetros). El EM presenta dimensión  $N$  (número de elementos de la muestra observada  $y^{\text{obs}}$ ). Se define así mismo la FDP  $f(\underline{y} | \underline{\theta})$ , que proporciona la probabilidad de  $\underline{y}$  dado  $\underline{\theta}$ . Esta función forma una superficie definida en todo el universo de inferencia. Como se observa en la figura 3.1, es una FDP en el EM pero no lo es en el EP.

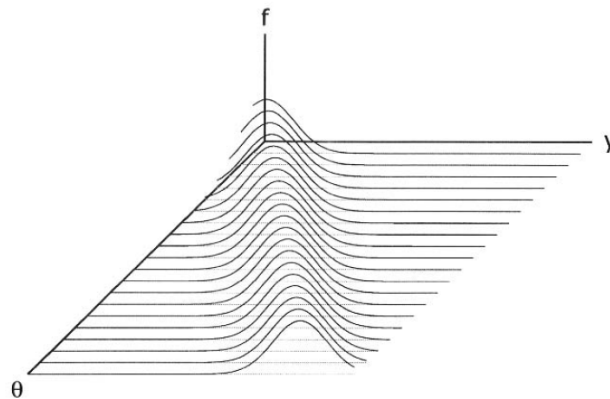


Figura 3.1. Superficie de Densidad de Observaciones (Fuente: Bolstad 2010)

Se define la Función de Verosimilitud tal y como la definió R.A. Fisher en 1922, como una función de los parámetros con idéntica expresión analítica que la FDP de las observaciones. Sin embargo, la Función de Verosimilitud no es una FDP. De forma gráfica, dicha función queda definida cortando la superficie de probabilidad de observaciones, mediante un plano que pasa por  $\underline{y}^{obs}$ , como se observa en la figura 3.2.

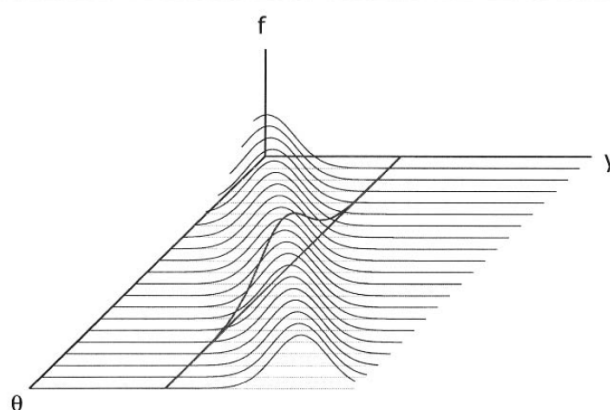


Figura 3.2. Superficie de Densidad de Observaciones con la Función de Verosimilitud (Fuente: Bolstad 2010)

### 3.1.4.1.-ESTIMACIÓN MÁXIMO-VEROSÍMIL

Para obtener una estimación del desconocido valor de los parámetros, y dado que la función de verosimilitud no es una FDP, Fisher decidió que la mejor manera de estimar el parámetro era elegirlo de forma que tuviese el valor más alto de la función de verosimilitud (la Moda). Este es el valor que hace que la muestra observada tenga máxima probabilidad de ocurrencia. Fisher llamaría a ese valor, estimador máximo verosímil (MLE).

### 3.1.4.2.-ESTIMACIÓN BAYESIANA

Con la aproximación bayesiana se debe tener una FDP prior definida en el EP antes de haber observado los datos. Esto requiere permitir dos **interpretaciones del concepto de probabilidad:**

1. En el espacio de parámetros, se define probabilidad en el sentido subjetivo de posibilidad o creencia.
2. En el espacio muestral de observaciones se define probabilidad en el tradicional sentido frecuentista.

La distribución conjunta de probabilidad de parámetros y observaciones se obtendrá con la multiplicación de los valores de la FDP observacional y los valores de las FDP prior. La superficie resultante se denomina “el modelo bayesiano completo” (*Figura 3.3*)

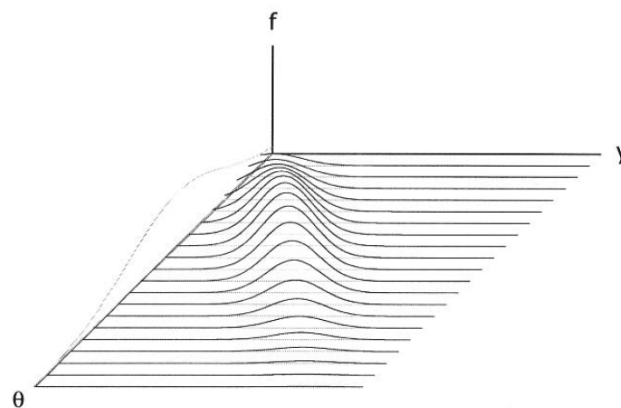


Figura 3.3. Distribución Conjunta de los parámetros y las observaciones. En el plano lateral se dibuja la Prior de los parámetros (Fuente: Bolstad 2010)

La FDP posterior se obtendrá mediante el corte de esa superficie con un plano que pasa por las observaciones  $y^{obs}$  (Figura 3.4).

Es decir, la función de verosimilitud y la posterior son el resultado de cortar con el mismo hiperplano dos superficies diferentes. Dado que la posterior sí es una FDP, se podrá emplear como estimador del parámetro a la Media de dicha distribución. Se considera que la posterior es el objetivo de inferencia en la aproximación bayesiana, mientras que en la aproximación frecuentista coexisten varios tipos de inferencia (estimación puntual, estimación de intervalo, test de hipótesis).

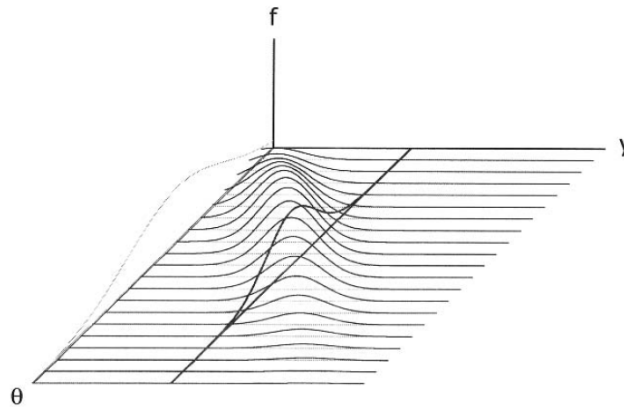


Figura 3.4. Distribución posterior de los parámetros (Fuente: Bolstad 2010)

Se da la circunstancia especial, de que la superficie de la FDP observacional y la superficie del modelo bayesiano completo, coinciden cuando la prior es de tipo no informativo. En este caso la función de verosimilitud y la posterior presentan idéntica forma. Sin embargo, el MLE será la Moda mientras que el estimador bayesiano será la Media.

#### 3.1.4.3.-INFERENCIA EN PRESENCIA DE PARÁMETROS NUISANCE

Se denominan parámetros *Nuisance* (incómodos) a aquellos parámetros sobre los que no interesa realizar inferencia, si bien son parte de la FDP conjunta. Solo se necesita estar seguros de que dichos parámetros no interfieren en el proceso de inferencia de los parámetros de interés.



La solución a este problema en la aproximación bayesiana es la marginalización de los parámetros de interés, mediante la integración de la FDP conjunta respecto al parámetro *Nuisance*.





## 3.2.- CADENAS DE MARKOV

### 3.2.1.- INTRODUCCIÓN

Las cadenas de Markov son un tipo de proceso estocástico discreto en el tiempo. Dicho proceso se mueve en torno a un conjunto de posibles valores llamado Espacio de Estados del Proceso. La principal propiedad de las cadenas de Markov es su falta de memoria: dados los estados, pasado y presente del proceso, el estado futuro del proceso solo depende del estado presente. A esa propiedad se la conoce como Propiedad de Markov.

### 3.2.2.- CADENAS DE MARKOV INVARIANTES EN EL TIEMPO Y CON ESPACIO DE ESTADOS FINITO

Las probabilidades de transición entre los K estados de un proceso de Markov pueden expresarse en forma de matriz, donde  $p_{ij}^{[n-1,n]} = P[x^n=j|x^{n-1}=i]$  es la probabilidad de que el proceso vaya del estado “i” en el paso temporal “n-1” al estado “j” en el paso temporal “n”.

Se denominan cadenas de Markov Homogéneas a aquellas en las que las probabilidades de transición solo dependen de los estados y no del tiempo. En este tipo de cadenas la matriz de transición se compone de elementos  $p_{ij} = P[x^{n+1}=j|x^n=i]$  que indican la probabilidad de ir al estado “j” desde el estado “i”, en un paso, para cualquier instante “n”.

Se define matriz de probabilidades de transición en N pasos, a la matriz  $P^{[N]}$  cuyo elemento  $p_{ij}$  representa la probabilidad del proceso, de pasar del estado “i” al estado “j” en el transcurso de N pasos o instantes. Se puede demostrar (Bolstad 2010) que gracias a la propiedad de Markov se cumple que  $P^{[N]} = P^N$ , conocida como ecuación de Chapman-Kolmogorov, la cual indica que la matriz de probabilidad de transición en N pasos es igual a la N-sima potencia de la matriz de probabilidad de transición en un paso.

### 3.2.3.- DISTRIBUCIÓN DE EQUILIBRIO Y ECUACIÓN DE ESTADO ESTACIONARIO

Sea una cadena o proceso de Markov que puede adoptar un número “m” de estados. Se define como **Distribución de Probabilidad de Ocupación (DPO)** en el instante “t”, al vector  $\alpha^{(t)} = [\alpha_1^{(t)}, \dots, \alpha_m^{(t)}]$  donde  $\alpha_i^{(t)}$  es la probabilidad de que la cadena esté en el estado “i”, en el instante “t”.

Se demuestra (Bolstad 2010) que dada la DPO  $\alpha^{(0)}$ , se puede obtener la correspondiente DPO de los estados en un instante “n” mediante la expresión  $\alpha^{(n)} = \alpha^{(0)} \times P^n$ . Es fácil demostrar ( $P^n = I \times P^n$ ) que la i-ésima fila de  $P^n$  es la DPO en el instante “n”, si suponemos que en el instante inicial la cadena adoptaba el valor del estado “i”.

Si una cadena de Markov posee una DPO que es independiente de su estado inicial, a esta DPO se la denomina **Distribución de Equilibrio (DE)** de la cadena.

$$\pi = \lim_{n \rightarrow \infty} \alpha^{(n)} \quad \text{Ec. 3.4}$$

No todas las cadenas (p.e. las cadenas periódicas) presentan una DE. Igualmente distintas cadenas pueden presentar la misma DE. Se tiene que cuando una cadena presenta una distribución de equilibrio  $\pi$ , entonces cumple con la **Ecuación de Estado Estacionario**:

$$\underline{\pi} = \underline{\pi} \times \underline{P} \quad \text{Ec. 3.5}$$

La anterior afirmación no tiene por qué satisfacerse en sentido inverso. La Ecuación de Estado Estacionario expresa que la probabilidad de “caer” en cada estado en régimen estacionario es igual a la probabilidad estacionaria de dicho estado. De la anterior ecuación se deduce que la probabilidad estacionaria de entrar en el estado “j” desde el estado “i” será  $\pi_i \times P_{ij}$ . Igualmente la probabilidad estacionaria de salir del estado “j” hacia el estado “i” será  $\pi_j \times P_{ji}$ . De igual

manera se puede deducir que la probabilidad estacionaria de estar en el estado “j” será  $\pi_j$ , igual al sumatorio para todos los “m” estados, del producto  $\pi_i \times P_{ij}$ .

Todo lo establecido para cadenas con espacios finitos, es igualmente válido en general para cadenas con espacios de estados contables e infinitos.

### 3.2.4.- CLASIFICACIÓN DE ESTADOS DE UNA CADENA DE MARKOV

Según el comportamiento de la distribución de equilibrio, podremos clasificar los estados de la cadena. Se puede clasificar los estados de una cadena de Markov por la probabilidad de primer retorno (probabilidad de volver una vez al mismo estado en “n” pasos).

- Estado transitorio: existe una probabilidad finita mayor que cero de que la cadena no retorne a dicho estado. La cadena puede visitar el estado un número finito de veces antes de abandonarlo definitivamente.
- Estado persistente o recurrente: ocurre cuando hay una probabilidad igual a uno de que la cadena vuelva a dicho estado.
  - Estado recurrente nulo: se presenta cuando el tiempo medio de retorno se incrementa con la evolución de la cadena.
  - Estado recurrente positivo: estado cuyo tiempo medio de retorno es finito.

La citada clasificación de estados se puede explicar en función de la matriz de probabilidades de transición en “n” pasos, como se demuestra en la *Figura 3.5*

Type	Sum	Limit as $n \rightarrow \infty$
Transient	$\sum p_{i,i}^{[n]} < \infty$	$\lim p_{i,i}^{[n]} = 0$
Null recurrent	$\sum p_{i,i}^{[n]} = \infty$	$\lim p_{i,i}^{[n]} = 1/\pi_i = 0$
Positive recurrent	$\sum p_{i,i}^{[n]} = \infty$	$\lim p_{i,i}^{[n]} = 1/\pi_i > 0$

Figura 3.5. Clasificación de estados de una Cadena de Markov (Fuente: Bolstad 2010)

### 3.2.5.- ERGODICIDAD DE LAS CADENAS DE MARKOV

#### 3.2.5.1.-REDUCIBILIDAD E IRREDUCIBILIDAD

Se sabe que el estado “j” es alcanzable desde el estado “i” en “n” pasos cuando  $p_{ij}^n$  es mayor que cero, para algún “n”.

Se dice que un par de estados se comunica entre sí cuando se cumple que cada estado del par es alcanzable desde el otro. Un conjunto de estados dentro del cual son todos alcanzables entre sí pero no pueden alcanzar otro estado fuera del conjunto, se dice que es una clase comunicada de estados.

Se define como cadena de Markov Irreducible a aquella cadena compuesta por una clase comunicada de estados. En otro caso será una cadena Reducible. Se puede demostrar (Bolstad 2010) que todos los estados de una cadena Irreducible son del mismo tipo.

#### 3.2.5.2.-ERGODICIDAD

Se definen como cadenas de Markov Ergódicas a aquellas que cumplen:

- Tener probabilidades de transición invariantes en el tiempo.
- Ser aperiódicas. Por tanto tienen DE.
- Ser Irreducibles.
- Sus estados son Recurrentes Positivos.

Las cadenas Ergódicas son muy importantes dado que en ellas se basan los métodos MCMC. Existen teoremas que demuestran que para este tipo de cadenas, el promedio en el tiempo de las características estadísticas de una realización simple, se aproxima al promedio de dichas características, en un instante particular, sobre todas las posibles realizaciones (el *Ensemble* o conjunto, *Figura 3.6*).

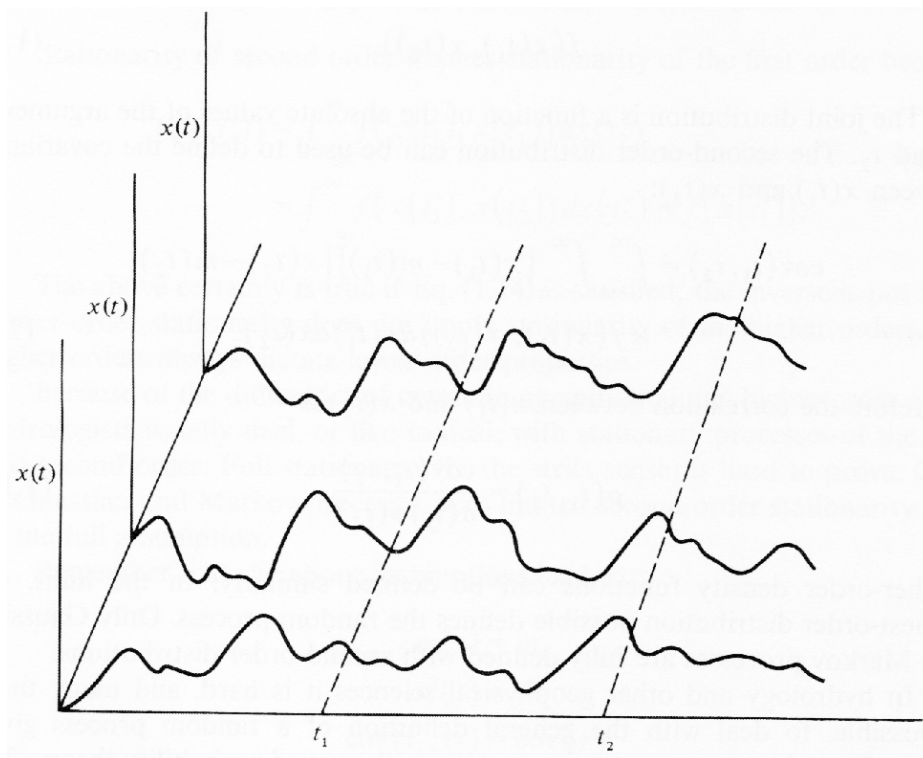


Figura 3.6. Concepto de *Ensemble* de un proceso aleatorio (Fuente: Bras and Rodríguez-Iturbe 1985)

### 3.2.6.- MUESTREO DE UNA CADENA DE MARKOV

La distribución estacionaria de una cadena de Markov Ergódica debe ser pensada como el promedio sobre todas las realizaciones, de cualquier estadístico, calculado en un instante futuro lejano.

Para que la cadena alcance el estado estacionario, y por consiguiente su DE, se debe permitir que la cadena sea simulada durante un tiempo de calentamiento suficientemente largo. ¿Cuál es el tiempo que necesita una



cadena de Markov para que el muestreo sobre su realización represente al promedio sobre el conjunto de realizaciones?

La respuesta es que no hay manera de saber “a priori” cuál es el número necesario de simulaciones para alcanzar el equilibrio de la cadena. Más adelante se verán métodos “a posteriori” para delimitar el número de simulaciones de calentamiento necesarias.

### 3.2.7.- CADENAS DE MARKOV REVERSIBLES EN EL TIEMPO. CONDICIÓN DE BALANCE DETALLADO

Sea una cadena de Markov, se define cadena de Markov “hacia atrás” a la cadena formada tomando los elementos de la cadena original y colocándolos en orden temporal inverso. Sea  $q_{ij}$  su matriz de probabilidades de transición (en adelante MPT). Cuando la cadena está en estado estacionario se tiene que (Bolstad 2010):

$$\pi_i \cdot q_{ij} = \pi_j \cdot P_{ji} \quad \text{Ec. 3.6}$$

Se dice que una cadena de Markov es reversible cuando las cadenas “hacia adelante” y “hacia atrás” tienen las mismas MPT, es decir,  $q_{ij}=p_{ij}$  para todos los estados “i”, “j”. De ello se deduce que en una **cadena reversible, en su estado estacionario**, se cumple la denominada **Ecuación de Balance Detallado (EBD)**:

$$\pi_i \cdot P_{ij} = \pi_j \cdot P_{ji} \quad \text{Ec. 3.7}$$

Se tiene que  $\pi_i \times p_{ij}$  es la probabilidad que permite ir del estado “i” al “j” en estado estacionario. Similarmente  $\pi_j \times p_{ji}$  es la probabilidad de ir del estado “j” al “i” en el régimen estacionario. La EDB expresa que, para cada par de estados, el flujo entre ellos se iguala. Y lo que es fundamental para resolver el problema en concreto que se plantea en esta tesina: **una cadena de Markov reversible**,

cuya matriz de transición  $P$  satisfaga la EDB, tendrá como distribución de equilibrio a  $\pi$ .

### 3.2.8.- CADENAS DE MARKOV CON ESPACIO DE ESTADOS CONTINUO

Se puede definir un espacio de estados que contenga todos los posibles valores (infinitos) en una dimensión (intervalo), en dos dimensiones o un hiperespacio  $n$ -dimensional. En esos casos no es posible emplear el concepto de probabilidad de transición entre pares de estados. En su lugar se definen las probabilidades de transición desde cada estado  $X$  hacia cada *conjunto medible de estados*  $A$ .

Se llama Kernel de Transición de la cadena Markov a la función que proporciona la probabilidad de que la cadena adopte un estado perteneciente al conjunto  $A$ , dado que actualmente adopta el estado  $X$ , para todos los posibles conjuntos de estados  $A$  y estados  $X$ . Se asume, al igual que se hizo con la matriz de probabilidades de transición que el Kernel es invariante, es decir, no dependerá del tiempo.

Todos los resultados y propiedades obtenidos para el caso discreto siguen siendo válidos para el caso de espacio de estados continuo, si bien las expresiones presentan modificaciones. En la práctica, para construir el Kernel de transición habrá que tratar con FDP's continuas junto con valores de probabilidad de transición discretos, empleados para expresar la probabilidad de permanencia en un mismo estado. Al igual que en el caso discreto, se tendrá un Kernel de transición en un paso, dos pasos, en "n" pasos. Así mismo, la distribución de equilibrio  $g(v)$  será una FDP (Bolstad 2010). En el **caso continuo** la **Ecuación de Balance Detallado**, se puede expresar de la forma:

$$\boxed{g(x) \cdot f(v|x) = g(v) \cdot f(x|v)} \quad \text{Ec. 3.8}$$

Siendo  $g(x)$  la densidad de probabilidad estacionaria del estado  $X$ , y  $f(v|x)$  la densidad de probabilidad de las transiciones en un paso, para todos los estados  $X$  y  $V$ .



### 3.3.- MUESTREO DE LA POSTERIOR. EL ALGORITMO MH

#### 3.3.1.- INTRODUCCIÓN

Los métodos de muestreo de tipo Monte Carlo directo (*ARS*, *SIR*, etc.) son métodos ineficientes cuando se tienen muchos parámetros y/o cuando la Prior es muy difusa respecto a la verosimilitud (Bolstad 2010).

Como alternativa a esos métodos para la generación de muestras de la posterior, un algoritmo inteligente de los años 50, incluido por la revista *Computing in Science & Engineering*, en el “TOP TEN” de los algoritmos más importantes del S.XX, llamado **algoritmo de Metrópolis-Hastings** (MH) por sus dos principales autores, proporcionó un método eficaz para tratar ese tipo de problemas. Se inició así una familia de algoritmos, conocida como métodos Markov Chain Monte Carlo (MCMC), la cual sigue mejorándose en la actualidad (Haario et al. 2006; Laloy and Vrugt 2012; Vrugt and Braak 2011; Vrugt, Braak, Diks, et al. 2009; Vrugt, Braak, and Gupta 2009).

Estos métodos entre los cuales destacan el mencionado algoritmo de Metrópolis-Hastings (MH), el muestreador de Gibbs y el muestreador por sustitución (algoritmo de aumentación de datos), no extraen la muestra de forma directa desde la posterior, sino que configuran una cadena de Markov en la cual su distribución de equilibrio sea la distribución posterior.

De esa manera, después de un tiempo suficiente de simulación de la cadena, necesario para alcanzar el equilibrio, cualquier valor extraído de la cadena aproximará una muestra aleatoria de la posterior de los parámetros, dadas las observaciones. Por tanto, el espacio de estados de la cadena coincidirá con el espacio de valores de los parámetros.

Habrá que tener en cuenta, que valores tomados de la cadena en posiciones temporales cercanas, estarán altamente correlacionados (autocorrelación de orden 1 o Markoviana). Por esa razón para que la muestra de la posterior se aproxime a una muestra independiente e idénticamente



distribuida (i.i.d.), se deberá muestrear los valores de la cadena en puntos temporales suficientemente separados.

### 3.3.2.- ECUACIÓN DE ESTADO ESTACIONARIO EN UN ESPACIO DE PARÁMETROS CONTINUO

Se ha visto que la distribución de equilibrio de una cadena de Markov Ergódica es solución de la Ecuación de Estado Estacionario discreta (Ec. 3.5) donde  $\underline{\pi}$  es la distribución de equilibrio y  $\underline{P}$  es la matriz de transición en un paso.

La **Ecuación de Estado Estacionario en el espacio de parámetros ( $\theta$ )** se escribe:

$$\int_A \pi(\theta) d\theta = \int \pi(\theta) \cdot P(\theta, A) d\theta \quad \text{Ec. 3.9}$$

Donde:

- $\pi(\theta)$  es la distribución de probabilidades en el equilibrio, del parámetro ( $\theta$ ).
- La parte izquierda de la ecuación es la probabilidad total estacionaria del subconjunto medible A, de valores del parámetro.
- $P(\theta, A)$  es el **Kernel de Transición** de la cadena. Es una función que proporciona la probabilidad de pasar desde un punto en el espacio p-dimensional de parámetros hacia un subconjunto medible A. Los **subconjuntos medibles** son todos aquellos que se puedan construir juntando intervalos de la forma  $(\theta_1 \leq x_1, \theta_2 \leq x_2, \dots, \theta_p \leq x_p)$ . Así, la probabilidad del conjunto medible se podrá calcular mediante las funciones de distribución acumulada conjunta de los parámetros, empleando las reglas de la probabilidad.

En general, un Kernel de Transición estará compuesto por dos partes. La primera parte “moviliza” puntos del espacio de parámetros

hacia otros puntos del mismo espacio, siendo similar a una FDP. La segunda parte “moviliza” cada punto hacia sí mismo con una probabilidad finita.

- La parte derecha de la ecuación representa el flujo de probabilidad hacia el conjunto A, en estado estacionario.

Se puede explicar la Ecuación de Estado Estacionario de la siguiente manera: la probabilidad de un conjunto de estados A, es igual a la suma (integral) ponderada de las probabilidades de adoptar ese conjunto A de estados (en un paso), partiendo desde todos los restantes posibles estados. El coeficiente de ponderación que corresponde a cada estado posible de partida, es su densidad de probabilidad estacionaria.

La clave que va a permitir muestrear la posterior, está en considerar como Distribución de Equilibrio  $\pi(\theta)$ , a la densidad posterior  $g(\theta|y)$ . En este punto cabe recordar que generalmente, solo se conocerá la posterior en su versión no escalada.

$$g(\theta | y) \propto g(\theta) \cdot L(\theta) \quad \text{Ec. 3.10}$$

Afortunadamente se verá que mediante el algoritmo MH, la versión no escalada es lo único que se necesita conocer.

### 3.3.3.- REVERSIBILIDAD Y ECUACIÓN DE BALANCE

Para conseguir que la distribución estacionaria de la cadena de Markov coincida con la posterior, se debe encontrar un Kernel de Transición  $P(\theta, A)$  que satisfaga la Ecuación de Estado Estacionario:

$$\int_A g(\theta | y) d\theta = \int g(\theta | y) P(\theta, A) d\theta \quad \text{Ec. 3.11}$$

En apartados anteriores se estableció que encontrar una matriz de probabilidades de transición  $\underline{P}$  que cumpliera con la Ecuación de Balance Detallado (Ec.3.7), era condición suficiente para que la cadena de Markov

tuviese a  $\underline{\pi}$  como distribución de equilibrio. Por tanto, si se encuentra un Kernel de Transición que balancee el flujo estacionario entre estados, entonces la distribución de equilibrio de la cadena Ergódica será la posterior  $g(\theta|y)$ .

Sea  $q(\theta, \theta')$  una distribución de probabilidad **candidata (o propuesta)** para ser parte del Kernel de Transición. Así,  $q(\theta, \theta')$  generará un valor  $\theta'$  partiendo de un valor inicial  $\theta$ . Se puede demostrar (Bolstad 2010) que si para todo par de estados  $\theta, \theta'$  la distribución candidata satisface la **Condición de Reversibilidad**:

$$\boxed{g(\theta | y) \cdot q(\theta, \theta') = g(\theta' | y) \cdot q(\theta', \theta)} \quad \text{Ec. 3.12}$$

Entonces  $g(\theta|y)$  será la distribución de equilibrio de la cadena cuyo Kernel de Transición es

$$\boxed{P(\theta, A) = \int_A q(\theta, \theta') d\theta' + r(\theta) \cdot \delta_A(\theta)} \quad \text{Ec. 3.13}$$

$$\boxed{r(\theta) = 1 - \int_A q(\theta, \theta') d\theta'} \quad \text{Ec. 3.14}$$

$$\delta_A(\theta) = 1, \theta \in A \quad \delta_A(\theta) = 0, \theta \notin A$$

Donde  $r(\theta)$  es la probabilidad de que la cadena de Markov permanezca en el estado  $\theta$ .  $\delta_A(\theta)$  es la función indicadora del conjunto A, la cual indica si el estado inicial pertenece o no a dicho conjunto.

### 3.3.4.- EL ALGORITMO METROPOLIS-HASTINGS

La mayoría de distribuciones candidatas o propuestas, no satisfacen la condición de Reversibilidad, es decir, la probabilidad de moverse de  $\theta$  a  $\theta'$  no es la misma que la de moverse en la dirección inversa.

Metropolis and Rosenbluth (1953) descubrieron un algoritmo capaz de encontrar el Kernel de una cadena de Markov con una distribución de equilibrio

deseada. Metropolis introdujo el concepto de **Probabilidad de Movimiento**  $\alpha(\theta, \theta')$  de manera que se equilibrase la Ecuación de Balance.

$$\alpha(\theta, \theta') = \min \left[ 1, \frac{g(\theta' | y) \cdot q(\theta, \theta')}{g(\theta | y) \cdot q(\theta', \theta)} \right] \quad \text{Ec. 3.15}$$

Con esta genial solución, se consigue que no se necesite conocer la posterior  $g()$  en su forma exacta, ya que el algoritmo funciona igualmente empleando la posterior no escalada, puesto que las constantes de escala se anulan al estar en numerador y denominador. Lo mismo ocurre con la distribución candidata  $q()$ , de la que solamente se necesitará su forma.

Por tanto, la distribución **candidata revisada**  $\alpha(\theta, \theta') \times q(\theta, \theta')$  satisface la Condición de Reversibilidad de la cadena, siendo así  $g(\theta|y)$  su Distribución de Equilibrio. La expresión del Kernel de Transición será similar a la de la Ec.3.13, pero con la distribución candidata revisada.

Los pasos del **algoritmo MH** son:

1. Comienzo con un punto inicial  $\theta_0$  en el espacio paramétrico.
2. Bucle  $n=1, \dots, n^0$  simulaciones cadena
  - a. Extrae un valor  $\theta'$  candidato mediante la distribución candidata  $q[\theta(n-1), \theta]$
  - b. Calcula la probabilidad de movimiento (o probabilidad de aceptación) hacia el valor candidato, mediante  $\alpha[\theta(n-1), \theta']$ . Para ello se necesitan la distribución candidata y la posterior no escalada ( Ec. 3.10).
  - c. Extrae muestra  $u \in U(0, 1)$ .
  - d. Decisión de cambio de estado (valor de  $\theta$ ) en la cadena. Si  $u < \alpha[\theta(n-1), \theta']$  entonces  $\theta(n) = \theta'$ . En otro caso,  $\theta(n) = \theta(n-1)$ . Es inmediato advertir que si la distribución candidata  $q(\theta, \theta')$  coincide con la forma

de la posterior, entonces se tiene que  $\alpha(\theta, \theta')=1$ . En ese caso todos los valores propuestos (candidatos) serán aceptados. En general, cuanto más se parezca la candidata a la posterior, más valores propuestos serán aceptados y mayor será el **Ratio de Aceptación**.

### 3.3.4.1.-CANDIDATAS DE TIPO *RANDOM WALK*

Metropolis (1953) consideró que el algoritmo que desplegaba los valores de la cadena de Markov, tuviese como distribución candidata a una distribución que quedaba centrada en el actual valor de la cadena. De esa forma,  $q(\theta, \theta')=q_1(\theta'-\theta)=q_1(\theta-\theta')$  es simétrica respecto del valor actual del parámetro  $\theta$ .

Por tanto la expresión de la probabilidad de aceptación queda simplificada como sigue:

$$\alpha(\theta, \theta') = \min \left[ 1, \frac{g(\theta' | y)}{g(\theta | y)} \right] \quad \text{Ec. 3.16}$$

Esta expresión implica que:

1. Si el valor de probabilidad posterior del valor candidato  $\theta'$  es mayor que el del valor actual  $\theta$ , entonces la probabilidad de aceptación será 1. Esta condición obliga a que la cadena se mueva en sentido *uphill* dentro de la distribución posterior del parámetro.
2. Si la probabilidad posterior de  $\theta'$  es menor que la de  $\theta$ , la probabilidad de aceptación será menor que 1 y tanto mayor cuanto mayor sea el ratio  $\theta'/\theta$  de dichas probabilidades. Habrá por tanto alguna probabilidad de que la cadena evolucione en sentido *downhill*.

Este comportamiento dota a la cadena de Markov de tipo *Random Walk* con las siguientes características:

- La cadena es capaz de moverse a lo largo de todo el espacio de parámetros, dado que presenta movimientos *uphill* y *downhill*.

- La cadena presentará un buen ratio de aceptación, si bien los movimientos en el espacio de parámetros serán cortos. Esto repercutirá en un mayor tiempo (nº de simulaciones) necesario para que la cadena explore todo el espacio de parámetros.
- La evolución de la cadena hará que tome ocasionalmente valores en las colas de la distribución posterior, si bien no permanecerá mucho tiempo en ellas debido a la principal tendencia *uphill* que la devolverá hacia las regiones de mayor densidad de probabilidad.

#### 3.3.4.2.-CANDIDATAS INDEPENDIENTES

Hastings (1970) introdujo las cadenas de Markov en las que la distribución candidata generadora, no dependía de la posición actual de la cadena, es decir,  $q(\theta, \theta') = q_2(\theta')$ .

La función candidata  $q_2(\theta')$  debe dominar, incluso en las colas, a la distribución posterior no escalada, requerimiento similar al que se debe cumplir en el método directo *ARS* (Bolstad 2010). Es decir, la distribución candidata permanecerá invariable en posición durante toda la simulación de la cadena.

Este tipo de cadenas presentan las siguientes características:

- Tienen un ratio de aceptación mucho menor que en el caso *Random Walk*.
- La evolución sobre el espacio de parámetros se realiza mediante saltos mayores que en el caso *Random Walk*. Por tanto, se explorará el espacio de parámetros en menos tiempo.

En las siguientes figuras se pueden comparar ambos tipos de algoritmos, observando cómo variables diferenciadoras principales, la capacidad exploratoria del espacio de parámetros y el ratio de aceptación.

Draw	Current value	Candidate	$\alpha$	$u$	Accept
1	2.000	1.767	1.000	.773	yes
2	1.767	1.975	.804	.933	no
3	1.767	.547	1.000	.720	yes
4	.547	1.134	.659	.240	yes
5	1.134	1.704	.553	.633	no
6	1.134	-.836	1.000	.748	yes

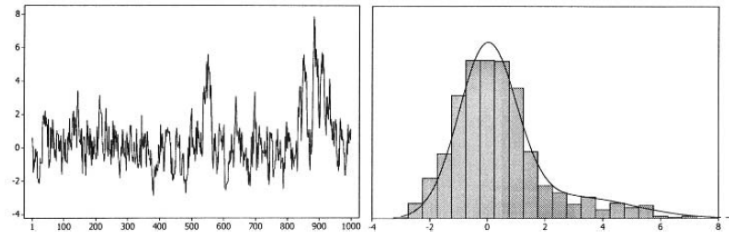


Figura 3.7. Resumen de las 6 primeras extracciones de una cadena Random Walk, y aspecto de la Traza de las 1000 simulaciones (Fuente: Bolstad 2010)

Draw	Current value	Candidate	$\alpha$	$u$	Accept
1	.4448	-2.7350	.0402	.5871	no
2	.4448	2.3553	.2537	.0987	yes
3	2.3553	.1611	1.000	.7896	yes
5	.1611	-1.5480	.3437	.6357	no
4	.1611	-1.3118	.4630	.1951	yes
6	-1.3118	-2.6299	.1039	.4752	no

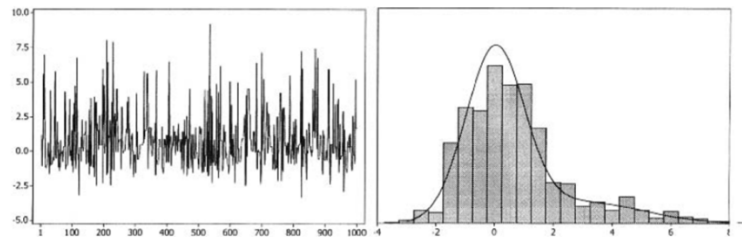


Figura 3.8. Resumen de las 6 primeras extracciones de una cadena con candidata Independiente y aspecto de la Traza de las 1000 simulaciones (Fuente: Bolstad 2010)

### 3.4.- INFERENCIA ESTADÍSTICA A PARTIR DE MUESTRAS MCMC

#### 3.4.1.- INTRODUCCIÓN

Como se ha visto con anterioridad, una vez ha comenzado la simulación de una cadena de Markov mediante el algoritmo MH, y después de un tiempo lo



suficientemente largo al cual llamaremos **periodo de calentamiento**, el estado de la cadena no dependerá ya del estado inicial con el que se inició la simulación. La cadena habrá alcanzado su distribución de equilibrio, y cualquier muestra extraída de la cadena será una muestra de la posterior.

Sin embargo, una secuencia de extracciones de la cadena, tal cual se va generando no es una muestra aleatoria, sino una muestra dependiente a partir de la cual no se deben realizar cálculos de inferencia.

Para conseguir de forma aproximada una muestra aleatoria deberemos realizar un proceso de adelgazamiento (*Thinning*) de la cadena.

#### 3.4.2.- PROPIEDADES DE MEZCLADO DE LA CADENA

Se dice que una cadena presenta unas **buenas propiedades de mezclado**, si puede alcanzar en pocos pasos cualquier punto del espacio de parámetros, sea cual sea el punto inicial de partida de la cadena. En general las cadenas cuyo tipo de muestreo emplea candidatas de tipo Independiente presentan mejor mezclado que las que utilizan el *Random Walk*.

Es importante advertir que para que las operaciones de inferencia realizadas sobre la posterior sean correctas, la cadena de Markov debe haber tenido tiempo de explorar todo el espacio de parámetros, de manera que las colas de la distribución posterior hayan podido ser muestreadas convenientemente.

En cadenas cuyas propiedades de mezclado no son buenas, el proceso exploratorio requerirá más tiempo. Así mismo, en distribuciones posteriores multimodales, también se necesitará más tiempo de exploración, para que la cadena sea capaz de visitar todas las zonas de alta densidad de probabilidad, teniendo para ello que atravesar zonas de probabilidad reducida. En la siguiente figura se presenta un ejemplo de lo que son un buen y mal mezclado.



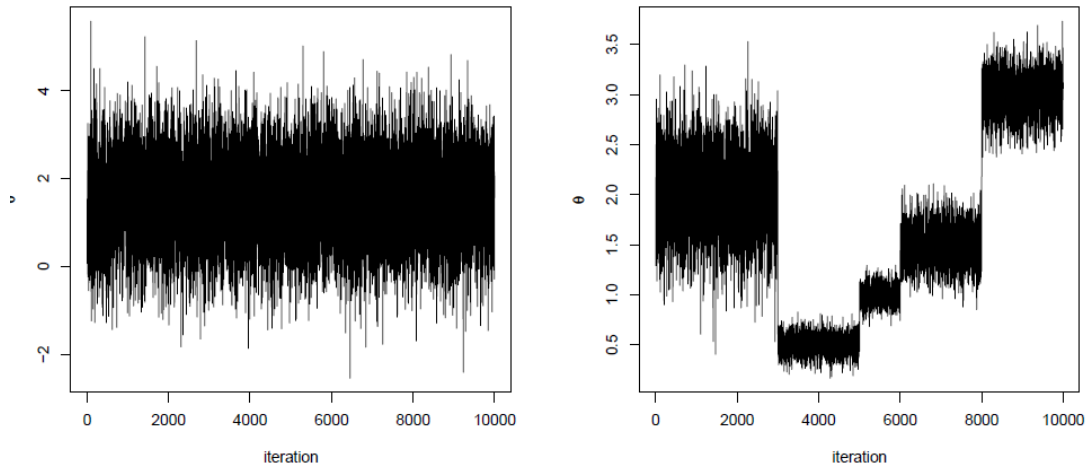


Figura 3.9. Mezclado de la cadena bueno (izq) y malo (der). (Fuente: Web)

Pero, ¿cuánto tiempo o número de iteraciones es necesario para que una muestra de la cadena de Markov se aproxime lo suficiente a una muestra de la distribución objetivo (la posterior)? Lo cierto es que no se puede decidir a priori cuántas iteraciones harán falta para alcanzar el equilibrio de la cadena, es decir, su periodo de calentamiento, a partir del cual la cadena muestrea la posterior.

### 3.4.3.- OBTENCIÓN DE UNA MUESTRA APROXIMADAMENTE ALEATORIA

Las muestras MCMC tomadas de forma consecutiva no son muestras aleatorias, debido a que se están generando mediante un proceso de Markov. Un muestreo de los de tipo directo (*ARS*, *SIR*, etc.) si que proporcionaría muestras aleatorias. Sin embargo, un método MCMC presenta grandes ventajas respecto al muestreo aleatorio (i.i.d.) en espacios paramétricos altamente dimensionales, ya que en estos espacios, las regiones con alta significancia estadística pueden estar muy dispersas (Saltelli et al. 2006), lo que provoca su difícil localización y exploración por métodos de tipo directo y no así mediante las cadenas de Markov.

Una escuela de pensamiento considera que, a pesar del carácter no aleatorio del muestreo Markoviano, se deberían emplear todas las extracciones

de la cadena, una vez transcurrido el periodo de calentamiento. Se puede demostrar que el resultado de esa forma de proceder no es tan preciso como realizar la inferencia a partir de una muestra puramente aleatoria de idéntico tamaño, tal y como se explica en el siguiente subapartado. Lo más correcto es por tanto, obtener una muestra de la posterior que resulte lo más parecida a una muestra aleatoria.

### 3.4.3.1.-PROCESOS ESTOCÁSTICOS LINEALES DE TIPO AUTORREGRESIVO

Son procesos cuyos valores, discretizados a intervalos  $\Delta t$ , son linealmente dependientes entre sí de la forma:

$$Z_i = \delta_1 \cdot Z_{i-1} + \delta_2 \cdot Z_{i-2} + \dots + \delta_p Z_{i-p} + \sigma_\epsilon^2 \cdot \xi \quad \text{Ec. 3.17}$$

$$\sigma_\epsilon^2 = \sigma_x^2 \left[ 1 - \sum_{j=1}^p \delta_j \times \rho(j) \right] \quad \text{Ec. 3.18}$$

Donde:

- $x_i$  es la serie estacionaria de valores aleatorios cuya varianza es  $\sigma_x^2$ .
- $z_i$  es la variable aleatoria resultado de tipificar  $x_i$ . Con Media  $E(z_i)=0$ .  
Con  $E(z_i \times z_{i-k})=\rho(k)$  función de autocorrelación o correlograma.
- $\delta_i$  son los parámetros autorregresivos.
- $p$  es el orden autorregresivo.
- $\xi$  es ruido blanco tipificado.
- $\sigma_\epsilon^2$  es la varianza del ruido blanco (Ec. 3.18).

Se demuestra en Bras and Rodríguez-Iturbe (1985) que la **función de autocorrelación** se puede expresar en función del desfase (*Lag*)  $k$  entre valores como:

$$\rho(k) = \delta_1 \cdot \rho(k-1) + \delta_2 \cdot \rho(k-2) + \dots + \delta_p \cdot \rho(k-p) \quad \text{Ec. 3.19}$$

Para el caso de un proceso en el que el estado actual se resume en el valor antecedente (**Proceso Markoviano**), la dependencia es autorregresiva de orden 1 (**AR1**). Los estados del proceso se obtienen según:

$$Z_i = \delta_1 \cdot Z_{i-1} + \sigma_\epsilon^2 \cdot \xi \quad \text{Ec. 3.20}$$

$$\sigma_\epsilon^2 = \sigma_x^2 (1 - \rho(1)^2) \quad \text{Ec. 3.21}$$

Y la función de autocorrelación de este tipo de procesos es exponencial, de forma que:

$$\rho(k) = \rho(1)^k \quad \text{Ec. 3.22}$$

Siendo  $\rho(0)=1$ ,  $\rho(1)=\delta_1$ ,  $k \geq 0$ .

#### 3.4.3.2.-TAMAÑO EFECTIVO DE UNA MUESTRA

Es conocido que al estimar la media de una variable aleatoria independiente (proceso de punto), la distribución muestral de la media tiene como esperanza de su primer momento a la media poblacional de la variable aleatoria subyacente y como esperanza del segundo momento a la varianza de la variable subyacente dividida por el tamaño de la muestra, es decir:  $E(\bar{x})=\mu$  y  $E(\bar{x}-\mu)^2=\sigma^2/N$ . Por tanto la incertidumbre con la que se estima la media de una variable aleatoria independiente es inversamente proporcional al tamaño de la muestra.

En el caso de una variable serial y positivamente dependiente, el aumento de esa incertidumbre va ligado además al grado de autocorrelación de la serie. Sea  $\rho(k)$  la función de autocorrelación de la serie, la incertidumbre con la que se estima la media se expresa según Bras and Rodríguez-Iturbe (1985):

$$E(\bar{x} - \mu)^2 = \frac{\sigma^2}{N} \cdot \left[ 1 + \frac{2}{N} \cdot \sum_{k=1}^{N-1} (N-k) \cdot \rho(k) \right] = \frac{\sigma^2}{N_e} \quad \text{Ec. 3.23}$$

Donde  $N_e < N$ , es el tamaño de la **Muestra Efectiva**, es decir, el tamaño de una muestra aleatoria que contendría la misma información que la serie dependiente.

Por tanto, la correlación entre los valores de la serie provoca que la información marginal que aporta cada nuevo valor dependiente es menor que la información que aportaría si ese valor fuese aleatorio sin estructura de dependencia. La consecuencia en el ámbito que abarca este trabajo es que, dada una precisión requerida para el cálculo de la  $\bar{x}$  posterior, **serán necesarias más simulaciones** ( $N_d$ ) cuanto mayor sea la autocorrelación de la cadena de Markov, o lo que es lo mismo, **cuanto peores sean sus propiedades de mezclado**.

#### 3.4.3.3.-EXTRACCIÓN DE UNA MUESTRA ALEATORIA A PARTIR DE LA CADENA DE MARKOV

Para obtener una muestra aproximadamente aleatoria, a partir de la cadena de Markov, se pueden plantear dos maneras de proceder.

1. Generar múltiples cadenas hasta el periodo de calentamiento. Consiste en simular una cadena de Markov hasta concluir el periodo de calentamiento. El siguiente elemento extraído de la cadena, ya será una muestra aleatoria de la posterior. Seguidamente se simulará de nuevo otra cadena de Markov hasta el periodo de calentamiento y se extraerá de nuevo el siguiente valor de la cadena, obteniendo así una segunda muestra aleatoria. El procedimiento se repetirá tantas veces como sea necesario para obtener el tamaño de muestra aleatoria requerido.
2. Generación de una cadena más allá del periodo de calentamiento. Con este procedimiento se debe realizar en una cadena, un número de simulaciones igual al periodo de calentamiento ( $N_b$ ) más las simulaciones necesarias ( $N_d$ )



para obtener el suficiente número de muestras dependientes que contengan la misma información que una muestra aleatoria independiente del tamaño requerido (la muestra efectiva, de tamaño  $N_e$ ). Seguidamente sobre los elementos de la cadena a partir del periodo de calentamiento, se realizará la operación de adelgazamiento (*Thinning*). Tras dicha operación la correlación entre dos muestras cualesquiera de la cadena, deberá resultar estadísticamente indistinguible de cero. Se hace necesario por tanto conocer el desfase o hueco que hay que dejar entre valores de la cadena, para que formen parte de una muestra aproximadamente aleatoria de la posterior.

#### 3.4.4.- TESTEO DE LA CADENA DE MARKOV

En el caso general, para obtener una muestra MCMC sobre la que poder realizar inferencia de la posterior, se hace imprescindible:

1. Estimar el periodo de calentamiento a partir del cual se considera que se ha alcanzado el régimen estacionario de la cadena, y por tanto la distribución de equilibrio.
2. Conocer las características de autocorrelación de la cadena, para poder adelgazarla y obtener así una muestra aleatoria.

Son diversos los métodos que nos ayudan a estimar el periodo de calentamiento, si bien ninguno de ellos es exacto, tampoco nos aseguran que se haya alcanzado el equilibrio y además necesitan de una simulación de testeo de la cadena, de cuyos resultados se podrán obtener los estadísticos apropiados que nos indicarán el número de simulaciones mínimas totales necesarias.

De igual modo habrá que estimar los estadísticos que nos proporcionen el *lag* para el cual ya no hay correlación significativa entre valores de la cadena.

Se puede encontrar un gran volumen de investigación desarrollada desde la última década del siglo pasado (Cowles and B. Carlin 1996; Gelman and Rubin



1992; Gelman et al. 2003; Gilks and Richardson 1995) sobre los diagnósticos que nos ayudan a determinar si la cadena puede haber alcanzado el equilibrio. Una colección de los mejores diagnósticos se ha recopilado en un conjunto de funciones del paquete estadístico de R, denominado CODA (Best et al. 1995). Destacan el Test de Geweke (1992), el Test de los Factores de Reducción de Gelman-Rubin (1992), el Test de Raftery-Lewis (1992) y los Test de carácter visual, como la inspección de la función de autocorrelación, o la visualización de la traza de valores de la cadena.

Hay que insistir en que ninguno de los métodos citados para determinar el periodo de calentamiento nos da garantías absolutas de que la cadena haya alcanzado el equilibrio. Solamente son útiles para tener evidencias si no se ha alcanzado dicho equilibrio. Cuando no aparecen esas evidencias se podrá afirmar que se ha alcanzado el equilibrio, si bien no será una conclusión fuerte.

### **3.5.- UN ALGORITMO MCMC ADAPTATIVO: DRAM**

#### **3.5.1.- LA ADAPTACIÓN COMO MEJORA AL ALGORITMO MH**

El algoritmo MH es el algoritmo de simulación MCMC más general, y de él han evolucionado diversas y distintas implementaciones para distintos propósitos.

En el algoritmo MH, la distribución “Candidata” o propuesta puede ser arbitraria, pero escoger una distribución lo más parecida a la distribución objetivo (la posterior de los parámetros) acelera considerablemente la convergencia hacia el muestreo de dicha distribución objetivo. Así mismo, cuanto más se parezcan ambas distribuciones, mejor propiedad de mezclado tendrá la cadena (puntos distantes de la distribución objetivo serán explorados) y por tanto mejor representará una secuencia corta de la cadena, a una muestra de la posterior.

Sin embargo, el modelador no conoce necesariamente las propiedades (matriz de covarianza) que debe tener la distribución propuesta para que se



parezca a la posterior de los parámetros. Este es el motivo del planteamiento de ir adaptando la distribución propuesta (cuya matriz de covarianzas inicial se ha de estimar, quizá de forma grosera) en base a los puntos del espacio de parámetros que ya han sido aceptados en los anteriores pasos de la cadena. De esta manera el algoritmo va “aprendiendo” qué forma tiene la posterior.

La propiedad de Ergodicidad que debe poseer la cadena de Markov, establece que la distribución empírica calculada a partir del muestreo de la propia cadena, es un estimador consistente de la distribución objetivo. Es decir, la Ergodicidad permite la convergencia de la cadena hacia una distribución de equilibrio coincidente con la distribución posterior de los parámetros.

Hay que tener en consideración un problema que puede surgir con los métodos adaptativos en general. Al adaptar la distribución propuesta empleando para ello, alguna parte de la información de la cadena generada hasta el momento, provoca que el muestreador pierda su carácter Markoviano y los resultados basados en la propiedad de la Ergodicidad ya no sean aplicables.

Una solución es usar la adaptación solamente durante un periodo inicial de calentamiento, en el que la cadena todavía converge hacia su distribución de equilibrio. De esta forma se asegura que se descartan de la muestra, esos valores de la cadena (Haario et al. 2006; Laine 2008).

### 3.5.2.- *DELAYED REJECTION ADAPTIVE METROPOLIS (DRAM)*

Son dos los procedimientos que incluidos en DRAM, mejoran notablemente la eficiencia del algoritmo básico MH: El proceso adaptativo (AM) y el método de rechazo retrasado (DR).

En **el proceso AM**, cada nuevo set de parámetros es generado a partir de una distribución propuesta del tipo Gaussiana Multidimensional, cuya matriz de covarianza se va actualizando de forma dinámica, a partir de los parámetros generados y aceptados hasta ese momento. La idea es que el algoritmo MCMC sea más eficiente en la medida en que la distribución propuesta, con la que se



van generando las muestras de parámetros, se va auto-ajustando a la forma y escala de la distribución posterior.

En el **método DR**, antes de proceder al rechazo de un punto no aceptado del espacio de parámetros (al que corresponderá un conjunto de parámetros), se aplica un escalamiento a la matriz de covarianzas de la distribución propuesta, con el fin de volver a intentar, en el mismo paso de la cadena, la aceptación de un nuevo punto generado con la nueva distribución propuesta escalada, preservando la reversibilidad de la cadena de Markov. Dicho escalamiento podrá ser aplicado sucesivas veces, aunque no suele mejorar los resultados aplicar DR en más de dos iteraciones seguidas (Haario et al. 2006; Laine 2008).

En conjunto, el algoritmo DRAM presenta un buen comportamiento incluso cuando la distribución posterior está lejos de ser Gaussiana y/o presenta una estructura de correlación no-lineal entre los parámetros (Laine 2008).

Se puede interpretar el proceso AM como un procedimiento de adaptación global a la distribución posterior, mientras que el método DR del procedimiento, se puede ver como una adaptación local a la distribución posterior, en el actual punto del espacio paramétrico en el que se ubica la cadena de Markov.

### 3.5.3.- IMPLEMENTACIÓN DE DRAM

El objetivo final de aplicar un método MCMC, es la aplicación de un análisis estadístico a un modelo no-lineal (Laine 2008). Dicho análisis estadístico consiste en evaluar la incertidumbre en los parámetros del modelo, así como la parte restante de incertidumbre total de las predicciones del modelo, que no es debida a los parámetros, la cual queda acumulada en el término de error.

Para evaluar la incertidumbre que los parámetros generan en las predicciones del modelo, se deberá propagar la incertidumbre de los parámetros a través del modelo, mediante un muestreo Monte Carlo de las posteriores de los parámetros.





Es posible que, en mayor o menor grado, la incertidumbre obtenida para los parámetros mediante un algoritmo MCMC, haya “absorbido” parte de la incertidumbre de cualquiera otra de las fuentes posibles de incertidumbre (estructural, de los datos, de las observaciones,...) (Laine 2008)

Es importante advertir, que la incertidumbre total asociada a cualquier predicción del modelo debe considerar ambas componentes de incertidumbre: La de los parámetros y la del término de los residuos.

Para explicar la implementación del algoritmo DRAM se establecerán las siguientes hipótesis:

1. Respecto al **modelo de Error**, se asume que las observaciones de la variable respuesta del modelo, tienen un error aditivo independiente, con distribución Gaussiana y con varianza desconocida y homocedástica  $\sigma^2$ . La modelación del término de error es una parte importante del proceso de modelización. El empleo del modelo Gaussiano resulta adecuado en muchas ocasiones, si bien su uso deberá estar justificado.
2. Para la inversa de la varianza del error del modelo ( $1/\sigma^2$ , **parámetro Nuisance** independiente del resto de parámetros  $\theta$ ), se considera el empleo una **distribución Prior del tipo Gamma**. Esta elección será útil para poder establecer con facilidad, la expresión (Ec. 3.27) de la distribución posterior de esta varianza, condicionada a las observaciones, dados los parámetros  $p(\sigma^{-2}|y, \theta)$ .

$$p(\sigma^{-2}) \sim \Gamma\left(\frac{n_0}{2}, \frac{n_0}{2} S_0^2\right)$$

Ec. 3.24

Donde,

- $S_0^2$  es la estimación Prior, de la media de la varianza Residual del modelo pre-calibrado, la cual se puede aproximar mediante el Error medio cuadrático (*MSE*) del modelo.

- $n_0=W*n$  es una medida del peso que se otorga a esa varianza residual Prior, siendo “n”, el nº de observaciones. Un valor alto de  $n_0$  mantendrá los valores muestreados desde la posterior de la varianza residual, próximos a su valor inicial.
3. Para el vector de **parámetros  $\theta$** , se pueden emplear **distribuciones Prior** Gaussianas independientes (no una prior multi-gaussiana) o de tipo no-informativo.
  4. Para el modelo de error asumido (Gaussiano aditivo homocedástico e independiente), **la función de verosimilitud de los parámetros  $\theta$** , para las “n” observaciones de la variable respuesta considerada, supuestas i.i.d., tiene la expresión:

$$p(y | \theta, \sigma^2) = (2\pi)^{-\frac{n}{2}} \sigma^{-n} \exp\left\{\frac{-1}{2\sigma^2} ss(\theta)\right\} \quad \text{Ec. 3.25}$$

Donde,  $SS(\theta)$  es la función de suma de cuadrados, definida como:

$$ss(\theta) = \sum_{i=1}^n (y_i - f(x_i; \theta))^2 \quad \text{Ec. 3.26}$$

Se debe considerar que para otro tipo de modelo de error, la función general a minimizar en la estimación de los parámetros del modelo, no sería la suma de cuadrados sino la función “menos dos veces la log-verosimilitud”. **Solamente para un modelo de ruido Gaussiano homocedástico no correlacionado**, resultan coincidir ambas optimizaciones, es decir, **minimizar  $SS(\theta)$  es equivalente a maximizar la verosimilitud** (o minimizar  $-2*\log [L(\theta)]$ ).

5. Si la Ec. 3.25 se considera como una función de la  $\sigma^2$ , se tiene la **función de verosimilitud de  $\sigma^2$** , la cual se puede demostrar por la teoría de las distribuciones conjugadas (Bolstad 2010) que se corresponde con una distribución Gamma para  $1/\sigma^2$ .

Puesto que la Prior para  $1/\sigma^2$  se escogió también una distribución de tipo Gamma, el producto de la verosimilitud por la Prior da como resultado igualmente una distribución de tipo Gamma, que será la distribución **Posterior de la inversa de la varianza residual del modelo**, cuya expresión es:

$$p(\sigma^{-2} | y, \theta) = \Gamma\left(\frac{n_0 + n}{2}, \frac{n_0 S_0^2 + SS(\theta)}{2}\right) \quad \text{Ec. 3.27}$$

5. Para las hipótesis expuestas, y considerando por ejemplo, una Prior de los parámetros Gaussiana (una distribución independiente para cada parámetro), la expresión de la **posterior de los parámetros no escalada**, será la siguiente:

$$p(\theta | y, \sigma^2) \propto \exp\left\{-\frac{1}{2}\left(\frac{ss(\theta)}{\sigma^2} + ss_{pri}(\theta)\right)\right\} \quad \text{Ec. 3.28}$$

6. Finalmente, se muestra la expresión que adopta la Probabilidad de Movimiento (o de aceptación),  $\alpha(\theta_1, \theta_2)$  necesaria para el algoritmo MH:

$$\frac{p(\theta_2 | y, \sigma^2)}{p(\theta_1 | y, \sigma^2)} = \exp\left\{-\frac{1}{2}\left(\frac{ss(\theta_2)}{\sigma^2} - \frac{ss(\theta_1)}{\sigma^2}\right) + \frac{1}{2}(ss_{pri}(\theta_2) - ss_{pri}(\theta_1))\right\} \quad \text{Ec. 3.29}$$



## 4.- APLICACIÓN A UNA CUENCA DE ALTA MONTAÑA: LA BAELLS, ALTO LLOBREGAT (BARCELONA)

### 4.1.- HERRAMIENTAS EMPLEADAS EN EL ANÁLISIS

#### 4.1.1.- EL MODELO HIDROLÓGICO TETIS

El modelo de simulación hidrológica TETIS desarrollado por el Departamento de Ingeniería Hidráulica y Medio Ambiente de la Universitat Politècnica de València pertenece a la clase de modelos distribuidos conceptuales y su objetivo es obtener de la mejor forma posible la respuesta hidrológica ocasionada por la precipitación de lluvia o de nieve, teniendo en cuenta los diferentes procesos físicos involucrados.

Es un modelo complejo que representa la cuenca como una malla de celdas interconectadas por la configuración topográfica, derivada de un Modelo de Elevación Digital, de manera que cada celda tiene asociado un valor para cada parámetro necesario para una correcta caracterización.

El modelo incorpora adecuadamente la variabilidad espacial de los ciclos hidrológicos y su base conceptual, por eso necesita de partida datos observados para poder permitir su calibración, cumple con el principio de parsimonia, esto es, ante funcionamientos similares, se selecciona el modelo con menos número de parámetros y con ecuaciones lo más simples posibles, representando todos los procesos significativos (Bussi 2010).

A continuación se realiza una descripción más detallada del modelo TETIS, presentando su estructura, los procesos internos, parámetros del modelo y las condiciones iniciales.

#### 4.1.1.1.-FORMULACIÓN DEL MODELO

La representación de la producción de escorrentía se hace mediante esquemas conceptuales simples adaptados a la escala de celda y al intervalo de tiempo que se considera en la modelación.

En TETIS, la producción de la escorrentía se basa en la realización de un balance hídrico en cada celda, asumiendo que el agua se distribuye en seis niveles o tanques de almacenamiento conceptuales y conectados entre sí (más un séptimo de cauces cuando lo haya en la celda), como puede observarse en la *Figura 4.1*.

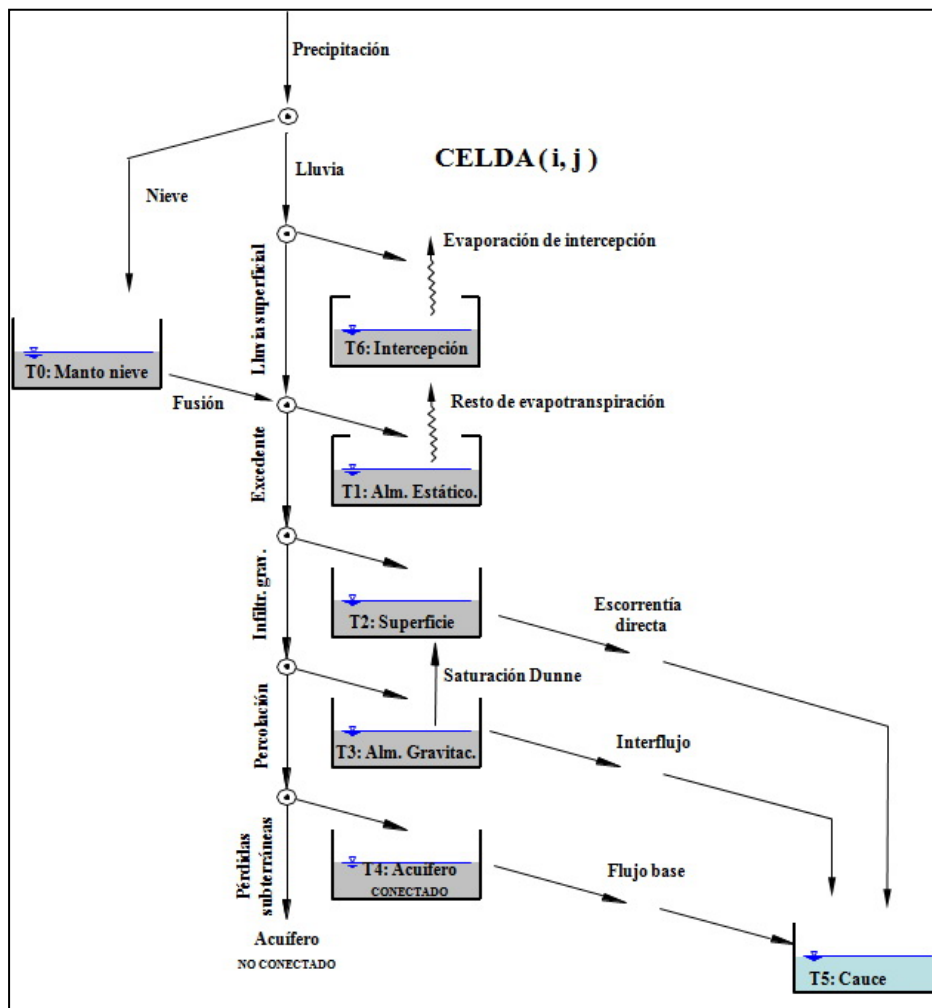


Figura 4.1 Esquema conceptual de tanques a nivel de celda del modelo TETIS

El flujo entre los tanques es función del agua almacenada en ellos, por lo que las **variables de estado** son los volúmenes almacenados en cada uno de



los tanques. La función que relaciona el flujo con estas variables de estado es función del esquema conceptual adoptado, del tipo de tanque y de las características morfológicas de la celda e hidrológicas del suelo en la misma.

Los flujos verticales de agua entre cada tanque representan los procesos hidrológicos de: lluvia ( $X_1$ ) y nieve ( $X_0$ ), correspondientes a la precipitación total, lluvia directa ( $X_6$ ), excedente ( $X_2$ ), infiltración ( $X_3$ ), exfiltración ( $Z_3$ ), percolación ( $X_4$ ) y pérdidas subterráneas ( $X_5$ ). Los horizontales representan respectivamente: escorrentía directa ( $Y_2$ ), interflujo ( $Y_3$ ) y flujo base ( $Y_4$ ). También existen otros dos flujos de salida: evapotranspiración desde el conjunto suelo-vegetación ( $Y_1$ ) y evaporación de agua interceptada ( $Y_6$ ).

El tanque  $T_0$  corresponde al almacenamiento de altura equivalente de agua (SWE), producto de las nevadas. La precipitación es considerada como nieve y es añadida a  $T_0$ , si la temperatura del aire es inferior a una temperatura crítica ( $T_c$ ). La nieve y la temperatura son interpoladas en cada celda, utilizando el método IDW. Para el proceso de fusión, TETIS utiliza el método índice de temperatura o grado-día (Martinec 1960), básicamente porque no se dispone de información completa de radiación solar neta, energía sensible, energía latente, calor del suelo y energía advectiva. El caudal proveniente de la fusión de nieve se divide entre el tanque de almacenamiento estático y el excedente.

El  $T_1$  representa el agua que transita por la cuenca y que sólo sale de ella por evapotranspiración ( $Y_1$ ), por lo tanto no hace parte de la escorrentía. Este almacenamiento tiene en cuenta la detención del agua en charcos y el agua que retiene el suelo por las fuerzas capilares. El almacenamiento por intercepción ocurre en  $T_6$ , cuya salida es la evaporación de intercepción ( $Y_6$ ). Se supone que el  $T_1$  tiene una capacidad máxima igual a la suma de la capacidad de almacenamiento capilar en el suelo y el almacenamiento en depresiones en la superficie del suelo, en función de la rugosidad y la pendiente. Por su parte en el  $T_6$  la capacidad depende de la cobertura de la superficie para almacenar agua, esta última relacionada con el tipo de vegetación.



El  $T_2$  representa el **agua en la superficie** y es susceptible de moverse superficialmente por la **ladera** (escorrentía directa) o de infiltrarse al nivel inferior. La cantidad de agua que se deriva para escorrentía superficial depende de la capacidad del suelo para dejar pasar el agua gravitacional y del estado del almacenamiento en la capa superior del suelo. La capacidad para dejar pasar el agua gravitacional se puede asociar a la conductividad hidráulica en la capa superior del suelo (en condiciones de saturación), conductividad en la que se tiene en cuenta tanto la macroporosidad como la microporosidad. Finalmente la cantidad de agua que sigue por el conducto distribuidor  $X_3$ , se relaciona con  $K_s$ , capacidad del suelo para dejar pasar el agua a su interior (conductividad hidráulica saturada del suelo asociada a la cobertura), con el paso temporal de modelación y con el flujo excedente del almacenamiento estático  $X_2$ .

El  $T_3$  se puede entender como el **agua gravitacional** almacenada en el suelo. Parte de este volumen pasa al nivel inferior por percolación, otra parte produce el flujo subsuperficial o Interflujo, y si se activa en el modelo el mecanismo de Dunne, se puede producir un flujo de exfiltración o saturación desde abajo (desde  $T_3$  hacia arriba), que contribuye directamente a la escorrentía directa. El **Interflujo** a través de la capa superior del suelo, se desarrolla inicialmente sobre una delgada capa que fluye lateralmente hacia abajo en la ladera y se va concentrando en pequeñas depresiones, grietas o pequeños conductos por el interior de esta capa, hasta que sale a los elementos de la red de drenaje. Se supone igualmente que la capa inferior del suelo a la escala de la celda tiene una capacidad de percolación representativa o característica y que se asocia al tipo de subsuelo y su estructura, lo cual puede estar estrechamente relacionado con las características geológicas (litológicas y estructurales) y geomorfológicas de las capas inferiores del suelo. La cantidad de agua que puede continuar por el canal  $X_4$ , hacia la zona inferior del suelo durante el intervalo de tiempo determinado es función del excedente del almacenamiento del flujo superficial en ladera  $X_3$  y de la conductividad hidráulica en la capa inferior del suelo (subsuelo) en condiciones de saturación que se conoce como capacidad de percolación  $K_p$ .

El  $T_4$  representa al **acuífero conectado** al punto de control de la cuenca. Las salidas corresponden a las pérdidas subterráneas (para el punto de control) y al flujo base de los cauces. Para la representación del flujo a través del almacenamiento subterráneo, en cada una de las celdas se utiliza la ecuación de continuidad y una ecuación que relaciona la tasa de flujo que sale de este almacenamiento con la cantidad de agua almacenada, del mismo modo que en los almacenamientos anteriores. El volumen de agua que durante el intervalo de tiempo ingresa por percolación  $X_4$  tiene la posibilidad de que una cantidad de agua siga hacia las pérdidas subterráneas  $X_5$  o que se derive hacia el almacenamiento subterráneo  $T_4$ .

Por último,  $T_5$  es el tanque que representa al almacenamiento en las celdas cauce (o en su caso, cárcava), en las cuales se suman los flujos de escorrentía directa, interflujo y flujo base, formando en su conjunto la escorrentía superficial.

La conceptualización del modelo TETIS es una malla interconectada en tres dimensiones según la Figura 4.2. Los tres tanques inferiores drenan hacia el correspondiente tanque aguas abajo, siguiendo las direcciones del flujo propuestas por el MED, hasta alcanzar la red de drenaje constituida por cárcavas y cauces.

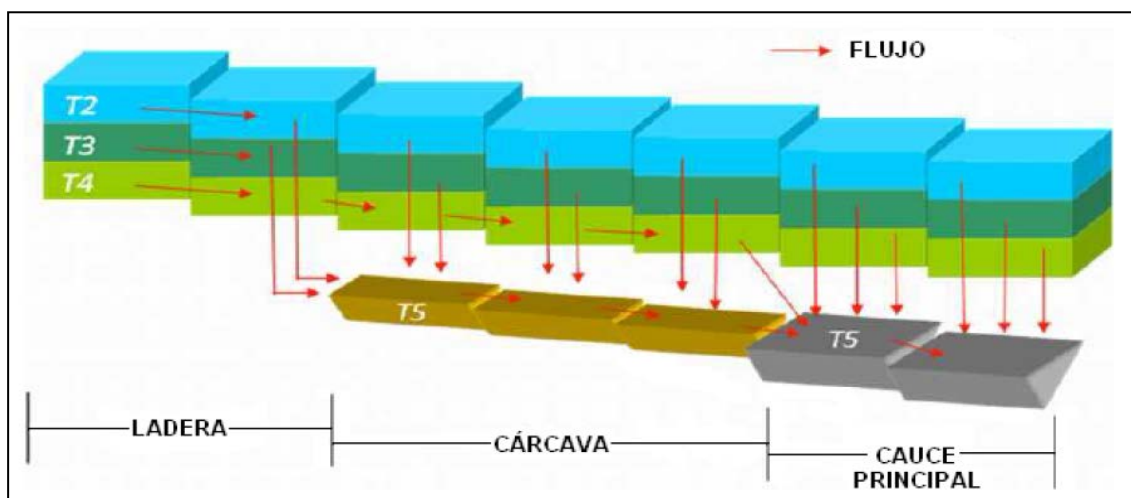


Figura 4.2. Movimiento horizontal propuesto por el modelo TETIS

(Fuente: Manual de Tetis)



Las salidas posibles de agua para cada celda se da por los tres componentes principales de la escorrentía: la escorrentía directa (producida por escorrentía hortoniana y escorrentía por saturación), el interflujo y el flujo base. Cada proceso involucrado se simula con ecuaciones simples y con pocos parámetros para estimar (I. Vélez 2001).

El sistema presenta tres elementos diferenciados, definidos por las **áreas umbrales** para que las diferentes componentes de la escorrentía salgan a la superficie: las laderas, las cárcavas y los cauces. Además, estas áreas coinciden con la presencia de una incisión permanente en el terreno en donde se concentran los flujos, ya sea en cárcavas o en cauces. En la *Figura 4.3* se observa un esquema de la relación entre las áreas umbrales para las diferentes componentes de la escorrentía y el área de captación.

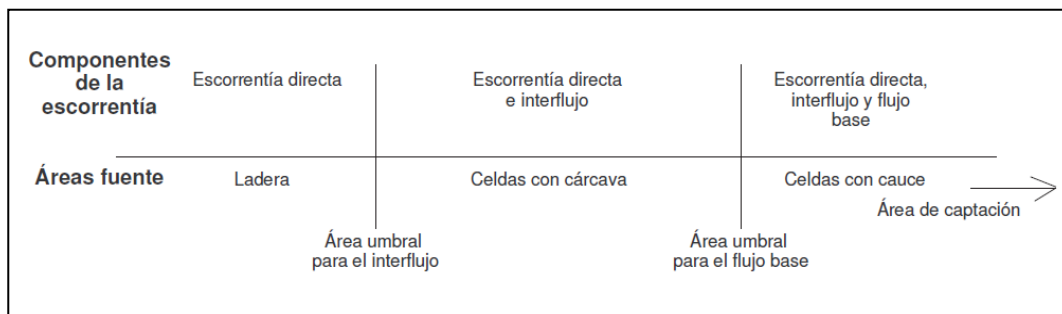


Figura 4.3. Elementos del sistema, áreas umbrales y componentes de la escorrentía (Fuente: Montoya 2008)

La formulación hidráulica para la **propagación en la red de cauces** utilizada en el modelo TETIS se basa en una aproximación de la onda cinemática, asumiendo un lecho fijo en cada celda en el cual se aplica una ecuación de balance. Dicha formulación se denomina Onda Cinemática Geomorfológica. La forma de la sección de flujo en cada uno de los elementos que conforman la red de drenaje puede obtenerse a partir de relaciones de geometría hidráulica previa validación de sus parámetros. Las relaciones han sido propuestas por (Leopold and Maddock 1953), que relacionan las dimensiones de la sección de flujo: profundidad, ancho, velocidad y caudal por medio de ecuaciones de tipo potencial. Las ecuaciones son las siguientes:

- Área de captación  $\Lambda$  y caudal a sección llena  $Q_b$ :

$$\Lambda = k \cdot Q_b^\phi \quad \text{Ec. 4.1}$$

- Ancho de la sección transversal a sección llena  $W_b$  y caudal a sección llena  $Q_b$ :

$$W_b = a_1 \cdot Q_b^{\alpha_1} \quad \text{Ec. 4.2}$$

- Ancho de la sección transversal  $W$  y caudal que pasa por la sección  $Q$ :

$$W = Q^{\alpha_2} \quad \text{Ec. 4.3}$$

- Diámetro del sedimento  $d$ , pendiente  $S$  y profundidad del flujo  $h$ :

$$d = c_d \cdot (Q \cdot h)^\theta \quad \text{Ec. 4.4}$$

- Coeficiente de rugosidad  $n$  y diámetro del sedimento  $d$ :

$$n = c_n \cdot d^\xi \quad \text{Ec. 4.5}$$

A partir de un número pequeños de secciones transversales medidas en campo usando regresiones lineales se pueden estimar los coeficientes y exponentes de las ecuaciones anteriores. En la Tabla 4.1 se muestran los intervalos de variación habituales en cauces naturales según literatura.

Tabla 4.1. Intervalos de variación de los coeficientes y exponentes geomorfológicos

(Fuente: Francés et al. 2007)

Parámetro	Intervalo de variación
k	0.5 – 0.75
$\phi$	0.65 – 0.8
$a_1$	0.5 – 5.75
$\alpha_1$	0.34 – 0.55
$\alpha_2$	0.05 – 0.2
$c_d$	0.5 – 50
$\theta$	0.5 – 2.75
$c_n$	0.025 – 0.07
$\xi$	0.125 – 0.18



#### 4.1.1.2.-PARÁMETROS DEL MODELO

Los parámetros de un modelo describen las características del sistema a modelar, en función de la estructura y conceptualización del mismo. Dado que el modelo se conceptualiza sobre una malla regular de celdas de un Modelo de Elevación Digital, es necesario caracterizar cada celda de la malla con un valor homogéneo del parámetro.

El enfoque más sencillo es asumir que un valor medido en un punto al interior de la celda es válido para toda su superficie. En la realidad, un área del tamaño de una celda (que puede variar desde algunos metros hasta un kilómetro) se comporta, desde el punto de vista hidrológico, de manera muy distinta a un punto; por ejemplo, en el caso de un parámetro de conductividad del suelo, la medición de laboratorio se efectúa sobre un bloque de suelo del tamaño de algunos centímetros, y el valor medido no suele ser representativo de la celda entera (Grayson and Blöschl 2001).

Otra posibilidad es ajustar el parámetro para que los resultados del modelo se acerquen a los valores observados de la variable output (es decir, calibrar el valor del parámetro), el parámetro pierde su significado físico, y representa algo similar a un valor medio en la celda; este parámetro se llama **parámetro efectivo**. El enfoque del parámetro efectivo intenta superar los problemas conceptuales asociados al valor puntual de un parámetro. Los parámetros efectivos son valores que reproducen el comportamiento de un área finita o de un volumen finito. No pueden entonces ser medidos puntualmente, y no necesitan estar relacionados con mediciones puntuales (Bussi 2010).

El modelo TETIS utiliza una estructura de parámetros efectivos separada en dos partes (Francés et al. 2007).

1. Las características físicas del medio natural, estimadas según la información disponible y medida a una escala espacial definida.
2. Los factores correctores.

Las características físicas expresan el valor medido o estimado del parámetro para cada celda de la cuenca, describen la variabilidad espacial del

mismo y poseen significado físico. Los factores correctores absorben los errores provenientes de todas las fuentes citadas antes. Es razonable suponer que el factor corrector sea común para todas las zonas de la cuenca, o al menos para un número limitado de regiones dentro de la cuenca. Además, siendo todas las celdas del mismo tamaño, los efectos de escala son iguales para toda la cuenca. La ecuación que describe la estructura separada del parámetro efectivo es la siguiente:

$$\theta_{i,j}^* \approx R_i \theta_{i,j} \quad j=1,\dots,n ; i=1,\dots,p \quad \text{Ec. 4.6}$$

donde  $\theta_{i,j}^*$  es el parámetro efectivo  $i$ -ésimo para la celda  $j$ ,  $R_i$  es el factor corrector del parámetro  $i$ , común para todas las celdas,  $\theta_{i,j}$  es el parámetro estimado para la celda  $j$ ,  $n$  es el número total de celdas y  $p$  es el número total de parámetros por celda.

La ventaja principal de esta estructura del parámetro es que, en la fase de calibración, se reduce notablemente el número de variables que deben ser ajustadas, siendo sólo necesario calibrar los  $p$  factores correctores, en vez de  $np$  valores (número de parámetros por número de celdas).

De este modo, el modelo hidrológico TETIS podrá ser calibrado a través del ajuste de 14 factores correctores según la Tabla 4.2. Son 10 factores correctores que influyen en los procesos de producción y propagación de la escorrentía (incluyendo el mecanismo de Dunne), un coeficiente, denominado  $\beta$ , para introducir la variabilidad espacial de la precipitación incorporando el efecto de la altitud y 3 factores del módulo de nieve.

Tabla 4.2. Factores correctores y respectivos parámetros ajustados.

<b>Factor corrector</b>	<b>Parámetro</b>
<i>F1</i>	Capacidad de almacenamiento hídrico
<i>F2</i>	Índice de cobertura de vegetación
<i>F3</i>	Capacidad de infiltración
<i>F4</i>	Velocidad del flujo superficial
<i>F5</i>	Capacidad de percolación

Factor corrector	Parámetro
<i>F6</i>	Velocidad del interflujo
<i>F7</i>	Capacidad de percolación profunda
<i>F8</i>	Velocidad del flujo base
<i>F9</i>	Velocidad del flujo en canal
<i>FN2</i>	Coeficiente de fusión
<i>FN3</i>	Coeficiente de fusión con lluvia
<i>FN4</i>	Temperatura base de fusión
<i>FP1</i>	Factor de interpolación de la lluvia con la cota, proporcional a la cantidad de precipitación observada en la estación (aproximación no lineal)
<i>FD2</i>	Capacidad máxima al tanque T3 expresada como un porcentaje de $H_u$ , máxima capacidad del tanque estático (mecanismo de Dunne de saturación del suelo desde abajo)

#### 4.1.1.3.-CONDICIONES INICIALES

Para simular un evento es necesario definir las condiciones iniciales de las variables de estado. En el modelo TETIS, las variables de estado son las siguientes:

1. El nivel de agua equivalente en el manto de nieve, tanque *H0*.
2. El nivel de almacenamiento por intercepción, tanque *H6*.
3. El nivel del almacenamiento estático, tanque *H1*.
4. El nivel del almacenamiento superficial, tanque *H2*.
5. El nivel del almacenamiento gravitatorio, tanque *H3*.
6. El nivel del almacenamiento subterráneo, tanque *H4*.
7. El caudal en los canales, tanque *H5*.

El nivel del almacenamiento estático presenta una capacidad máxima, definida por la capacidad de almacenamiento hídrico del suelo  $H_u$ . El modelo necesita como dato de entrada la condición inicial del tanque de almacenamiento estático representada por el porcentaje del total de la capacidad de almacenamiento hídrico del suelo.



Con respecto a los niveles del almacenamiento superficial, gravitatorio y subterráneo, el valor inicial de entrada se da como la altura de agua en milímetros de cada tanque. Para el caudal en los canales, se define el valor inicial como el porcentaje del caudal a sección llena.

Una posibilidad para considerar niveles distribuidos en el espacio (un valor para cada celda) es definir los valores iniciales de humedad como los producidos por una simulación anterior, por ejemplo de un periodo de calentamiento. El programa permite utilizar valores de humedad inicial definidos por el usuario o valores obtenidos de una simulación anterior. En el presente trabajo se ha utilizado un **periodo de calentamiento de 1 año** para establecer las condiciones iniciales en la cuenca.

#### 4.1.2.- EL PAQUETE ESTADÍSTICO FME

A diferencia de los métodos de estimación de parámetros (calibración), que proporcionan un conjunto de parámetros “óptimo” en el sentido mínimo cuadrático de mejor ajuste a las observaciones, los métodos bayesianos pueden ser empleados para obtener las PDF de los parámetros, condicionados por los datos (Soetaert and Petzoldt 2010, 2012; Soetaert 2010a; b; c). Es decir, emplean la información de los residuos (valores observados menos valores simulados) y también presentan la posibilidad de introducir la información que a priori se pueda tener sobre los parámetros.

**FME, Flexible Modeling Environment** (Soetaert and Petzoldt 2012; Wu and Liu 2012) es un paquete de R que proporciona funciones que ayudan a realizar análisis de identificabilidad, sensibilidad e incertidumbre de modelos. Contiene herramientas de calibración y herramientas que permiten la aplicación de métodos de Monte Carlo.

Si bien fue creado para ser empleado con modelos basados en ecuaciones diferenciales ordinarias, ecuaciones diferenciales en derivadas parciales o ecuaciones diferenciales algebraicas, podría trabajar con cualquier tipo de



modelos. Entre otras, las funciones de que dispone FME y que van a emplearse en el presente trabajo son:

- **modCost:** Dada una solución del modelo para cada variable simulada y dadas las observaciones de dichas variables, estima los residuos del modelo, así como el “coste” o la suma de cuadrados residual (SSR), por cada variable simulada y para el total del modelo.
- **modMCMC:** Herramienta de análisis bayesiano, que simula una cadena de Markov. FME implementa el algoritmo MCMC denominado DRAM (Haario et al. 2006). Esta herramienta se empleará como para el análisis de incertidumbre sobre los parámetros del modelo. Su resultado son las distribuciones posteriores de probabilidad de los parámetros.
- **sensRange:** Función que produce envolventes alrededor de las variables sobre las que se ha efectuado el análisis de sensibilidad de los parámetros. Con esta herramienta se puede realizar un análisis de incertidumbre (sensibilidad global) sobre una variable respuesta del modelo (con observaciones disponibles), tomando como datos de partida una muestra de las distribuciones posteriores de los parámetros.

FME también contiene otras funciones (Soetaert and Petzoldt 2012; Soetaert 2010b), como algoritmos para la calibración de modelos, o herramientas para la generación de valores de los parámetros por muestreo, según un diseño basado en grid, en distribuciones multi-normal o uniforme, o mediante muestreo tipo Hipercubo Latino.

## 4.2.- IMPLEMENTACIÓN DEL MODELO HIDROLÓGICO

En este capítulo se presenta una descripción de las principales características físicas de la cuenca en estudio, La Baells (Llobregat), específicamente las que determinan su comportamiento hidrológico y geomorfológico. A partir de esta caracterización se estimarán los parámetros necesarios para el funcionamiento del modelo.



#### 4.2.1.- CUENCA HIDROLÓGICA

La cuenca hidrográfica del Llobregat, en la que se ubica la subcuenca de La Baells, objeto de estudio de esta tesina, se encuentra situada al Noreste de la Península Ibérica, ubicada en la región de Catalunya y constituye el eje central de la provincia de Barcelona, exceptuando la cabecera del río Cardener que se encuentra en la provincia de Lleida.

El nacimiento del río Llobregat se encuentra en el extremo Norte de la provincia de Barcelona, en el municipio de Castellar de N'Hug, en un fondo rocoso a 1295 m de altitud, este nacimiento tiene su origen en un río subterráneo (manantiales) formado por rocas calcáreas permeables y el lecho del río reposa sobre rocas impermeables. Entre estas dos capas sale al exterior y hace su recorrido hasta desembocar en el mar Mediterráneo. No existe ningún estudio hidrogeológico que permita deducir el origen de estos manantiales, pero dada su situación y características es lógico suponer que debe tratarse de una surgencia natural. El río tiene una longitud de 156.5 Km y una extensión total de la cuenca de 4948 Km<sup>2</sup> que se reparte total o parcialmente en comarcas de características bien diferenciadas como son el Berguedà, el Bages y el Baix Llobregat.

En el curso del propio río Llobregat y, en el curso de sus afluentes, destacan las comarcas de Osona, el Solsonès, el Anoia, el Alt Penedès, la Segarra y el Vallès Occidental (Atlas de Catalunya, 1993).

En la *Figura 4.4* se representa gráficamente el perfil longitudinal del río Llobregat que es el resultado de los diferentes materiales que atraviesa y de la acción erosiva de sus aguas en el tiempo.



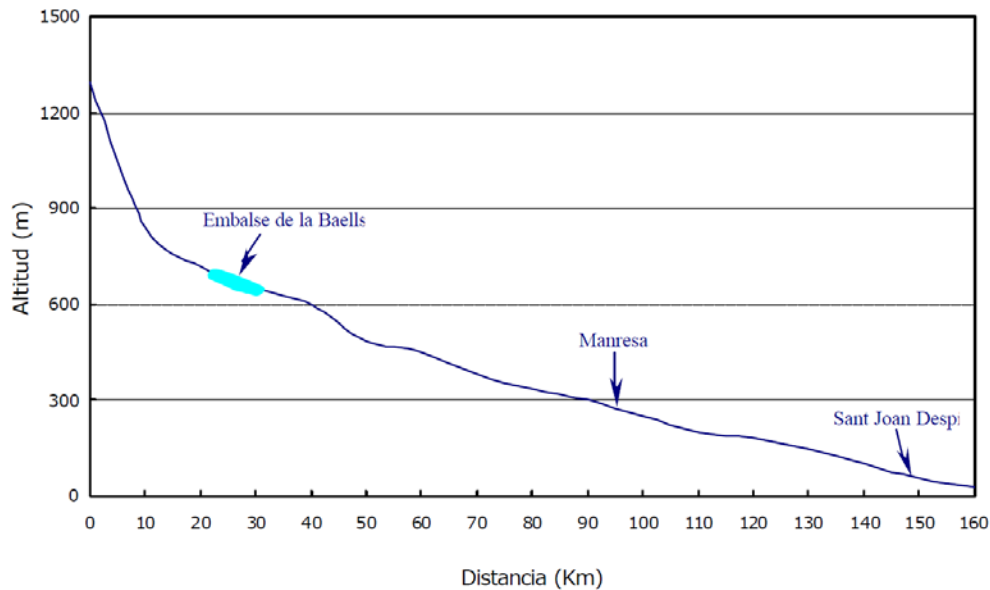


Figura 4.4. Perfil Longitudinal del Río Llobregat (Fuente: Web)

En su primer tramo de recorrido, dentro de la subcuenca La Baells, el río Llobregat salva fuertes desniveles, pasa a través de rocas calcáreas, formando un congosto y abriendo un estrecho y profundo valle hasta llegar a La Pobla de Lillet (860 m), primera población importante de su curso. Allí recibe por su margen izquierdo su primer afluente, el río Arija, el cual proviene de la sierra de Mogrony (2045 m). En ésta zona el río forma un ángulo y, por terrenos triásicos, inicia su único tramo de recorrido longitudinal siguiendo el valle de acuerdo con la estructura geológica dominante, sierras y plegamientos longitudinales Este-Oeste.

Llega de este modo a Guardiola de Berguedà (700 m) donde confluye con el río Bastareny, que discurre por el valle de Bagà, según la misma orientación (en este caso en sentido inverso Oeste-Este). A partir de ésta zona, el río Llobregat recobra de nuevo su característica orientación transversal Norte-Sud, que no abandonará hasta el inicio de su curso bajo al penetrar en las alienaciones paleozoicas y triásicas del sistema Mediterráneo en cuyo momento se inclinará ligeramente para proseguir en dirección Noroeste-Sudeste hasta su desembocadura.

Unos 2 Km aguas abajo de Guardiola de Berguedà recibe también por su margen derecho otro afluente, el río Saldes, procedente del Pedraforca, y después de otros 8 Km de recorrido, al pasar por la población de Cercs, situada sobre un altozano, a 650 m de altitud, recibe por su margen derecho, el río Peguera, de menor recorrido (9.6 Km) que los anteriores pero de caudal muy constante como todos ellos. El río Llobregat prosigue su curso y por su margen izquierdo recibe la riera de Margansol junto a cuya confluencia está el Embalse de La Baells que tiene una capacidad de  $115 \text{ hm}^3$ . Hasta el embalse que da nombre a la subcuenca de estudio, el área drenada es de  $504 \text{ km}^2$ .

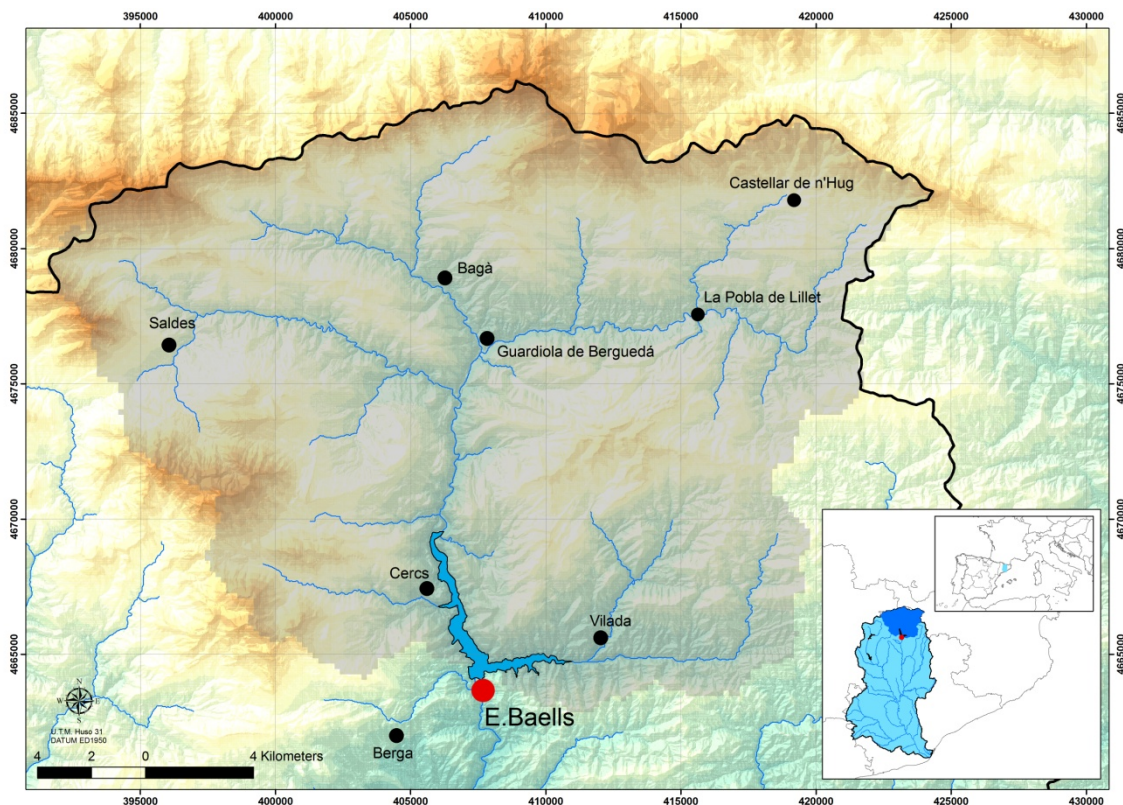


Figura 4.5. Localización de la cuenca de La Baells.

#### 4.2.2.- CLIMA

En la cuenca de La Baells el clima es de alta montaña-mediterráneo con las principales estaciones de lluvias en primavera y otoño, mientras que el invierno es la temporada con menor precipitación. Las nevadas ocurren pocas veces en las alturas, pero las bajas precipitaciones durante el invierno y las temperaturas



suaves provocan un papel hidrológico moderado del manto de nieve (Gallart et al. 2011).

La vegetación nativa en la mayoría de las altitudes medias es caducifolia, (*Quercus pubescens*), bosques con arbustos de hoja perenne de *Buxus sempervirens* y bosques de *Quercus ilex* en las zonas soleadas o rocosas.

En las regiones más altas o de más frío, las cubiertas dominantes son *Pinus sylvestris* y los bosques de *Fagus sylvatica*, aunque la mayor cubierta es de *Pinus uncinata* con algunos arbustos *Juniperus sp.* Por encima de 2300 msnm, la cubierta predominante es de pradera de tierras altas, favorecida por la actividad de pastoreo que redujo la altitud de la línea de árboles.

Como valores característicos anuales, la cuenca de La Baells presenta una temperatura media de 10.9 °C con un coeficiente de variación (CV) de 0.06. La precipitación media es de 905 mm con CV de 22.7. La escorrentía media es de 416 mm con CV 41.9 (Gallart et al. 2011)

#### 4.2.3.- PRECIPITACIÓN Y TEMPERATURA

*Spain02* (Herrera 2011; Herrera et al. 2012), proyecto desarrollado dentro del Grupo de Meteorología de Santander<sup>1</sup>, conformado por investigadores y profesores de la Universidad de Cantabria, el Instituto de Física de Cantabria y la Agencia Estatal de Meteorología, es una **base de datos diaria de precipitación y temperaturas** máximas y mínimas sobre una malla regular de alta resolución (0.2°x0.2°) para el territorio peninsular español y las Islas Baleares, interpolada a partir de un número elevado (miles) de estaciones que han sido sometidas a un estricto control de calidad. Dicha base datos es de libre acceso y está disponible online<sup>2</sup>, a distribución pública para ser aprovechada por los científicos que se encuentran al frente de las diferentes investigaciones relacionadas con el impacto que generan los cambios en la

<sup>1</sup> <http://www.meteo.unican.es/en/main>

<sup>2</sup> <http://www.meteo.unican.es/es/datasets/spain02>



precipitación y las temperaturas. Esta rejilla ha sido validada frente a las observaciones y ha sido utilizada para la validación de modelos globales y regionales del clima.

La base de datos observados para el desarrollo de Spain02 es la red secundaria de la Agencia Estatal de Meteorología (AEMET). Los primeros registros de esta red de observaciones datan del 1 de septiembre de 1859 y abarcan hasta nuestros días. El periodo considerado en este proyecto fue desde 1950 a 2003, ya que se encuentra contenido entre los dos periodos utilizados usualmente para calcular promedios climáticos (1961-1900 y 1971-2000).

La base de datos interpolada ha sufrido modificaciones y actualizaciones, además de incorporar nuevos métodos en su desarrollo. Las versiones van desde Spain02\_v1 donde el periodo es desde 1950 a 2003 y únicamente se dispone de precipitación, hasta la más reciente versión Spain02\_v3 donde se describe una nueva rejilla de datos de precipitación interpolados a partir de una red de unas 500 estaciones con al menos 40 años con el 90% de los datos en el periodo 1950-2011, apropiada para el estudio de tendencias y considerando la orografía como covariable.

Para el desarrollo de este trabajo se ha empleado la versión de *Spain02*, con fecha de actualización de 15 de Marzo de 2012.

#### **4.2.4.- ESTIMACIÓN DE PARÁMETROS ESPACIALES**

Además de la información hidrometeorológica, el modelo hidrológico distribuido TETIS requiere una estructura de parámetros que describan la variabilidad espacial de las características de los suelos, del sustrato y de la cubierta vegetal a escala de cuenca, representada mediante la estimación a priori de mapas de parámetros a partir de la información medioambiental disponible, la cual debe ser coherente con la topografía, el uso del suelo presente, la cubierta vegetal existente, la edafología, la litología y las demás características del medio físico susceptibles de representarse espacialmente.

La coherencia en la representación espacial está ligada con la correcta calibración del modelo, ya que en ésta se debe garantizar que el modelo refleje la realidad del medio sobre el que se aplica.

Para una descripción detallada del proceso de estimación de parámetros se puede consultar (Ramirez et al. 2012)

#### **4.2.4.1.-INFORMACIÓN DISPONIBLE DE LAS CARACTERÍSTICAS FÍSICAS DE LA CUENCA**

Para la creación de los mapas que representan las características físicas de cada una de las cuencas, se ha utilizado información disponible que tuviera representación cartográfica y que fuera relevante para la determinación de las variaciones espaciales en las áreas de estudio (Ramirez et al. 2012).

La Tabla 4.3 muestra los diferentes tipos de información cartográfica original, utilizada en función del tema tratado en el estudio.

*Tabla 4.3. Información cartográfica utilizada para la estimación de parámetros*

Fuente: (Ramirez et al. 2012)

<b>Mapas</b>	<b>Formato Original y Escala</b>	<b>Parámetros estimados</b>
Modelo de Elevación Digital	Digital, píxel de 25 x 25 m	Mapa de pendiente, Mapa de direcciones de flujo, Mapa de celdas acumuladas y Mapa de velocidad de flujo en la ladera.
CORINE <i>land cover</i> 2006	Digital, 1:100.000	Mapa de almacenamiento estático del suelo, Mapa de intercepción, Mapa de almacenamiento superficial por depresiones en el terreno.
Base de datos de Suelo Europea (ESDB)	Digital, 1:1.000.000	Mapa de almacenamiento estático del suelo y de conductividades hidráulicas horizontal y vertical del suelo.
Mapa litoestratigráfico y de Permeabilidades de España.	Digital, 1:200.000	Mapa de permeabilidad horizontal y vertical del acuífero y mapa de velocidad de recarga del acuífero.



#### 4.2.4.2.-MODELO DE ELEVACIÓN DIGITAL (MED)

La distribución espacial de algunos parámetros requeridos en el modelo TETIS, se pueden obtener mediante operaciones cartográficas, a partir de información topográfica disponible. El mapa base para la descripción de las características físicas de la cuenca y de los fenómenos de la naturaleza que en ella se originan es el Modelo de Elevación Digital (MED), de cuya calidad depende la fiabilidad de la topología de la red de drenaje, deducida de algunos mapas derivados del mismo. Para efectos de la implementación del modelo hidrológico distribuido TETIS, se ha tomado como mapa base el Modelo Digital del Terreno con tamaño de celda de 25 m adquirido a través de la página web del Instituto Geográfico Nacional. Los MED originales se han agregado a una **resolución espacial de 200 x 200 m** de tamaño de celda. Los respectivos tamaños de celdas se consideran adecuados para el área de las cuencas a modelar y es acorde a los tiempos de cálculo necesarios en el proceso de calibración automática de los parámetros del modelo. Todos los demás mapas de parámetros utilizados por el modelo tienen la misma resolución espacial.

### 4.3.- APLICACIÓN PRÁCTICA

#### 4.3.1.- PLANTEAMIENTO DEL PROBLEMA

La evaluación de la incertidumbre de los parámetros así como de la incertidumbre de la predicción en modelos hidrológicos, debe ser una parte esencial de cualquier estudio hidrológico. No es suficiente la estimación del valor “óptimo” de los parámetros, como tampoco lo es la presentación de los resultados que el modelo genera para una variable, sin asociar a dichos resultados el grado de fiabilidad que tienen dichos resultados.

El análisis de incertidumbre, tal como se ha planteado, está basado en un modelo de regresión aditivo y no lineal (Ec. 2.1) según el cual, las observaciones se representan mediante la suma de un componente determinístico (el modelo hidrológico) y un componente aleatorio el cual describe los errores o residuos, entendidos como la diferencia entre lo



observado y lo calculado por el modelo. Esa diferencia es debida a errores de medición y a todo lo que el modelo no contempla sobre los procesos subyacentes del sistema modelado. Los residuos están conformados por una combinación de los errores generados por las posibles fuentes, como son los datos de entrada (datos *forcing*), las observaciones, los parámetros y la estructura del modelo. Esta aproximación agregada a la modelación de los residuos, no permite dilucidar cuál es la contribución de cada fuente al error total, aunque sí permite estimar la incertidumbre de los parámetros así como la incertidumbre total de la predicción.

Así, según el modelo de error, la inferencia sobre los parámetros del modelo está conceptualmente basada en una función de verosimilitud mediante la cual se cuantifica la probabilidad de que las observaciones sobre la variable respuesta del modelo sean reproducidas por un particular conjunto o vector de parámetros del modelo. El mapeo desde el espacio de parámetros al espacio de verosimilitudes permite la identificación de una serie de conjuntos de parámetros aceptables dadas las observaciones. El tradicional proceso de estimación de parámetros o calibración, persigue el objetivo de encontrar, de entre todos los vectores plausibles, el vector de parámetros máximo-verosímil. Sin embargo, la consideración de todos los vectores de parámetros verosímiles permite la estimación de la incertidumbre paramétrica, así como su efecto sobre la incertidumbre predictiva del modelo.

En consecuencia con lo anterior, una aproximación formal a la estimación de la incertidumbre de los parámetros y a la incertidumbre de la predicción requiere de la asunción a priori de un modelo estadístico para el término aleatorio o residual del modelo de regresión mencionado. Es decir, se debe establecer a priori una forma funcional de la distribución de probabilidad conjunta de los residuos. El modelo estadístico de los residuos es la base para derivar de él la función de verosimilitud apropiada para la inferencia sobre los parámetros. La ventaja de la aproximación formal es que las hipótesis sobre el modelo de error quedan establecidas de forma explícita y su validez puede (y



debe) ser verificada a posteriori, tras la realización de la inferencia sobre los parámetros.

Como ejemplo, en el presente trabajo se asumirá un modelo estadístico de los residuos en el que los errores son (i.i.d.), independientes (sin estructura de correlación temporal) e idénticamente distribuidos según una distribución Normal con media cero y varianza estacionaria (homocedástica). Ese modelo estadístico conduce a una función de log-verosimilitud proporcional a la función cuadrática de los residuos del modelo (Ec. 3.25). O lo que es lo mismo, la inferencia sobre los parámetros realizada con este modelo de error, se puede llevar a cabo mediante la conocida y generalmente empleada técnica de minimizar los cuadrados de los residuos. El incumplimiento de las hipótesis establecidas para el modelo de error puede introducir sesgo en la estimación de los parámetros y afectar a la estimación de su incertidumbre. Igualmente se verá afectada la estimación de la incertidumbre predictiva. Por tanto se deberá investigar a posteriori, cuales son las características estadísticas de los residuos, verificando las 3 hipótesis: Normalidad, homocedasticidad e independencia.

#### **4.3.2.- OBJETIVOS DE LA APLICACIÓN PRÁCTICA**

Establecido el marco general en el que se desenvuelve el trabajo, a continuación se enuncian de forma esquemática los hitos que se van a ir desarrollando en los siguientes apartados, cuyos resultados van a contribuir a la consecución de los dos objetivos principales: Evaluar la incertidumbre de los parámetros condicionada a las observaciones y al modelo de error establecido, así como estimar el efecto de dicha incertidumbre sobre la incertidumbre predictiva. Como objetivo adicional se realizará una estimación de la incertidumbre predictiva total, según el modelo de error establecido a priori y se comprobará la validez de dicho modelo error, para el caso analizado.

Los hitos que se van a ir cumpliendo en los siguientes apartados son:





1. Estimar los rangos de los distintos parámetros así como su influencia sobre la respuesta del modelo mediante la realización de un **análisis local de sensibilidad** a partir de un punto del espacio paramétrico obtenido mediante una calibración manual preliminar del modelo.
2. Evaluar la **incertidumbre de los parámetros** mediante la obtención de su **distribución posterior** condicionada a las observaciones así como a los datos de entrada denominados *forcing*. Debido a que en el problema planteado, ni la verosimilitud (Ec.3.25) ni por tanto la posterior (Ec.3.28) pueden ser evaluadas directamente por medios analíticos o numéricos, se necesita recurrir a métodos como el muestreo por Monte Carlo para obtener una muestra de la posterior. El muestreo de la posterior se efectúa mediante la ejecución del algoritmo MCMC llamado DRAM, implementado en el paquete FME. Así mismo en este proceso se muestrea, aunque de forma analítica, basándose en la teoría de distribuciones conjugadas, la **distribución posterior de la varianza de los residuos** (Ec.3.27), de manera que a cada muestra obtenida de la posterior de los parámetros, le corresponda una muestra de la posterior de la varianza del modelo de error.

Además, para la correcta aplicación de los métodos MCMC se deben realizar unas comprobaciones y operaciones sobre la cadena de valores obtenida, las cuales permitirán considerarla como muestra de la distribución posterior de los parámetros, a saber:

- Comprobación de estacionariedad de la cadena
  - Comprobación de aleatoriedad de los valores de la cadena
  - Obtención en su caso, de una muestra aprox. aleatoria mediante adelgazamiento de la cadena obtenida con MCMC.
3. Evaluar el **efecto de la incertidumbre de los parámetros sobre la incertidumbre predictiva**.
  4. Comparación para cada parámetro entre la magnitud relativa de su incertidumbre y la magnitud relativa de la incertidumbre que genera sobre la

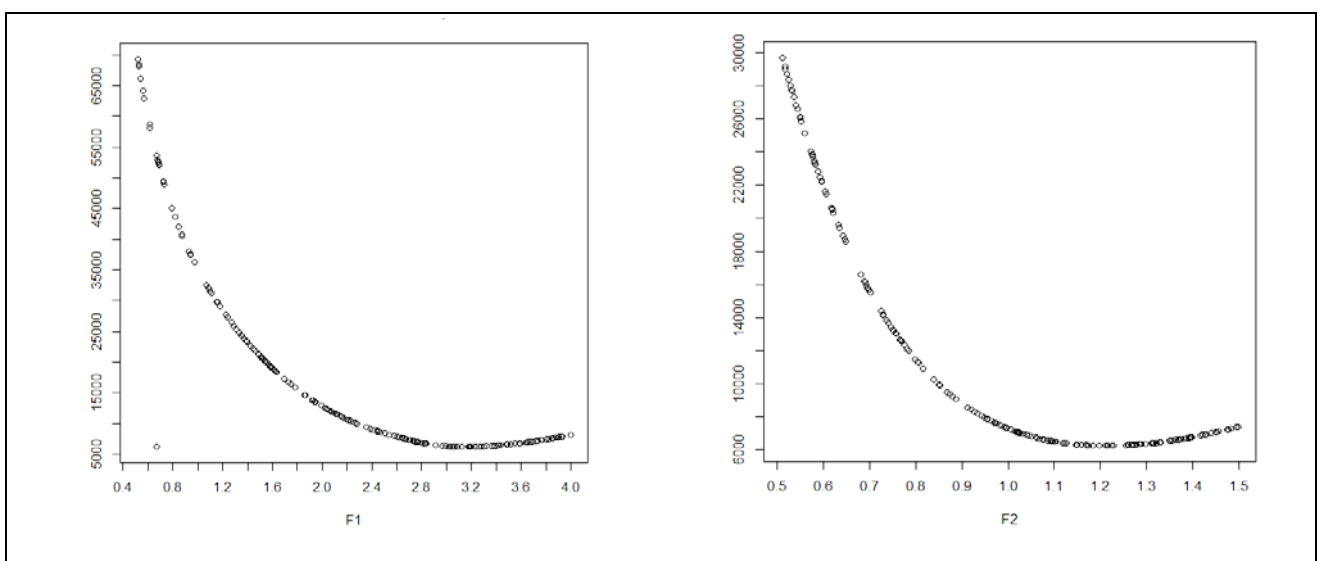
variable respuesta. Relación con el concepto de **Sensibilidad Global** del modelo a cada parámetro.

5. Estimar la **incertidumbre predictiva total** según el modelo de error preestablecido.
6. Comprobación del modelo de error establecido a priori.

#### 4.3.3.- ANÁLISIS DE SENSIBILIDAD LOCAL

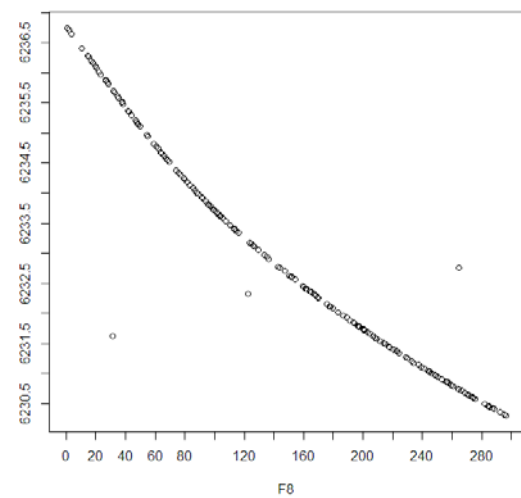
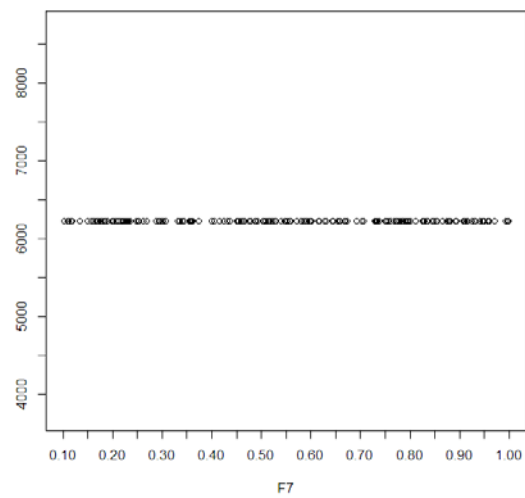
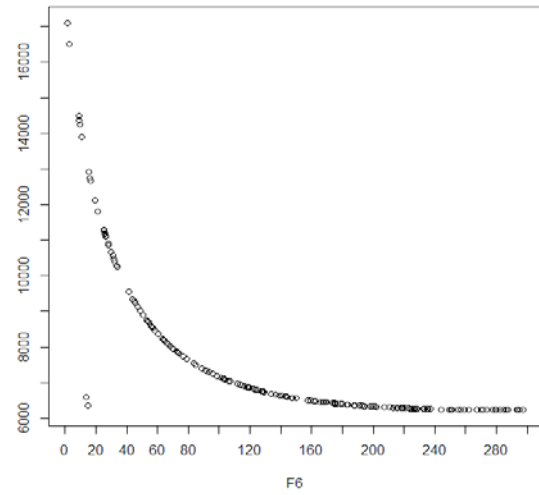
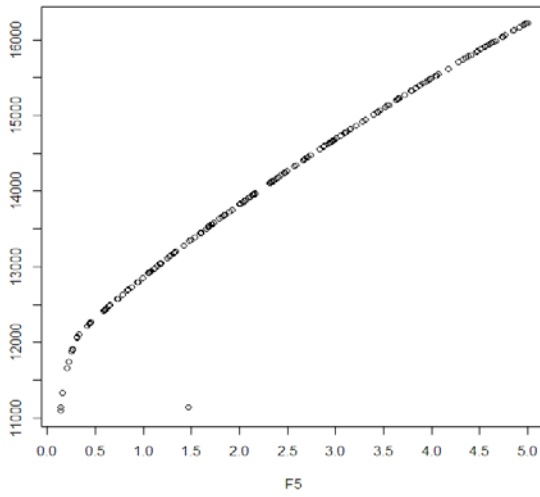
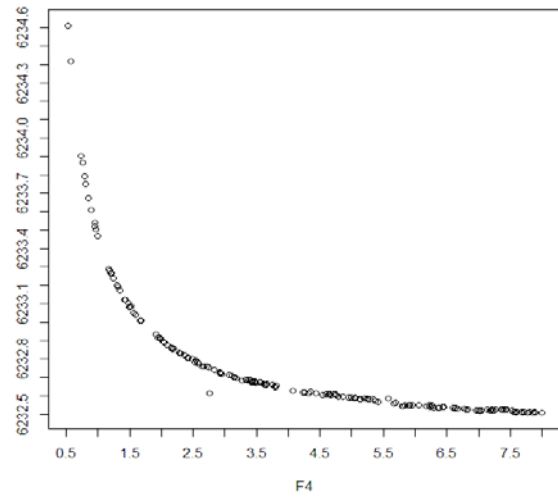
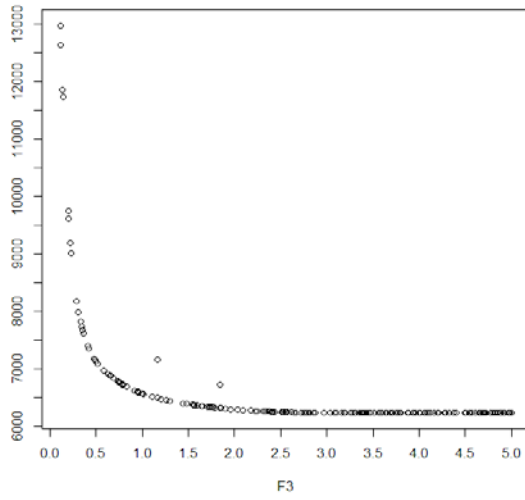
Una vez confeccionado el modelo hidrológico, se aborda en primer lugar una calibración manual que nos permite conocer en qué zona del espacio de los parámetros va a quedar ubicado el modelo.

A continuación se realiza un muestreo Monte Carlo directo sobre el espacio de parámetros, dentro del rango razonable en el que a priori se supone que se van a mover sus valores. Este muestreo se realiza partiendo del punto de calibración manual, manteniendo fijos en ese punto todos los parámetros excepto el que se está muestreando. El resultado se aprecia en las siguientes figuras para cada uno de los 14 parámetros del modelo, en las cuales se representa en ordenadas la suma de los residuos cuadráticos del modelo (SSR) y en abscisas el valor del parámetro.



Evaluación de la Incertidumbre de los Parámetros y su efecto sobre la Incertidumbre Predictiva de un Modelo Hidrológico Distribuido, mediante un algoritmo *Markov Chain Monte Carlo* de tipo Adaptativo.

Máster Universitario en Ingeniería Hidráulica y Medio Ambiente



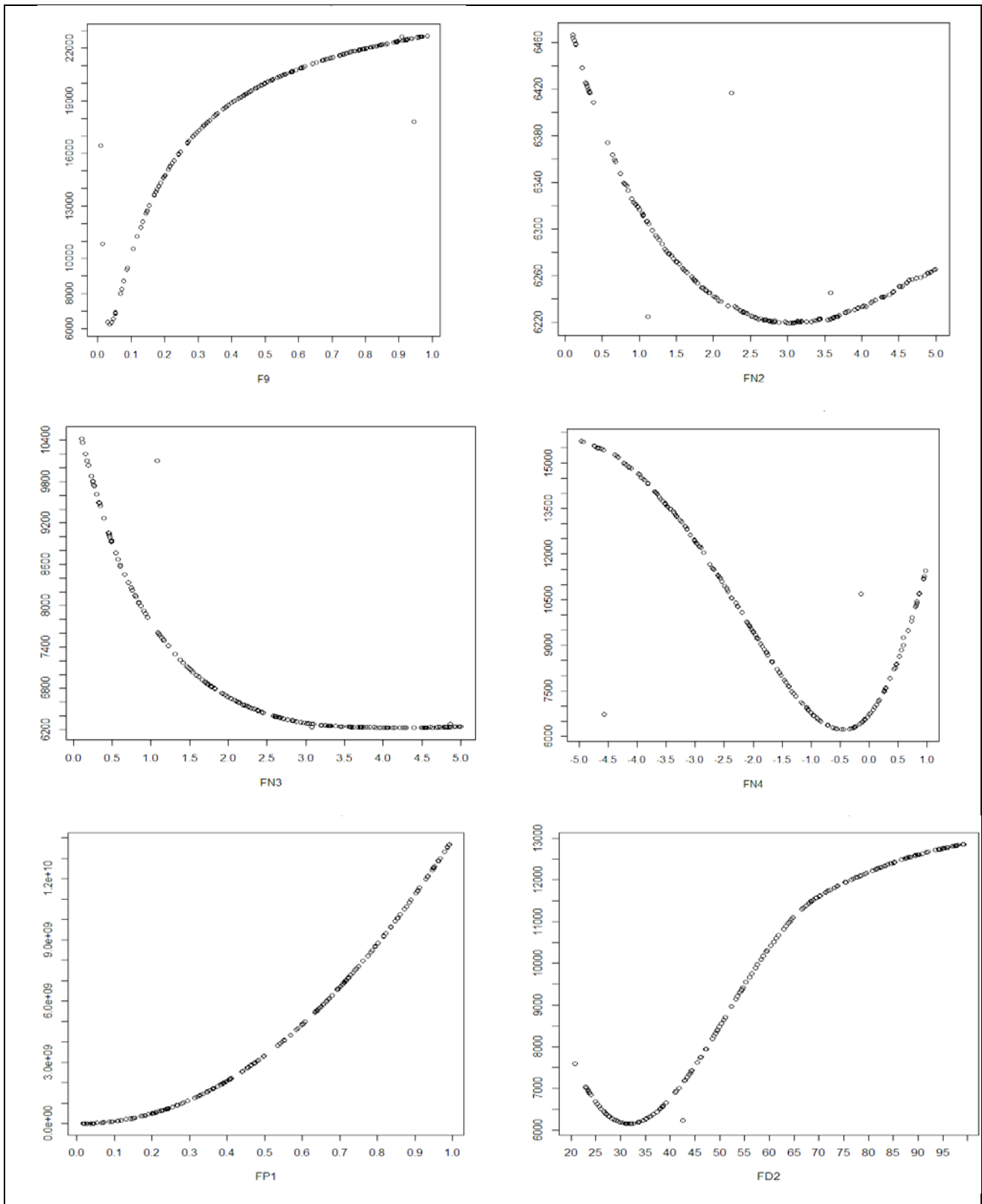


Figura 4.6: Seguimiento de los parámetros. Sensibilidad Local por Monte Carlo



A la vista de las gráficas y considerando el significado conceptual de cada parámetro, se pueden agrupar los 14 parámetros en tres grupos.

1. Parámetros que presentan un mínimo local en la SSR. Es importante advertir que este mínimo está condicionado al punto del espacio de parámetros desde el que se está realizando del análisis local. Los siguientes parámetros pertenecen a este grupo: F1, F2, F9, FN2, FN4, FP1 y FD2.
2. Parámetros que no presentan mínimo local aunque si tienen un valor umbral a partir del cual la SSR adopta un valor meseta mínimo. Estos parámetros pueden presentar falta de identificabilidad ya que a partir de su respectivo umbral, cualquier valor que adopten puede ser igual de válido desde el punto de vista del óptimo de la función objetivo. Son de este grupo, los parámetros de infiltración (F3), respuesta rápida del flujo (F4), respuesta intermedia (F6) y parámetro de fusión con lluvia (FN3).
3. Parámetros que tienden asintóticamente al valor mínimo de la SSR. Por tanto el valor óptimo del parámetro es aquel en el que se ubica la asíntota. El parámetro relacionado con la capacidad de percolación (F5) pertenece a este grupo, siendo su valor óptimo cero.
4. Parámetros que no presentan una variación significativa de la SSR en todo el rango de valores muestreado. El modelo resulta insensible a estos parámetros por lo que pueden ser eliminados de este. Pertenecen a este grupo el parámetro de percolación profunda (F7), y la respuesta lenta del flujo (F8).

La conclusión principal que se extrae del análisis de sensibilidad local, es la posibilidad de prescindir de ciertos parámetros en el análisis de incertidumbre que se va a realizar. Los parámetros que se pueden obviar en dicho análisis son tres: F5, F7 y F8. Sin ellos la cuenca se comportará sin respuesta lenta y sin pérdidas de recurso en el punto de control (embalse de La Baells).

#### 4.3.4.- INCERTIDUMBRE DE LOS PARÁMETROS

El análisis de incertidumbre se va a realizar sobre 11 parámetros. En la siguiente tabla se muestra la nomenclatura de parámetros que el software FME asignará a los resultados de los análisis y su correspondencia con la nomenclatura de TETIS.

Tabla 4.4 Equivalencia en la nomenclatura de los parámetros

P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11
F1	F2	F3	F4	F6	F9	FN2	FN3	FN4	FP1	FD2
ALM.ESTÁTICO	ET	INFILTRACION	ESCORRENTÍA	INTERFLUJO	VEL.CAUCE	FUSION	FUSION LLUVIA	TEMP. FUSION	PRECIPIT. ALTURA	COEF. DUNNE

##### 4.3.4.1.- CALIBRACIÓN Y SIMULACIÓN DE LA CADENA DE MARKOV

Para comenzar la simulación del algoritmo MCMC es necesario proporcionar una matriz inicial de covarianzas para la distribución candidata o propuesta de los parámetros, que en el caso del algoritmo empleado DRAM, es una distribución multi-gaussiana. Sin embargo a priori no se conoce dicha matriz por lo que se asume que sea una matriz diagonal (hipótesis de independencia entre los parámetros). Las varianzas de la diagonal se estiman de forma grosera por conocimiento experto. Finalmente la matriz de covarianzas se escala convenientemente según se indica en Laine (2008) por  $2.4^2/npar$ , siendo  $npar$  la dimensión del vector de parámetros.

Al estar empleando un algoritmo MCMC de tipo adaptativo, la matriz de covarianzas supuesta diagonal se irá modificando de forma dinámica en función de los valores de la cadena que van siendo aceptados, durante un periodo de calentamiento, con el que la cadena alcanza su distribución de equilibrio y en el cual ninguna de las muestras de parámetros extraídas del muestreo será guardada como parte de la distribución posterior. Tras un número de simulaciones que en este análisis ascendió a la cifra de 25.000, se da por finalizado el proceso de calentamiento y adaptación. Las simulaciones de la cadena que se realizan desde ese punto, ya son consideradas parte del muestreo de la distribución posterior de los parámetros. En este análisis se realizaron 5.000 simulaciones para el muestreo de la posterior de los



parámetros, así como de la distribución posterior de la varianza residual. Los resultados de esta simulación MCMC fueron los siguientes.

```
R> summary(MC)

Iterations = 1:5000
Thinning interval = 1
Number of chains = 1
Sample size per chain = 5000

1. Empirical mean and standard deviation for each variable,
   plus standard error of the mean:

      Mean      SD Naive SE Time-series SE
p1  4.197014 1.134e-01 1.604e-03      8.503e-03
p2  1.839625 5.035e-02 7.120e-04      4.044e-03
p3  6.923788 6.122e-01 8.657e-03      4.369e-02
p4  5.033751 1.069e+00 1.512e-02      7.435e-02
p5 241.179784 2.072e+01 2.931e-01      1.852e+00
p6  0.048093 1.791e-03 2.533e-05      1.357e-04
p7  3.223441 6.211e-01 8.783e-03      5.889e-02
p8  7.081434 4.761e-01 6.733e-03      4.176e-02
p9 -0.984668 4.469e-02 6.320e-04      3.093e-03
p10 0.004353 6.596e-05 9.329e-07      4.876e-06
p11 43.845504 1.717e+00 2.429e-02      1.485e-01

2. Quantiles for each variable:

      2.5%      25%      50%      75%      97.5%
p1  3.983237  4.117656  4.192587  4.267327  4.429281
p2  1.735466  1.808873  1.842349  1.874543  1.934949
p3  5.693817  6.523597  6.954430  7.415422  7.894978
p4  3.158247  4.202924  4.956089  5.792617  7.150535
p5 204.975272 225.821932 239.998664 255.217815 283.540004
p6  0.044449  0.046866  0.048160  0.049365  0.051416
p7  1.987604  2.808479  3.236030  3.665951  4.406630
p8  6.116830  6.750596  7.106272  7.441466  7.881216
p9 -1.072917 -1.013790 -0.983127 -0.954271 -0.895950
p10 0.004231  0.004308  0.004353  0.004399  0.004482
p11 40.389690 42.842877 43.840115 45.002904 46.942114

3. Varianza del Ruido

mean 7.4843404
sd 0.3803322
min 6.1878854
max 9.0260970
q025 7.2271093
q050 7.4835378
q075 7.7402213
```

Antes de considerar que estos resultados ya corresponden a la distribución posterior de los parámetros, se deben realizar los correspondientes diagnósticos sobre la cadena, que permiten verificar la estacionariedad de la misma así como la aleatoriedad de la muestra extraída.

No obstante a la vista de los resultados mostrados, ya se puede extraer alguna conclusión. En el apartado primero de los resultados se pueden observar dos columnas en las que se calcula la desviación típica (error estándar) de la media estimada para cada uno de los 11 parámetros. La primera columna estima ese error sin considerar ningún tipo de dependencia en



la serie de valores de la cadena, mientras que en la segunda columna se realiza la estimación del error considerando el efecto de la autocorrelación en la serie (Ec. 3.23). Comparando ambas columnas se observa que el error en la estimación en la media es mayor (hasta en un orden de magnitud) si se considera el efecto de posibles autocorrelaciones. Por tanto se concluye que la cadena tal cual se ha extraído del algoritmo MCMC no es una muestra aleatoria, por lo que se deberá someter a un proceso (*Thinning*) que la aproxime a ese tipo de muestras.

#### **4.3.4.2.-DIAGNÓSTICOS SOBRE LA SIMULACIÓN MCMC**

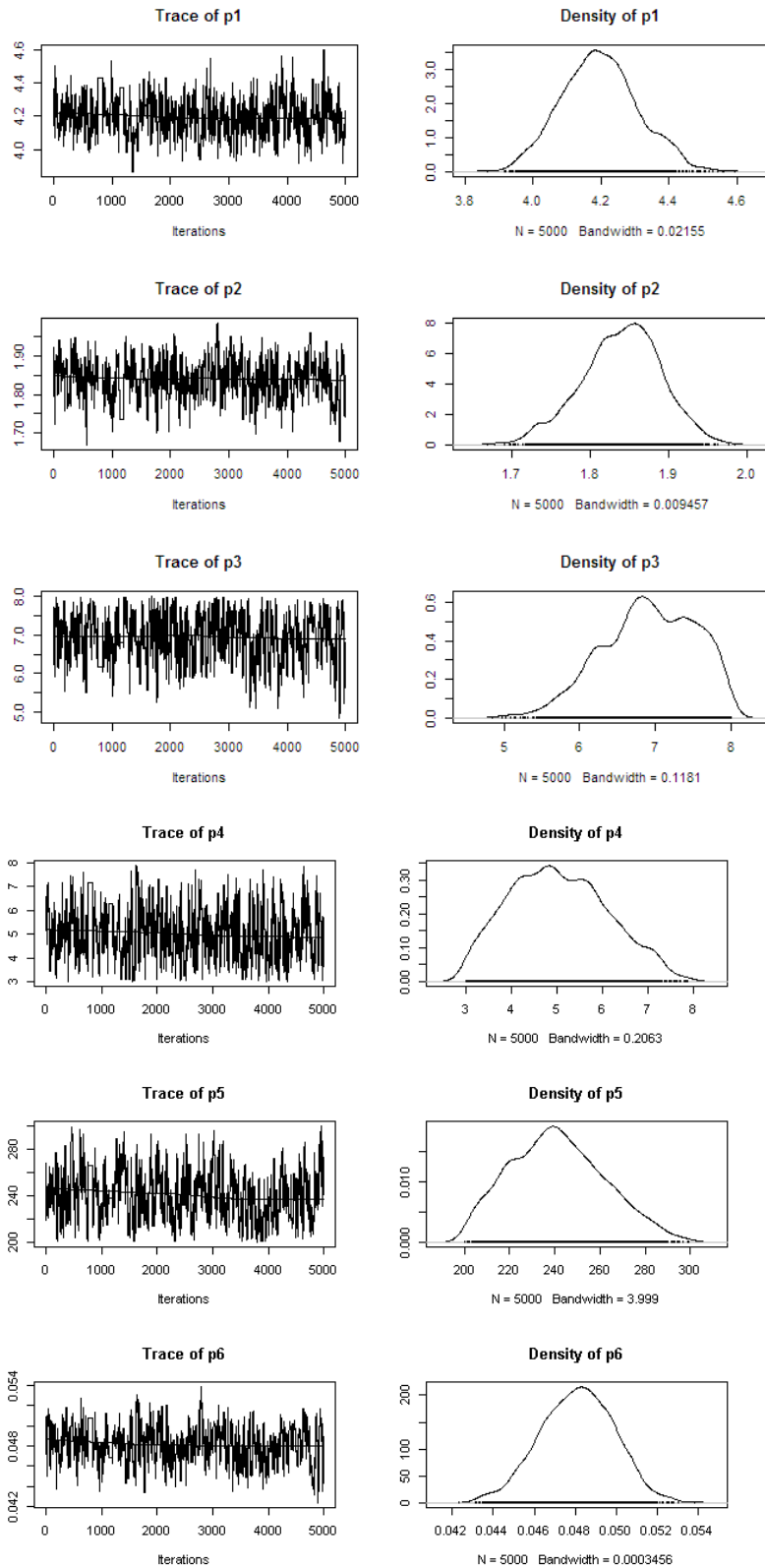
##### **4.3.4.2.1.-ESTACIONARIEDAD**

Con el objetivo de comprobar que la cadena de Markov de cada parámetro alcanzó el equilibrio se cuenta, entre otras, con las siguientes herramientas:

1. **Inspección visual** de la evolución de valores de la cadena. Es necesario que la cadena presente una tendencia media estabilizada y es deseable que presente unas buenas propiedades de mezclado (ver apartado 3.4.2). Si la cadena presenta un alto grado de autocorrelación no tendrá buenas propiedades de mezclado. En la siguiente figura se observa la evolución de la cadena de Markov para cada parámetro así como la distribución marginal posterior resultante.

La inspección visual parece indicar que las cadenas presentan una tendencia media estabilizada y un mezclado aceptable aunque no óptimo, lo cual es esperable si se tiene en cuenta que las cadenas presentan autocorrelación como se ha visto anteriormente.





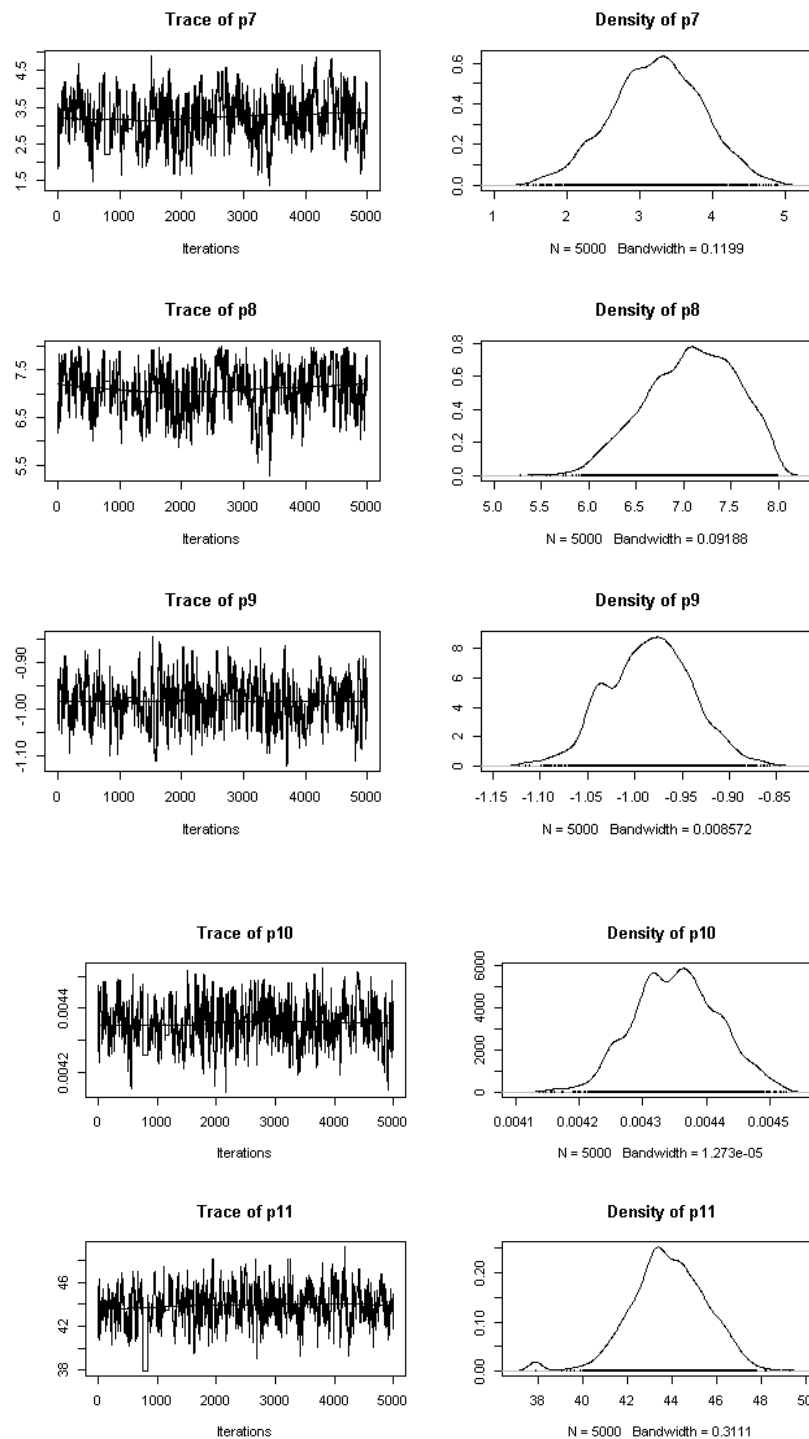


Figura 4.7. Trazas de la simulación MCMC

- De una manera más rigurosa se puede comprobar la estacionariedad de las cadenas. Mediante el test que lleva su nombre, **Geweke** (1991)

establece que si la cadena ha alcanzado el equilibrio, se debe cumplir que la media de un porcentaje (p.e. el 10%) de los valores del inicio de la cadena, debe ser aproximadamente igual a la media de un porcentaje de valores (50%) del final de la cadena. La comparación se realiza en términos de valor estandarizado (z-score), evaluándose el número de desviaciones típicas en que difieren ambas medias. Si dicho número es en valor absoluto menor o igual a 2, se considera que la cadena está en equilibrio. Se muestran a continuación los resultados del test, con los que se concluye que la cadena está en equilibrio.

```
R> geweke.diag(MC, frac1=0.1, frac2=0.5)

Fraction in 1st window = 0.1
Fraction in 2nd window = 0.5

      p1      p2      p3      p4      p5      p6      p7      p8      p9      p10     p11
1.3499  1.1748  1.1901  1.1774  1.5815  2.0341  0.6196  1.7123  0.9128  0.7883 -1.9873
```

3. Cabe mencionar un test, que aunque no se ha llevado a la práctica, es el que presenta mayor aceptación desde el punto de vista estadístico. Es el test de **Gelman and Rubin** (1992), el cual comprueba la estacionariedad, mediante la simulación de tres o más cadenas por cada parámetro. El test verifica que para cada parámetro, la varianza interna (promedio) de cada realización, se asemeja a la varianza a través del conjunto de las realizaciones. Es decir, se está comprobando directamente la ergodicidad de la cadena, siendo la estacionariedad una consecuencia directa de esa propiedad.

#### 4.3.4.2.2.-ALEATORIEDAD

El objetivo de estos diagnósticos es comprobar la aleatoriedad y en su caso, calcular el mayor desfase para el que hay correlación, dato que nos será de utilidad para realizar el proceso de adelgazamiento de la muestra. Para la comprobación de la aleatoriedad de las muestras obtenidas por simulación MCMC para cada parámetro se pueden emplear, además de la observación

realizada sobre el error estándar en la estimación de las medias, los siguientes diagnósticos:

1. **Inspección visual** de las funciones de autocorrelación. A modo de ejemplo se muestran dichas funciones para los 6 primeros parámetros. Tal y como se explica en el apartado 3.4.3.1, el tipo de dependencia entre los valores de una cadena de Markov se modela mediante una estructura del tipo AR1.

En las figuras se aprecia el alto grado de autocorrelación que existe para desfase=1, así como el alto valor de desfase para el que la autocorrelación se hace indistinguible de cero. La muestra efectiva de las cadenas se prevé que sea muy inferior a la muestra real, que es de 5.000 elementos por parámetro.

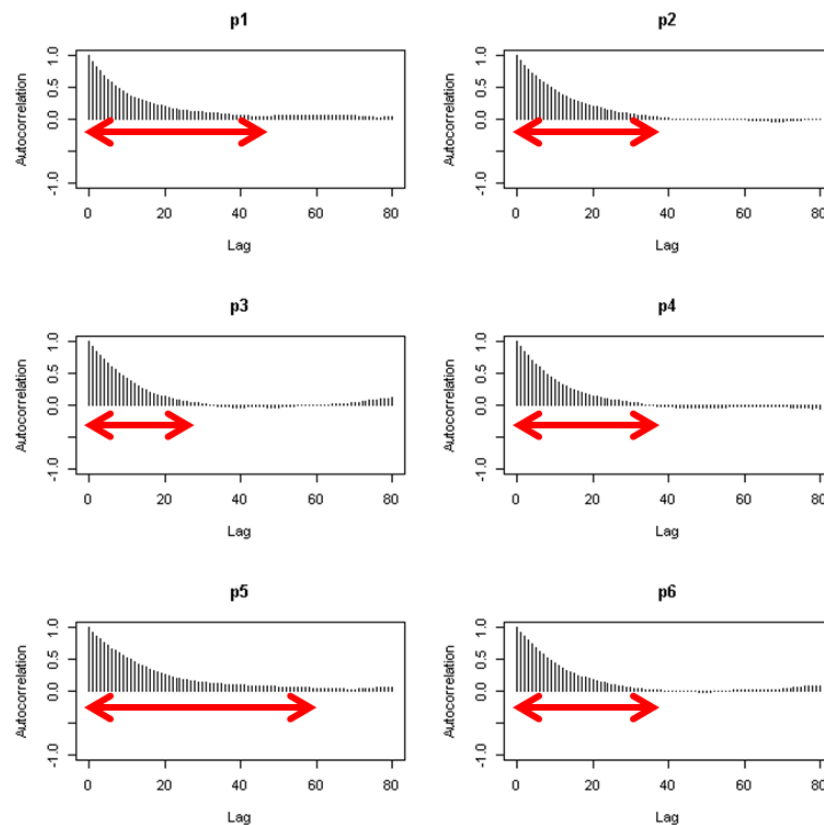


Figura 4.8. Testeo de aleatoriedad de la muestra mediante la inspección de los correlogramas



2. Test de **Raftery and Lewis** (1992) basado en la teoría de series temporales. Los resultados del test nos proporcionan la información necesaria para obtener una muestra aleatoria independiente e idénticamente distribuida, de cada uno de los parámetros. Dichos resultados ordenados por columnas son:

- Columna 1: Intervalo de **Thinning** necesario para quedarnos con una muestra, que en principio será aleatoria. Tras el proceso de adelgazamiento se hace necesario volver a pasar el test, para comprobarlo.
- Columna 2: Número de valores iniciales que deben ser descartados de la cadena, por considerar que pertenecen al periodo de calentamiento (**Burn-In**), en el que se alcanza el equilibrio.
- Columna 3: Muestra total (**N**) mínima que se necesita, considerando el efecto de autocorrelación. Es por tanto el número de simulaciones de la cadena necesarias para alcanzar con una probabilidad del 95%, la precisión deseada ( $\pm 0.005$ ) en la estimación de un cuantil de la distribución posterior del parámetro. Los cuantiles del 2.5% y del 97.5% serán los más interesantes si se quiere obtener un **intervalo de credibilidad** del 95% para el valor del parámetro.
- Columna 4: Indica el tamaño (**Lower Bound, LB**) de muestra que se necesitaría para el mismo fin anteriormente indicado, si la muestra obtenida de MCMC ya fuese puramente aleatoria.
- Columna 5: El **factor de dependencia** es el ratio  $N/LB$ . Debe ser menor de 5 para considerar que, de forma aproximada, se cuenta con una muestra aleatoria pura.

```
R> boa.randl(MC, 0.025, 0.05, 0.95, 0.001)

  Thin Burn-in Total Lower Bound Dependence Factor
p1    7      42   574         38      15.10526
p2   15     75  1425         38      37.50000
p3    9     45   684         38      18.00000
p4    7     35   434         38      11.42105
p5    8     40   536         38      14.10526
p6   11     66   902         38      23.73684
p7   11     44   605         38      15.92105
p8    9     54   648         38      17.05263
p9    7     42   490         38      12.89474
p10   8     32   480         38      12.63158
p11  15    180  1875         38      49.34211

R> boa.randl(MC, 0.975, 0.05, 0.95, 0.001)

  Thin Burn-in Total Lower Bound Dependence Factor
p1    6     30   252         38      6.631579
p2   11     66   726         38      19.105263
p3    6     24   282         38      7.421053
p4    4     20   212         38      5.578947
p5   10     50   470         38      12.368421
p6    9     54   648         38      17.052632
p7    9     45   540         38      14.210526
p8    5     25   295         38      7.763158
p9    6     30   330         38      8.684211
p10   6     30   312         38      8.210526
p11   5     20   225         38      5.921053
```

El test se pasa para cada cuantil que se quiera calcular a partir de la distribución posterior, ya que el resultado del test es sensible al cuantil de la posterior que se quiere obtener. Los resultados muestran los test para los cuantiles del 2.5% y 97.5%. La precisión se ha establecido en 0.05 en lugar del valor estándar 0.005, por motivos prácticos (reducir el número de simulaciones necesarias).

El mayor intervalo de *Thinning* requerido es 15 y será este valor el que ha de tomarse para adelgazar todas las cadenas. Además, según el test hacen falta 1875 muestras no aleatorias (parámetro P11) como la suministrada, para poder realizar los cálculos inferenciales con la precisión establecida. Dado que partimos de una muestra de 5000 vectores de parámetros, valor superior a 1875, podemos proceder a su adelgazamiento.

Tras el adelgazamiento la muestra resultante queda compuesta por 322 vectores de parámetros, los cuales deben ser una muestra aproximadamente

aleatoria de la posterior. Para comprobarlo se vuelve a realizar el test de **Raftery and Lewis (1992)**.

```
R> boa.randl(AA, 0.025, 0.05, 0.95, 0.001)

  Thin Burn-in Total Lower Bound Dependence Factor
p1    1      3   53      38          1.394737
p2    1      2   42      38          1.105263
p3    1      5   68      38          1.789474
p4    1      3   53      38          1.394737
p5    1      3   53      38          1.394737
p6    1      6   87      38          2.289474
p7    1      2   42      38          1.105263
p8    1      5   68      38          1.789474
p9    1      3   53      38          1.394737
p10   1      2   42      38          1.105263
p11   1      9  114      38          3.000000

R> boa.randl(AA, 0.975, 0.05, 0.95, 0.001)

  Thin Burn-in Total Lower Bound Dependence Factor
p1    1      9  114      38          3.000000
p2    1      5   68      38          1.789474
p3    1      2   42      38          1.105263
p4    1      5   40      38          1.052632
p5    1      5   68      38          1.789474
p6    1      3   53      38          1.394737
p7    1      5   68      38          1.789474
p8    1      3   53      38          1.394737
p9    1      2   42      38          1.105263
p10   1      5   68      38          1.789474
p11   1      2   42      38          1.105263
```

Efectivamente, las recomendaciones de adelgazamiento adoptan para todos los parámetros el valor de 1 (no adelgazamiento). Se necesita un mínimo de 114 muestras con las características estadísticas de correlación de la muestra que tenemos, que como tiene 322 elementos (>114) nos permite la realización de inferencia sobre la posterior. Se observa además que los factores de dependencia son inferiores a 5, para todos los parámetros.

#### 4.3.4.3.-EVALUACIÓN DE LA INCERTIDUMBRE DE LOS PARÁMETROS

Sobre la muestra aleatoria i.i.d. de 322 elementos de la posterior se realiza el siguiente cálculo de inferencia.



```
R> summary(AA)
Iterations = 1:322
Thinning interval = 1
Number of chains = 1
Sample size per chain = 322

1. Empirical mean and standard deviation for each
   variable, plus standard error of the mean:
2.      Mean      SD      Naive SE      Time-series SE
p1  4.199244 1.131e-01 6.304e-03 9.651e-03
p2  1.837322 5.059e-02 2.819e-03 4.278e-03
p3  6.913651 6.283e-01 3.501e-02 4.833e-02
p4  5.035049 1.083e+00 6.038e-02 8.069e-02
p5 240.828980 2.048e+01 1.141e+00 1.862e+00
p6  0.048034 1.833e-03 1.021e-04 1.512e-04
p7  3.218921 6.247e-01 3.481e-02 6.077e-02
p8  7.078877 4.799e-01 2.674e-02 4.841e-02
p9 -0.985387 4.531e-02 2.525e-03 3.427e-03
p10 0.004352 6.637e-05 3.699e-06 5.015e-06
p11 43.860377 1.704e+00 9.497e-02 1.580e-01

2. Quantiles for each variable:
      2.5%      25%      50%      75%      97.5%
p1  3.970288  4.12626  4.191025  4.269767  4.429059
p2  1.735210  1.80120  1.840135  1.872292  1.930961
p3  5.632615  6.47848  6.960535  7.428221  7.889694
p4  3.099452  4.21885  4.970229  5.797013  7.150535
p5 204.016963 225.87801 239.906816 254.773741 282.082903
p6  0.044306  0.04686  0.048100  0.049353  0.051171
p7  1.978643  2.78436  3.231754  3.656588  4.371547
p8  6.116999  6.74788  7.109351  7.439576  7.892651
p9 -1.069221 -1.01515 -0.986919 -0.954480 -0.897490
p10 0.004231  0.00431  0.004348  0.004393  0.004481
p11 40.690234 42.90559 43.798672 45.053126 46.784611
```

Ahora se puede observar que el error estándar en la estimación de la media es ya del mismo orden, se considere o no, el efecto de autocorrelación de la muestra. En la *Figura 4.9* se muestran las distribuciones posteriores de los 11 parámetros del modelo. Así mismo en la *Figura 4.10* y en la *Tabla 4.5* se ordenan los parámetros en función de la magnitud de la desviación típica de su distribución, o lo que es lo mismo, de su incertidumbre.



**DISTRIBUCIONES POSTERIORES DE LOS PARAMETROS**

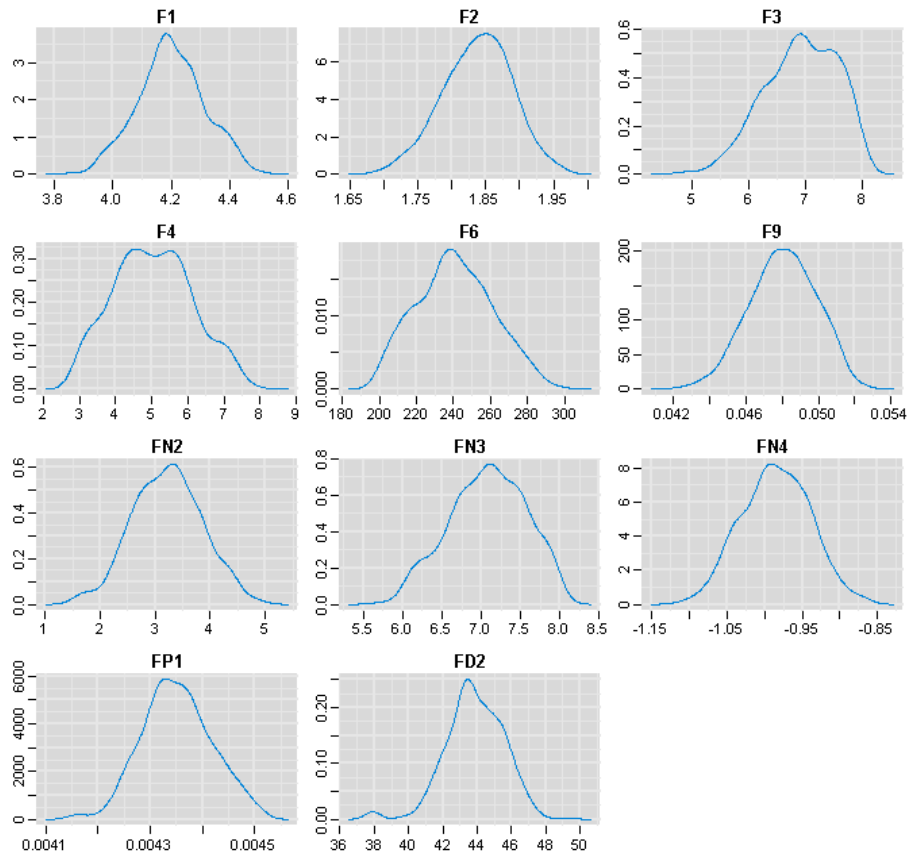


Figura 4.9. Posterior de los parámetros del modelo La Baells

Tabla 4.5. Parámetros ordenados de mayor a menor desv. típica

		Mean	SD
<b>F6</b>	<b>p5</b>	240.829	2.05E+01
<b>FD2</b>	<b>p11</b>	43.86038	1.70E+00
<b>F4</b>	<b>p4</b>	5.035049	1.08E+00
<b>F3</b>	<b>p3</b>	6.913651	6.28E-01
<b>FN2</b>	<b>p7</b>	3.218921	6.25E-01
<b>FN3</b>	<b>p8</b>	7.078877	4.80E-01
<b>F1</b>	<b>p1</b>	4.199244	1.13E-01
<b>F2</b>	<b>p2</b>	1.837322	5.06E-02
<b>FN4</b>	<b>p9</b>	-0.985387	4.53E-02
<b>F9</b>	<b>p6</b>	0.048034	1.83E-03
<b>FP1</b>	<b>p10</b>	0.004352	6.64E-05

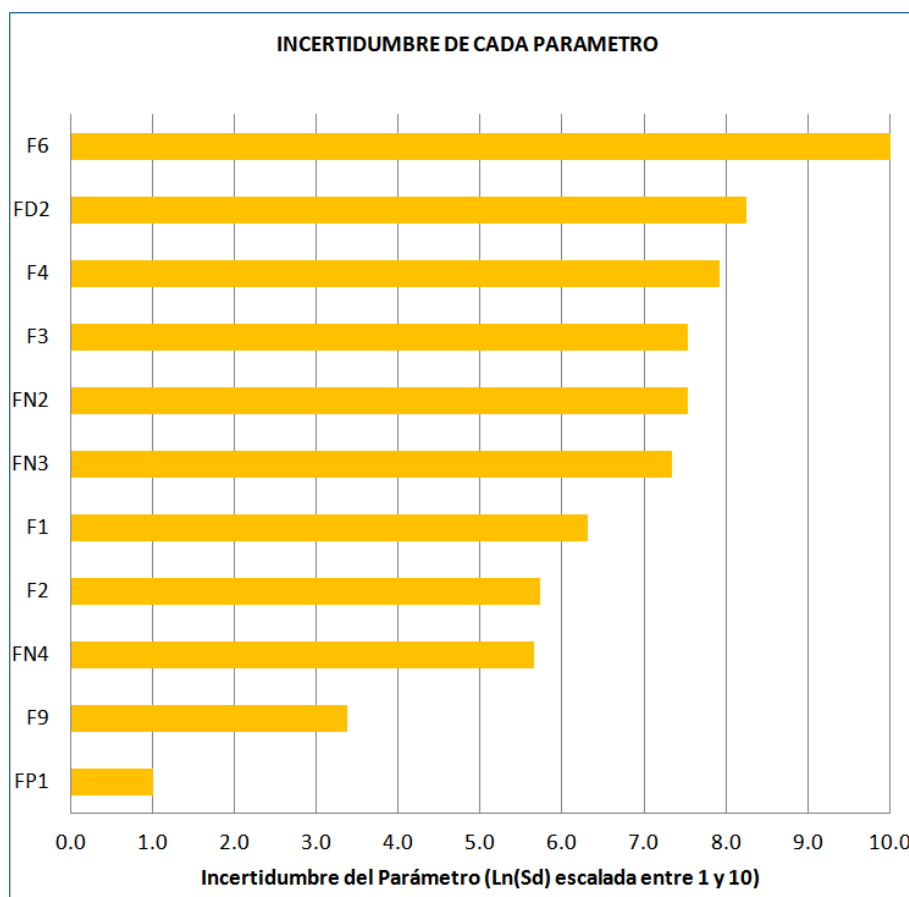


Figura 4.10. Incertidumbre de los parámetros de La Baells.

#### 4.3.5.- EFECTO DE LA INCERTIDUMBRE PARAMÉTRICA SOBRE LA PREDICCIÓN DEL MODELO

En el presente apartado se evalúa en qué medida afecta la incertidumbre de los parámetros a la incertidumbre de la predicción de la variable considerada, en este caso el caudal en el punto de control de la cuenca.

Una estimación de la distribución de la incertidumbre predictiva, que solo considere el efecto de la incertidumbre de los parámetros, se puede obtener mediante la evaluación de la predicción del modelo, para cada una de las “J” muestras aleatorias independientes extraídas mediante el algoritmo MCMC. Hecho esto, se podrá calcular para cada uno de los pasos temporales simulados por el modelo, los percentiles deseados sobre las “J” predicciones calculadas por el modelo.

Por motivos prácticos se adopta  $J=100$  en lugar de considerar los 322 vectores de parámetros. Los percentiles calculados son los correspondientes al 5%, 95%, y 50%. Así mismo se representan las envolventes de caudales máximos y mínimos. En la siguiente figura se observa el resultado.

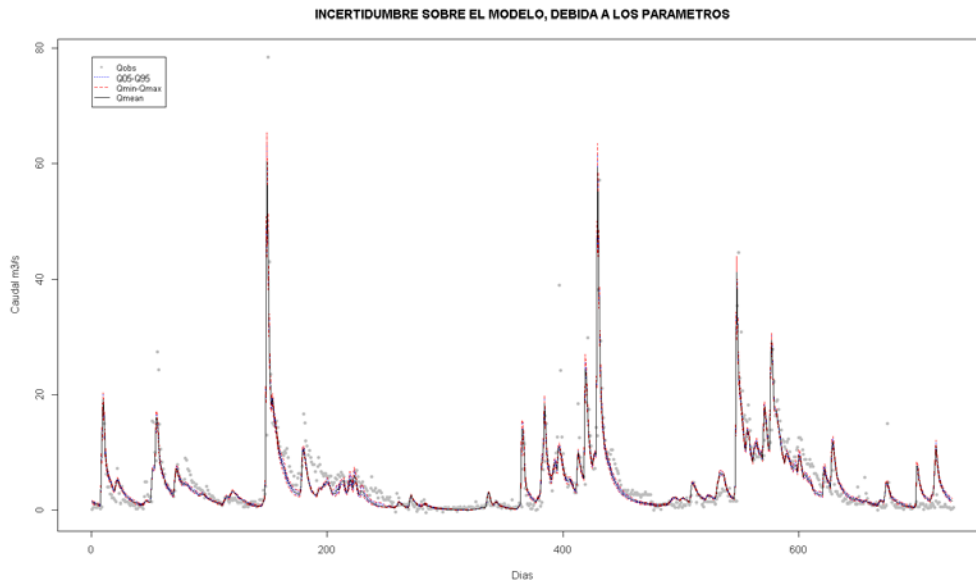


Figura 4.11. Incertidumbre predictiva debida a los parámetros.

La principal conclusión que se extrae del análisis es que el efecto de los parámetros sobre la incertidumbre predictiva del modelo es muy pequeño. De hecho las bandas de incertidumbre graficadas, que deberían contener al 90% de las observaciones, se alejan mucho de ese valor. La razón de que dichas bandas de incertidumbre de la predicción debida a los parámetros, no expliquen la mayor parte de la incertidumbre total de la predicción, es que la principal fuente de error no son los parámetros, sino los datos de entrada, las observaciones y la estructura del modelo. Todo ese error remanente queda agregado en el término de error de la Ec. 2.1. En las figuras 4.12 y 4.13 se representan dos detalles de la gráfica de incertidumbre predictiva debida a los parámetros. En ellas se aprecia que incluso las envolventes de máximos/mínimos resultan estrechas. Se observa así mismo que el ancho de las bandas es mayor en las recesiones que en los tramos de caudal bajo o en las ramas de crecida.

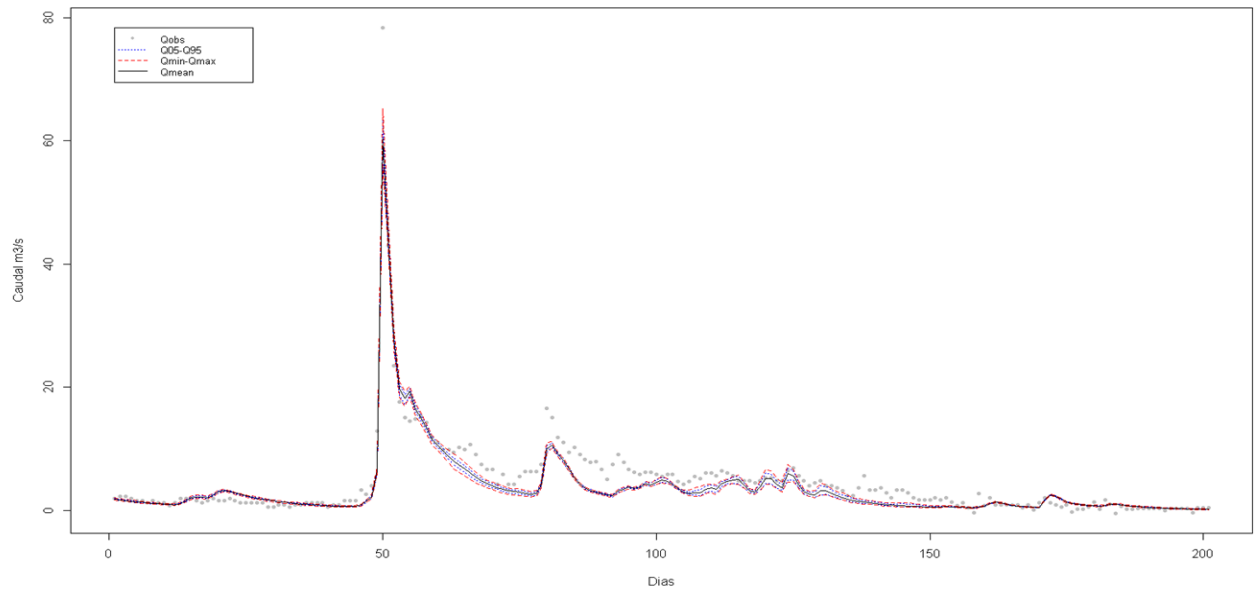


Figura 4.12. Incertidumbre predictiva debida a los parámetros. Detalle entre los días 100 y 300 de simulación.

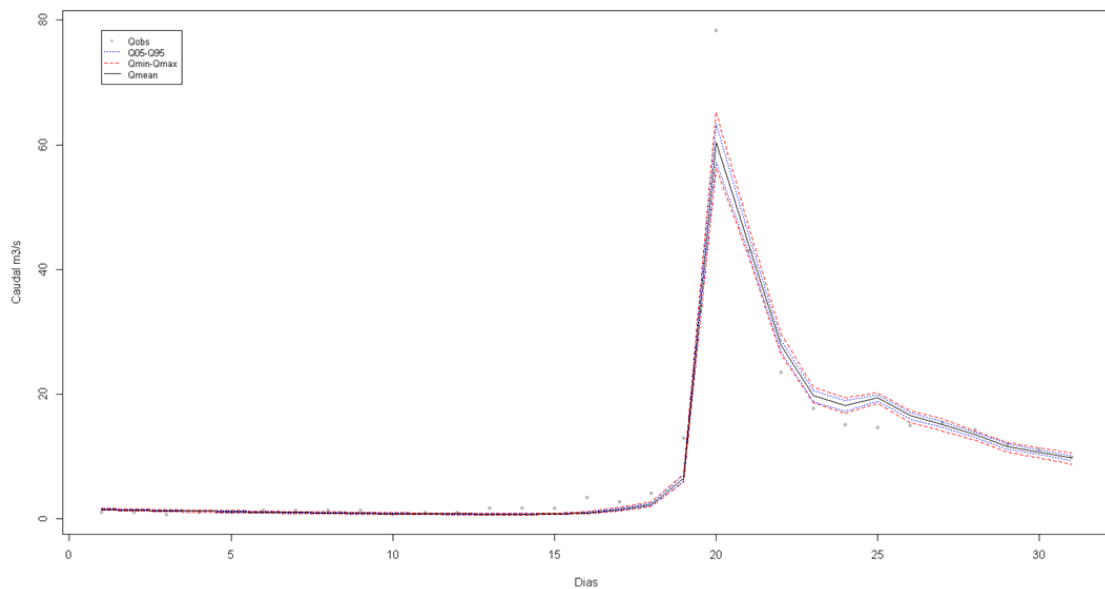


Figura 4.13. Incertidumbre predictiva debida a los parámetros. Detalle entre los días 130 y 160 de simulación.

#### 4.3.6.- SENSIBILIDAD GLOBAL DEL MODELO A LOS PARÁMETROS

En el anterior apartado se ha analizado como afecta la incertidumbre de todos los parámetros a la predicción del modelo. Sin embargo puede resultar interesante evaluar en qué medida es responsable cada parámetro de esa influencia sobre la incertidumbre de la predicción. En principio se espera que haya parámetros cuya incertidumbre influya sobre la incertidumbre de la predicción, mientras que otros no tengan un efecto significativo sobre el modelo. El resultado del análisis será una lista con los parámetros, ordenados de mayor a menor influencia sobre la incertidumbre predictiva.

El análisis que se va a efectuar se puede explicar con apoyo en la *Figura 4.14*. Se pretende evaluar el cambio que se produce para la predicción del modelo, entre un cálculo en el que se varían todos los parámetros de forma simultánea, y cada uno de los cálculos en los que se deja fijo uno de los parámetros, mientras se muestrean todos los demás. Dicha variación en la predicción es medida como variación del coeficiente de Nash.

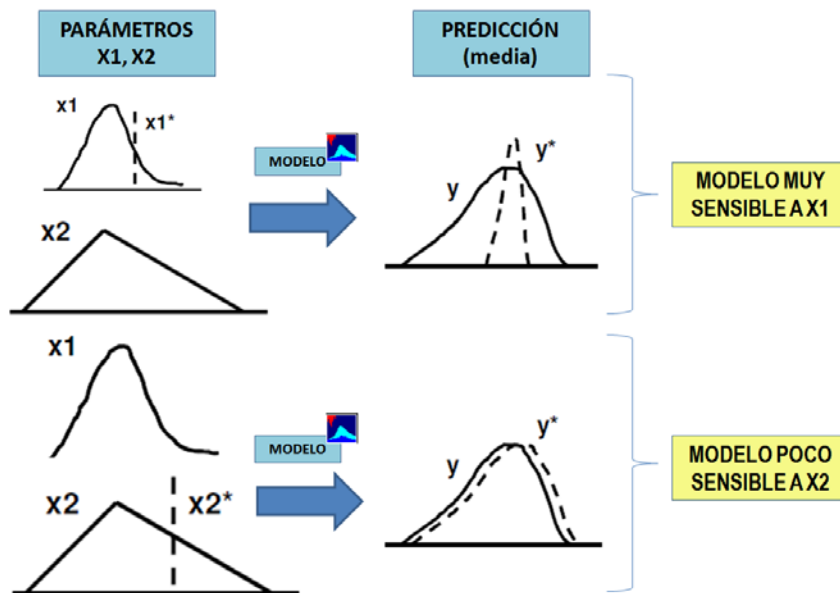


Figura 4.14. Sensibilidad Global a un parámetro.

En la tabla 4.6 se muestran los coeficientes de Nash respecto de las observaciones, para las 5 curvas (hidrograma medio,  $q_5$ ,  $q_{95}$ ,  $Q_{min}$  y  $Q_{max}$ ),

cuando se muestrea la posterior de todos los parámetros (fila A) y cuando se van muestreando todos los parámetros menos el propio de la fila (filas B).

Tabla 4.6. Coeficientes de Nash de los distintos cálculos

		MEAN	q5	q95	MIN	MAX
A)	ALL	0.8348	0.8288	0.8356	0.8243	0.8326
	F2 P2	0.8313	0.8299	0.8245	0.8272	0.8132
	FP1 P10	0.8335	0.8218	0.8364	0.8106	0.8316
	FD2 P11	0.8338	0.8259	0.8352	0.8170	0.8314
	F6 P5	0.8333	0.8280	0.8331	0.8211	0.8308
B)	FN2 P7	0.8347	0.8318	0.8334	0.8280	0.8332
	F1 P1	0.8346	0.8289	0.8337	0.8199	0.8305
	F9 P6	0.8344	0.8276	0.8360	0.8215	0.8354
	FN4 P9	0.8341	0.8298	0.8333	0.8233	0.8311
	FN3 P8	0.8343	0.8286	0.8335	0.8243	0.8309
	F4 P4	0.8348	0.8295	0.8343	0.8238	0.8346
	F3 P3	0.8347	0.8300	0.8357	0.8228	0.8329

Tabla 4.7. Variación de los coeficientes de Nash que provoca la fijación de cada parámetro.

	EFECTOS <sup>8</sup> DE CADA PARAMETRO SOBRE LA INCERTIDUMBRE PREDICTIVA DEL MODELO						EFECTO CONJUNTO	
	SOBRE MEDIA	SOBRE q5	SOBRE q95	SOBRE MIN	SOBRE MAX			
F2	0.0042	0.0013	0.0133	0.0035	0.0233	0.0456	10	
FP1	0.0015	0.0085	0.0009	0.0167	0.0012	0.0288	6.4	
FD2	0.0012	0.0036	0.0005	0.0088	0.0013	0.0154	3.5	
F6	0.0017	0.001	0.003	0.0039	0.0021	0.0118	2.7	
FN2	0.0001	0.0036	0.0027	0.0044	0.0008	0.0115	2.6	
F1	0.0002	0.0000	0.0023	0.0053	0.0025	0.0104	2.4	
F9	0.0005	0.0015	0.0004	0.0034	0.0034	0.0093	2.2	
FN4	0.0008	0.0012	0.0028	0.0013	0.0018	0.0078	1.8	
FN3	0.0006	0.0003	0.0026	0.0000	0.002	0.0056	1.4	
F4	0.0000	0.0008	0.0016	0.0007	0.0024	0.0055	1.3	
F3	0.0001	0.0015	0.0001	0.0019	0.0004	0.0039	1	

En la tabla 4.7 se representa la variación porcentual absoluta (entre filas B y fila A) en el coeficiente de Nash, que experimenta cada una de las 5 curvas. El efecto total de no considerar la incertidumbre de un determinado parámetro será la suma de los efectos sobre las 5 curvas. Dicho efecto conjunto escalado y ordenado de mayor a menor, se representa en la siguiente figura.

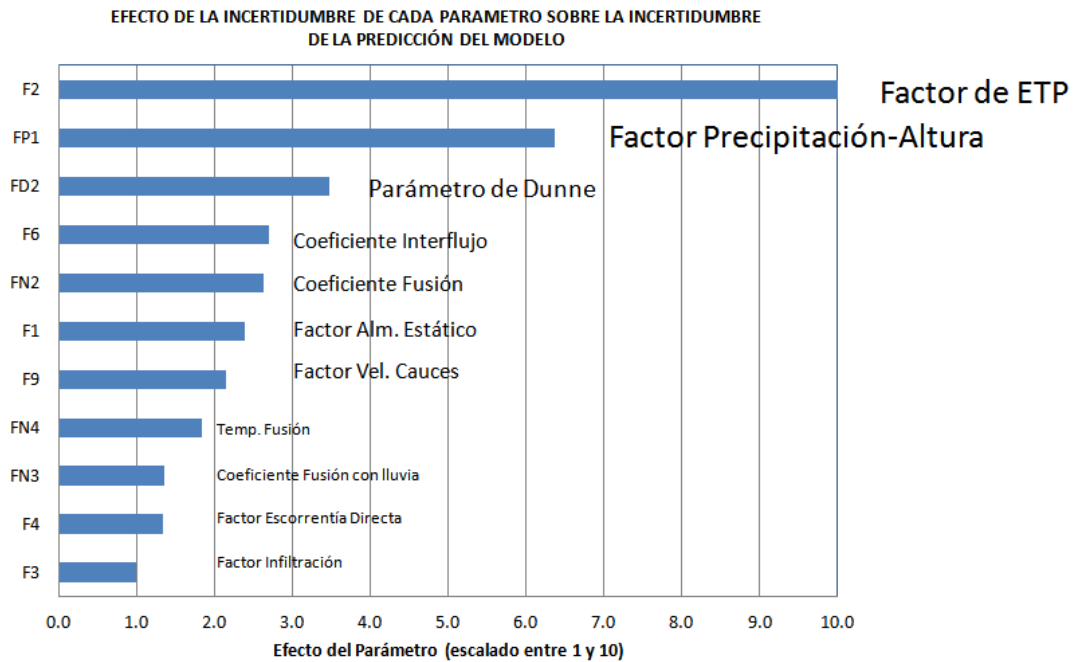


Figura 4.15. Efecto relativo de cada parámetro sobre la incertidumbre predictiva.

En la *Figura 4.16* se comparan las magnitudes relativas de la incertidumbre de los parámetros y las magnitudes relativas del efecto de los parámetros sobre la incertidumbre de la predicción.

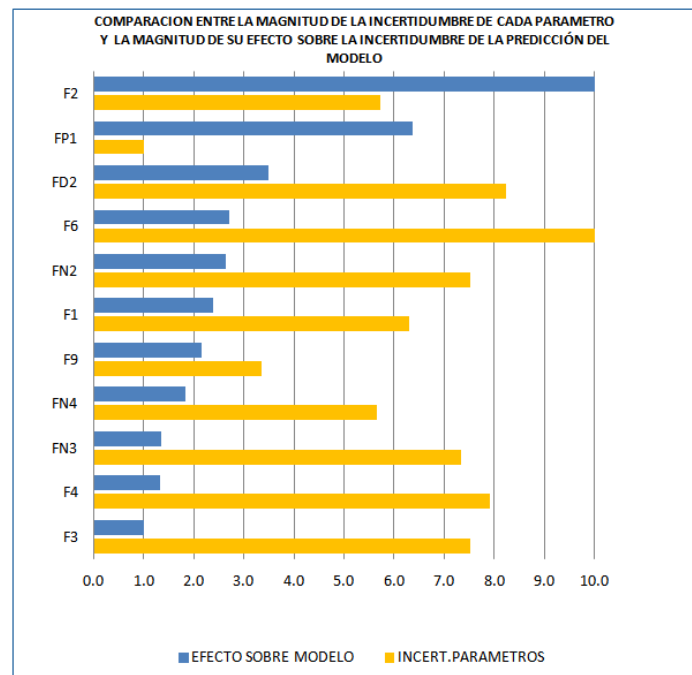


Figura 4.16. Comparación entre la incertidumbre de los parámetros y su efecto sobre la incertidumbre predictiva.

En los resultados obtenidos se observa que hay parámetros que presentan mayor incertidumbre que el resto, sin embargo su efecto sobre la predicción no es el más importante. De igual manera ocurre lo contrario, con parámetros con una incertidumbre relativa pequeña, que sin embargo son los que más influyen en la incertidumbre predictiva del modelo. Esa mayor o menor facilidad que tiene el modelo para proyectar la incertidumbre del parámetro sobre la incertidumbre de la predicción, es lo que tradicionalmente se ha conocido como sensibilidad global del modelo al parámetro.

En las siguientes figuras se explica de una manera gráfica esa relación entre la incertidumbre del parámetro, la incertidumbre de la predicción debida al parámetro y la sensibilidad global del modelo al parámetro.

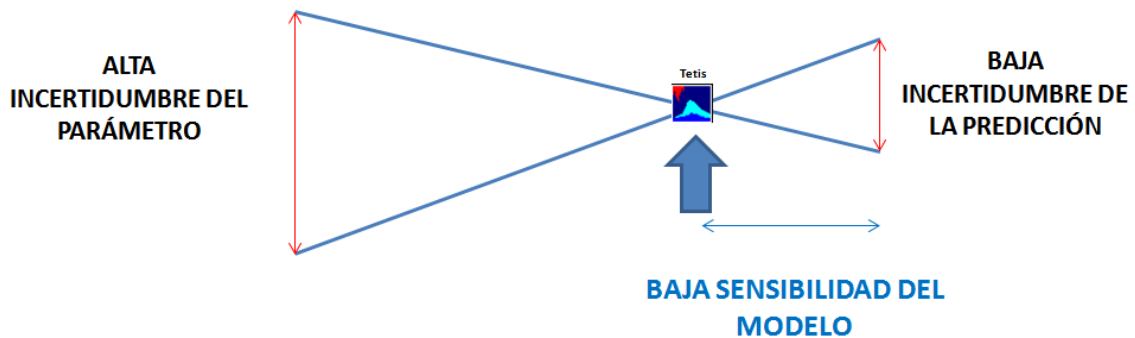


Figura 4.17. Proyección de la incertidumbre del parámetro sobre la incertidumbre predictiva (caso 1).

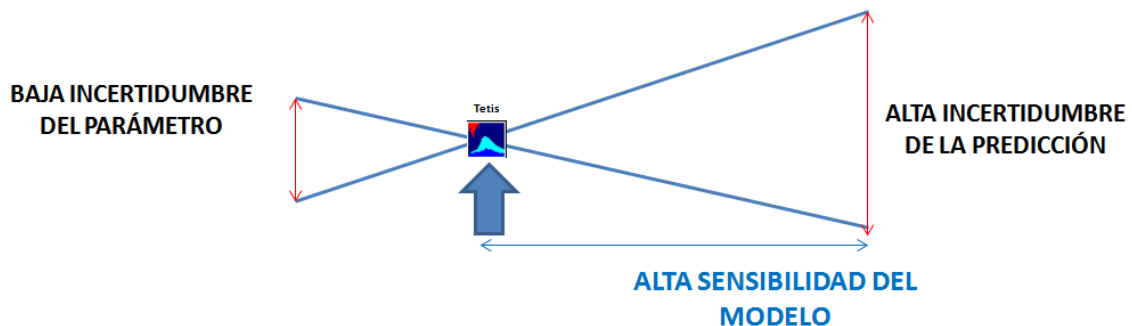


Figura 4.18. Proyección de la incertidumbre del parámetro sobre la incertidumbre predictiva (caso 2).



#### 4.3.7.- INCERTIDUMBRE TOTAL EN LA PREDICCIÓN DEL MODELO

Como se ha visto en el anterior apartado, la parte de la incertidumbre predictiva total que es debida a los parámetros resulta pequeña en comparación con la totalidad de la incertidumbre. Tal como se ha planteado el problema, la modelación del resto de la incertidumbre queda representada por el término de error de la ecuación de regresión no lineal planteada (Ec. 2.1).

El modelo de error establecido presenta un solo parámetro, la varianza residual, la cual se ha considerado estacionaria. Dicho parámetro ha sido tratado como un parámetro *nuisance* por el algoritmo DRAM, de manera que el muestreo de su distribución posterior se realiza, tal y como se explica en el apartado 3.5.3 mediante una distribución Gamma. En definitiva, a cada vector de parámetros muestreado de la posterior de los parámetros, le corresponde un valor de varianza residual muestreado de su distribución posterior Gamma.

Por tanto a cada una de las “J” predicciones que el modelo proporciona para un determinado paso temporal, se puede agregar su correspondiente término de error muestreado según el modelo de error establecido. En este caso, un modelo Gaussiano, independiente, con media cero y varianza homocedástica.

Hecho esto, se podrá calcular para cada uno de los pasos temporales simulados por el modelo, los percentiles deseados sobre las “J” predicciones calculadas por el modelo a las que se les ha añadido el correspondiente error.

Al igual que en el apartado 4.3.5, por motivos prácticos se adopta  $J=100$  en lugar de considerar los 322 vectores de parámetros. Los percentiles calculados son los correspondientes al 5%, 25%, 75%, 95%, así como la media. En las siguientes figuras se observa el resultado.

Si comparamos las bandas de credibilidad ahora representadas con las que se representaron para la incertidumbre debida a los parámetros, se observa que estas sí que recogen la mayor parte de las observaciones, con lo que dichas bandas si están explicando la incertidumbre. La cuestión es si el modelo de error es el correcto y si las citadas bandas presentan la forma adecuada.

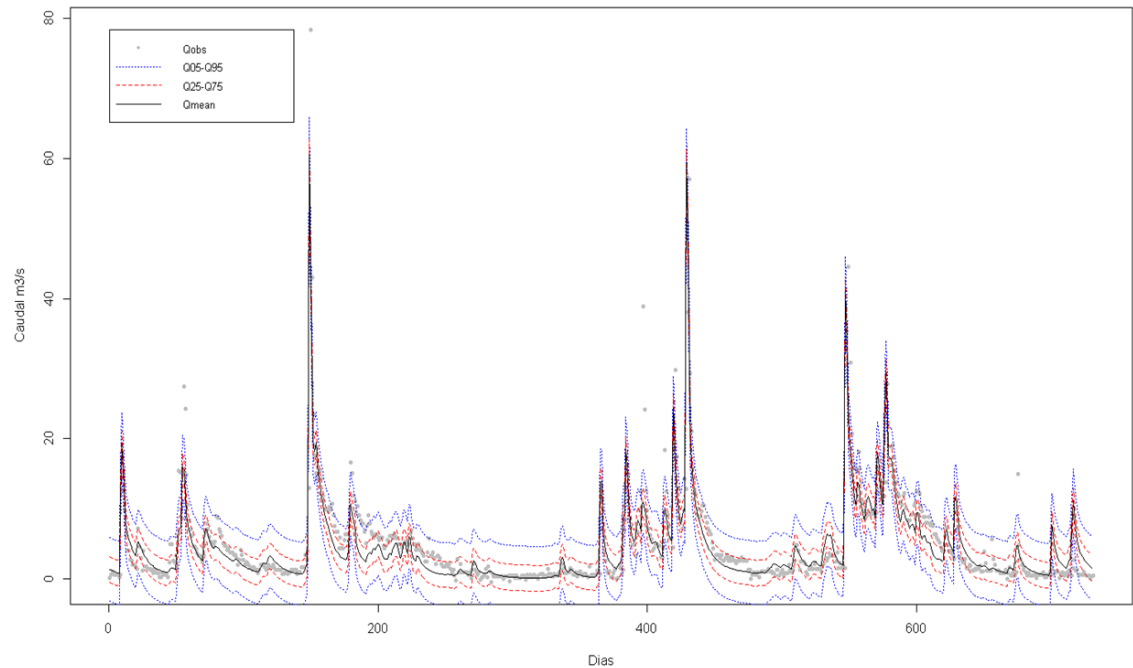


Figura 4.19. Incertidumbre total de la predicción.

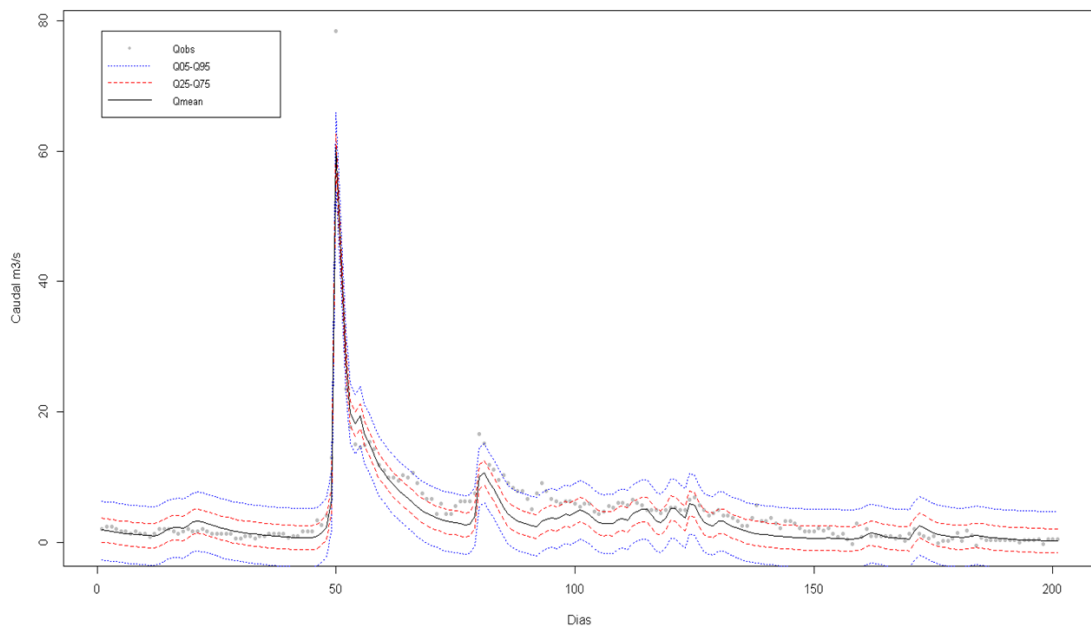


Figura 4.20. Incertidumbre total de la predicción. Detalle entre los días 100 y 300 de simulación.

#### 4.3.8.- VERIFICACIÓN DEL MODELO DE ERROR

Debido a que todos los resultados de incertidumbre quedan condicionados al modelo de error establecido a priori, debe ser contrastada la mayor o menor bondad de las hipótesis asumidas por dicho modelo.

##### 4.3.8.1.-DISTRIBUCIÓN GAUSSIANA DE LOS ERRORES

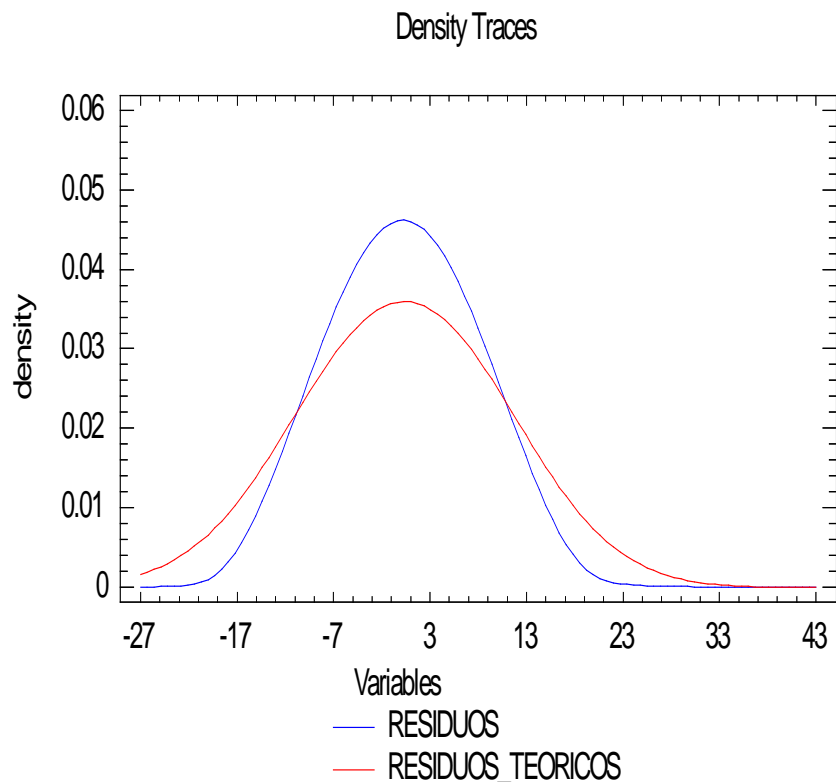


Figura 4.21. Comparación entre las distribuciones teórica y real de los residuos.

Tal y como muestra la figura, la distribución real de los errores presenta una forma más picuda (kurtosis) que la distribución Normal teórica. En cuanto a la asimetría (sesgo) de la distribución real, es prácticamente nula. Por tanto la hipótesis de Normalidad no se cumple aunque la separación de la hipótesis no es excesiva.

#### 4.3.8.2.-HOMOCEASTICIDAD

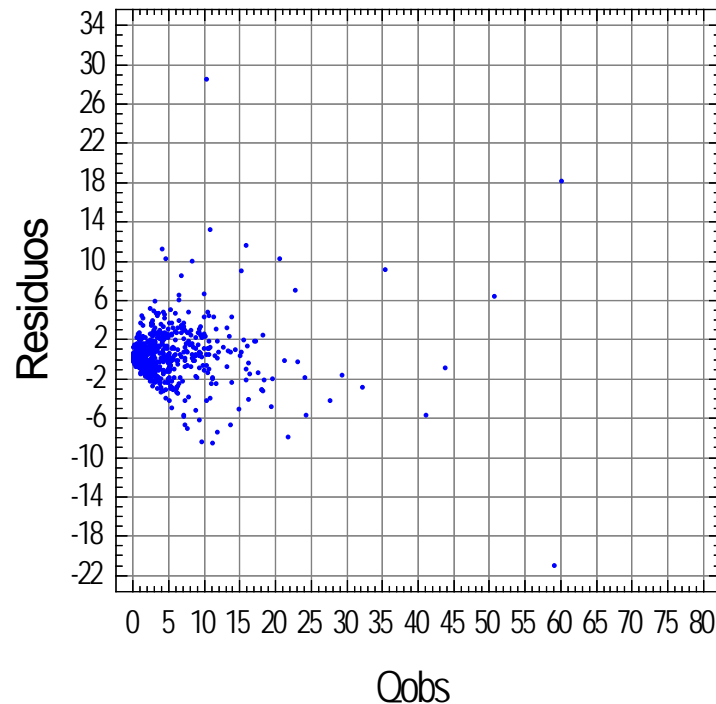


Figura 4.22. Diagrama de dispersión de los residuos en el que se aprecia Heterocedasticidad.

Como muestra la figura, con carácter general la varianza que presentan los caudales bajos es inferior a la que presentan los picos de caudal. Por tanto los errores exhiben una clara heterocedasticidad.

En lo que refiere a la incertidumbre predictiva, ignorar la heterocedasticidad considerando en su lugar una varianza del error constante (promediada), produce una **sobrevaloración de la incertidumbre para los caudales bajos**.

La sobrevaloración de la incertidumbre con caudales bajos se hace visible en el hecho de que las bandas de credibilidad llegan a adquirir valores negativos como se observa en la *Figura 4.20*.

#### 4.3.8.3.-INDEPENDENCIA DE LOS RESIDUOS

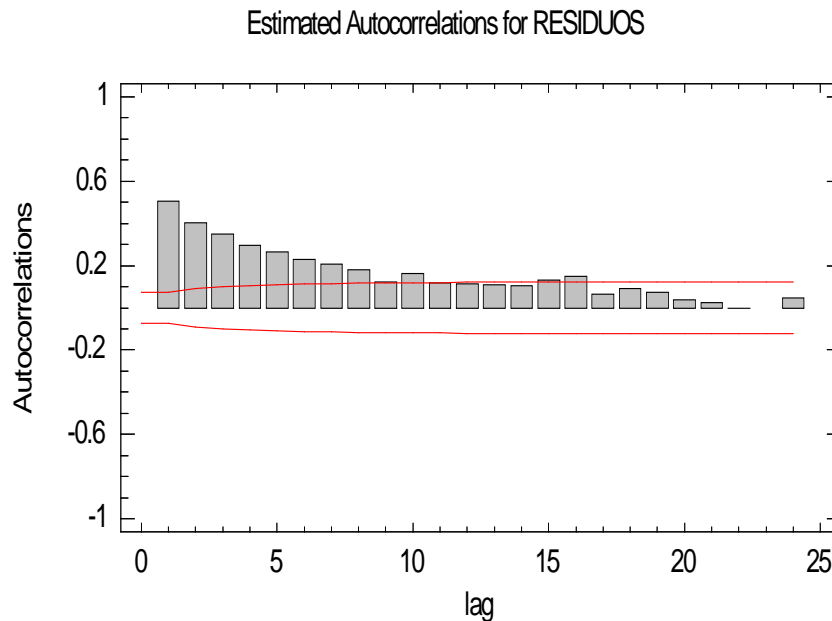


Figura 4.23. Función de Autocorrelación de los residuos.

Como muestra la función de autocorrelación de los residuos del modelo, estos presentan una estructura de dependencia temporal, en lugar de ser un perfecto ruido blanco. Las razones de que en los residuos se presenten estas correlaciones, deben fundamentarse en la no consideración de forma correcta de ciertos procesos del modelo y/o en los posibles errores de los datos de entrada y/o de las observaciones.

#### 4.3.8.4.-FORMA DE LAS BANDAS DE CREDIBILIDAD

El incumplimiento de las hipótesis del modelo de error se puede ver reflejado en la incorrecta representación de las bandas de incertidumbre.

No solo se observa en la aparición de bandas con valores negativos. Se puede verificar que el porcentaje de observaciones que contiene las bandas del p% de credibilidad, difiere significativamente del p%.

En el caso analizado, se comprueba que las bandas del 50% de credibilidad contienen el 72% de las observaciones. Por su parte, las bandas del 90% de credibilidad contienen al 94% de las observaciones.



## 5.-CONCLUSIONES

El objetivo principal de esta tesina ha sido investigar sobre el potencial de los métodos MCMC en la evaluación de la incertidumbre de los parámetros de un modelo distribuido conceptual de simulación hidrológica (**TETIS**, Francés et al. 2007) mediante la integración de dicho modelo hidrológico, programado en Fortran, con el paquete de **R** denominado **FME**, *Flexible Modeling Environment* (Soetaert and Petzoldt 2010; Soetaert 2010a; b).

Se ha seguido una aproximación formal a la estimación de la incertidumbre de los parámetros y a la incertidumbre de la predicción, asumiendo a priori un modelo estadístico de los errores cuyas hipótesis son: Normalidad, homocedasticidad e independencia. Definido el modelo de error, queda definida la función de verosimilitud de los parámetros, y por tanto mediante el teorema de Bayes, la expresión de la distribución posterior (no escalada) de los mismos, habiendo asumido Priors no-informativas. Para la evaluación de la distribución posterior, se ha empleado un método de muestreo por Monte Carlo, guiado por una cadena de Markov, métodos conocidos como algoritmos *Markov Chain Monte Carlo* (MCMC), con un algoritmo de tipo adaptativo (DRAM) que mejora la eficiencia respecto al algoritmo original MH. Obtenida una muestra de la posterior, asegurando su estacionariedad y aleatoriedad, queda evaluada la incertidumbre de los parámetros. A continuación se procede a la estimación del efecto que esa incertidumbre de los parámetros puede tener sobre la incertidumbre predictiva del modelo. Así mismo se realiza una estimación de la incertidumbre predictiva total, según el modelo de error supuesto. Dado que todos los resultados obtenidos sobre la incertidumbre están condicionados por el modelo de error asumido a priori, se verifica la bondad de las hipótesis de dicho modelo, obteniendo un resultado que confirma el incumplimiento de las hipótesis asumidas sobre la estructura del error.

La aplicación de estimación de incertidumbre se ha realizado sobre el modelo de la cuenca La Baells (Llobregat), de 504 Km<sup>2</sup> para un periodo de 731 días que abarca del 1 Octubre de 2002 hasta el 31 de Septiembre de 2004. Se

considera además un periodo previo de calentamiento de 1 año, para minimizar la influencia de las condiciones iniciales sobre los resultados del modelo.

Las conclusiones que se derivan sobre el ejercicio realizado se explican a continuación.

## 5.1.- SOBRE LA METODOLOGÍA

### 5.1.1.- DEBILIDADES

Se pueden considerar los siguientes, como puntos débiles de la metodología empleada:

- Los algoritmos MCMC pese a la continua mejora en su eficiencia, siguen siendo una herramienta con alto coste computacional, más aún cuando se emplean conjuntamente con modelos hidrológicos distribuidos.
- La herramienta (software) empleada para la simulación MCMC, no permite la evolución simultánea de varias cadenas de Markov, lo cual permitiría mejorar la eficiencia en la consecución de la distribución de equilibrio, además de poder evaluar en línea si la cadena ha alcanzado el equilibrio, evitando la ejecución a posteriori de los correspondientes Test.

### 5.1.2.- FORTALEZAS

Se pueden considerar las siguientes fortalezas de la metodología empleada:

- Los métodos MCMC empleados como métodos bayesianos formales, permiten alcanzar múltiples objetivos, a saber:
  - Pueden proporcionar un conjunto de parámetros de calibración más eficiente que otros métodos que se basan en la optimización de una función (MSE, RMSE, Nash, etc.). Este hecho será tanto



- más cierto cuanto mayor sea la interrelación (p.ej. colinealidad) entre los parámetros.
- Muestran la distribución Posterior de los parámetros, lo que permite evaluar su incertidumbre.
  - Posibilitan la evaluación de la incertidumbre de la predicción que es debida a los parámetros.
  - Los métodos MCMC permiten el empleo de conocimiento previo sobre los parámetros mediante su distribución Prior.
  - Son métodos que están abiertos a la evaluación por separado de distintas fuentes de incertidumbre.

## 5.2.- SOBRE LOS RESULTADOS

- **Respecto al modelo de error**, resulta evidente a la vista de los resultados, que se están incumpliendo las 3 hipótesis: Normalidad, independencia y homocedasticidad. La solución es, o bien aplicar alguna transformación a los datos para forzar en alguna medida su normalidad y la estacionariedad de la varianza (p.e. transformación Box-Cox), o abordar el problema de una manera más conceptual e intuitiva, reformulando el modelo estadístico de error mediante una función de verosimilitud más general, que permita considerar las tres hipótesis según se aborda en Schoups and Vrugt (2010).
- La incertidumbre predictiva del modelo que es debida a los parámetros resulta ser en general reducida. Además, la incertidumbre obtenida para los parámetros no se propaga de igual manera a través del modelo. Esto es equivalente a decir que la **Sensibilidad Global del modelo** a cada parámetro resulta diferente. Se observan 3 modos de comportamiento: Parámetros con mucha incertidumbre cuyo efecto sobre la incertidumbre del modelo es reducido, parámetros con poca incertidumbre y con un efecto





amplificado sobre la incertidumbre del modelo, y el caso intermedio entre esos dos. Como ejemplos concretos del caso estudiado se detallan los siguientes:

- Los factores F3, F4 y FN3 afectan a la predicción del modelo en menor medida que el resto de parámetros. Son parámetros que presentan umbral de insensibilización, siendo esta la causa.
- El factor que presenta menor incertidumbre es el relacionado con el incremento de lluvia con la altura (FP1). Por el contrario, es el segundo parámetro que más contribuye a la incertidumbre predictiva.
- Los parámetros F2 (ET) y FN4 (temperatura de fusión) presentan una incertidumbre parecida. Sin embargo, F2 es el parámetro que más contribuye a la incertidumbre predictiva del modelo, mientras que FN4 es de los que menos influye en dicha incertidumbre.
- El factor relacionado con el mecanismo de Dunne (FD2) es el tercer parámetro que más influye en la incertidumbre predictiva, por lo que se puede deducir que este mecanismo de escorrentía resulta relevante en el modelo analizado.

Como conclusión, en lo que se refiere a la incertidumbre de los parámetros, cabe decir que **la incertidumbre de un parámetro solo será importante cuando afecte de forma significativa a la predicción del modelo**. Por tanto, es importante conocer como de sensible resulta el modelo a cada parámetro cuando este se mueve dentro de su rango probable de valores (su distribución posterior).



## 6.-APORTES DEL TRABAJO Y FUTURAS LINEAS DE INVESTIGACIÓN

### 6.1.- APORTES DEL TRABAJO

- Realización mediante un método bayesiano formal, de un análisis de incertidumbre de los parámetros de un modelo hidrológico, así como de su incertidumbre predictiva, tanto la debida a los parámetros como la total.
- Se ha podido comprobar un mejor comportamiento del algoritmo adaptativo DRAM, respecto al algoritmo de optimización SCE-UA, en la calibración del modelo analizado.
- Desarrollo del código necesario (R y ©Matlab) que permite la utilización conjunta del paquete FME y el modelo TETIS.

### 6.2.- LINEAS FUTURAS

- En primer lugar, la programación de un algoritmo MCMC cuyo código abierto permita total libertad de operación y modificación. Dicho algoritmo MCMC debe permitir la ejecución en paralelo de múltiples cadenas, siendo esta, una forma más eficaz de alcanzar la distribución de equilibrio de la cadena, incluso con distribuciones posterior multimodales.
- Se estima necesaria la consideración de un **modelo de error más acorde** con la estructura que, con carácter general, presentan los residuos de un modelo hidrológico. Dicho modelo de error debe contemplar el modelado de residuos con varianza heterocedástica, con estructura de correlación temporal y con distribuciones más realistas (sesgadas, o con colas pesadas) que no necesariamente se corresponden con la Normal.



- Extender el análisis de incertidumbre sobre los parámetros, a otros parámetros del modelo TETIS que forman parte de módulos en desarrollo (sedimentación, nieve, etc.)
- Una de las ventajas de los métodos bayesianos formales, es que permiten la posibilidad de poder **considerar explícitamente, las distintas fuentes de error** por separado. En el presente trabajo solo se han considerado de forma explícita en el análisis de incertidumbre, los parámetros hidrológicos propios del modelo. El resto de incertidumbre debida a observaciones, datos de entrada y estructura del modelo hidrológico, se ha considerado de forma agregada en el modelo de error. Sin embargo, si es posible idear maneras de desagregar esas fuentes de incertidumbre, por ejemplo parametrizando procesos o datos relacionados con esas fuentes de incertidumbre, se hará posible su consideración en el análisis de incertidumbre. En definitiva, poder separar en alguna medida, y evaluar por separado la incertidumbre de las distintas fuentes, según Vrugt et al. (2009), "...es la clave para mejorar el conocimiento y teoría hidrológicos, que permiten predecir los flujos de agua a través de las cuencas".
- En relación con el punto segundo y el punto anterior, resulta interesante y necesario, llevar a cabo las comprobaciones necesarias para confirmar las siguientes afirmaciones encontradas en la literatura, y que están relacionadas con la bondad del modelo de error así como con la consideración explícita de algunas de las fuentes de error, en lugar de considerarlas de forma agregada en el término de error. Dichas afirmaciones son:
  1. La **incertidumbre de los parámetros de un modelo hidrológico, es infraestimada** cuando se considera un modelo de error Gaussiano i.i.d. Se observa mayor incertidumbre en los parámetros cuando en el modelo de error se consideran la correlación, la heterocedasticidad y los valores reales de asimetría y kurtosis de la distribución de errores.



2. Los **valores de los parámetros, durante la calibración, son forzados** a compensar parcialmente los errores debidos a una combinación de las distintas fuentes de error, cuando estas no se consideran de forma explícita dentro del modelo. Según Vrugt, et al. (2009), la proyección de la incertidumbre estructural así como de la incertidumbre en los datos de entrada al modelo, sobre los parámetros del modelo, es lo que da lugar a lo que Beven et al. han dado en llamar en una serie de artículos desde 1992, Equifinalidad.
  3. Las estimaciones de parámetros obtenidas bajo la hipótesis del modelo de error Gaussiano i.i.d. no son robustas en cuanto que resultan muy sensibles al periodo de calibración considerado. El empleo de un modelo de error que considere la correlación, la heterocedasticidad así como la forma real de la distribución (p.e. con colas más pesadas que la distribución Normal, algo que parece ser común en modelación hidrológica), permite una inferencia menos sensible a los *outliers* en pequeños o grandes eventos, lo que resulta en una **estimación de parámetros menos sesgada y más consistente**.
  4. En lo que refiere a la incertidumbre predictiva, ignorar la heterocedasticidad considerando en su lugar una varianza del error constante (promediada), produce una **sobrevaloración de la incertidumbre para los caudales bajos, al tiempo que infravalora la incertidumbre en la predicción de los picos**.
- Como complemento a lo anterior, y dado que los métodos de muestreo Monte Carlo requieren un alto coste computacional, puede ser muy interesante la paralelización del código del modelo hidrológico TETIS, lo que permitiría el abordaje en tiempos más razonables de este tipo de análisis.



## 7.-REFERENCIAS BIBLIOGRÁFICAS

- Aguilera, P., Fernández, A., Fernández, R., Rumí, R., and Salmerón, A. (2011). “Bayesian networks in environmental modelling.” *Environmental Modelling & Software*, Elsevier Ltd, 26(12), 1376–1388.
- Beck, M. (1987). “Water quality modeling: a review of the analysis of uncertainty.” *Water Resources Research*.
- Best, N., Cowles, M., and Vines, S. (1995). “CODA Manual version 0.30.” *MRC Biostatistics Unit, Cambridge, UK*.
- Beven, K. (1987). “Towards a new paradigm in hydrology.” *Water for the Future: Hydrology in Perspective. IAHS n° 164*.
- Beven, K. (1989). “Changing ideas in hydrology—The case of physically-based models.” *Journal of hydrology*, Elsevier, 105(1-2), 157–172.
- Beven, K. (1993). “Prophecy, reality and uncertainty in distributed hydrological modelling.” *Advances in Water Resources*, 16(1), 41–51.
- Beven, K. (2006). “A manifesto for the equifinality thesis.” *Journal of Hydrology*, 320(1-2), 18–36.
- Beven, K. (2007). “Towards integrated environmental models of everywhere: uncertainty, data and modelling as a learning process.” *Hydrology and Earth System Sciences*, 11(1), 460–467.
- Beven, K., and Binley, A. (1992). “The future of distributed models: model calibration and uncertainty prediction.” *Hydrological processes*, 6(3), 279–298.
- Beven, K., and Freer, J. (2001). “Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology.” *Journal of Hydrology*, 249(1-4), 11–29.
- Bolstad, W. (2010). *Understanding computational Bayesian statistics*. Wiley.
- Bras, R., and Rodríguez-Iturbe, I. (1985). *Random functions and hydrology*. Addison-Wesley.
- Brendecke, C. (2012). “Supplemental Comments on Model Uncertainty.”
- Brun, R., and Reichert, P. (2001). “Practical identifiability analysis of large environmental simulation models.” *Water Resources Research*, 37(4), 1015–1030.



- Bussi, G. (2010). “Calibración de un modelo conceptual distribuido del ciclo de sedimentos. Aplicación a la cuenca experimental de Goodwin Creek (EEUU).” *MSc Tesis. Universidad Politécnica de Valencia.*
- Clark, J. (2004). “Why environmental scientists are becoming Bayesians.” *Ecology letters.*
- Clark, J., and Gelfand, A. (2006). “A future for models and data in environmental science.” *Trends in ecology & evolution*, 21(7), 375–80.
- Cowles, M., and Carlin, B. (1996). “Markov chain Monte Carlo convergence diagnostics: a comparative review.” *Journal of the American Statistical ...*
- Cullen, A., and Frey, H. (1999). *Probabilistic techniques in exposure assessment: a handbook for dealing with variability and uncertainty in models and inputs.*
- Dennis, B. (1996). “Discussion: should ecologists become Bayesians?” *Ecological Applications.*
- Duan, Q., and Sorooshian, S. (1992). “Effective and efficient global optimization for conceptual rainfall-runoff models.” *Water Resources Research.*
- Francés, F., Vélez, I., and Vélez, J. (2007). “Split-parameter structure for the automatic calibration of distributed hydrological models.” *Journal of Hydrology*, 332(1-2), 226–240.
- Gallart, F., Delgado, J., Beatson, S., Posner, H., Llorens, P., and Marcé, R. (2011). “Analysing the effect of global change on the historical trends of water resources in the headwaters of the Llobregat and Ter river basins (Catalonia, Spain).” *Physics and Chemistry of the Earth, Parts A/B/C*, Elsevier Ltd, 36(13), 655–661.
- Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2003). *Bayesian data analysis.* (Chapman and Hall, eds.).
- Gelman, A., and Rubin, D. (1992). “Inference from iterative simulation using multiple sequences.” *Statistical science.*
- Geweke, J. (1991). “Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments.” *Bayesian statistics.*
- Gilks, W., and Richardson, S. (1995). *Markov Chain Monte Carlo in Practice.* (Chapman and Hall, eds.), London.
- Gourley, J., and Vieux, B. (2006). “A method for identifying sources of model uncertainty in rainfall-runoff simulations.” *Journal of Hydrology*, 327(1-2), 68–80.
- Grayson, R., and Blöschl, G. (2001). “Spatial modelling of catchment dynamics.” *Spatial Patterns in Catchment Hydrology: ...*



- Haario, H., Laine, M., Mira, A., and Saksman, E. (2006). “DRAM: Efficient adaptive MCMC.” *Statistics and Computing*, 16(4), 339–354.
- Hastings, W. (1970). “Monte Carlo sampling methods using Markov chains and their applications.” *Biometrika*.
- He, M., Hogue, T., Franz, K., Margulis, S., and Vrugt, J. (2011). “Characterizing parameter sensitivity and uncertainty for a snow model across hydroclimatic regimes.” *Advances in Water Resources*, Elsevier Ltd, 34(1), 114–127.
- Herrera, S. (2011). “Desarrollo, validación y aplicaciones de Spain02: Una rejilla de alta resolución de observaciones interpoladas para precipitación y temperatura en España.” *PhD Tesis. Universidad de Cantabria*.
- Herrera, S., Gutiérrez, J., Ancell, R., Pons, M., Frías, M., and Fernández, J. (2012). “Development and analysis of a 50-year high-resolution daily gridded precipitation dataset over Spain (Spain02).” *International Journal of Climatology*, 32(1), 74–85.
- Kuczera, G., Kavetski, D., Franks, S., and Thyer, M. (2006). “Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters.” *Journal of Hydrology*, 331(1-2), 161–177.
- Kuczera, George, and Parent, E. (1998). “Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm.” *Journal of Hydrology*, 211(1-4), 69–85.
- Laine, M. (2008). “Adaptive MCMC methods with applications in environmental and geophysical models.” *Finnish meteorological institute contributions*.
- Laloy, E., and Vrugt, J. (2012). “High-dimensional posterior exploration of hydrologic models using multiple-try DREAM (ZS) and high-performance computing.” *Water Resources Research*, 48(1), 1–18.
- Leopold, L., and Maddock, T. (1953). “The hydraulic geometry of stream channels and some physiographic implications.” *United States Government Printing Office*.
- Li, L., Xia, J., Xu, C., and Singh, V. (2010). “Evaluation of the subjective factors of the GLUE method and comparison with the formal Bayesian method in uncertainty assessment of hydrological models.” *Journal of Hydrology*, 390(3-4), 210–221.
- Martinec, J. (1960). “The degree-day factor for snowmelt runoff forecasting.” *IUGG General Assembly of Helsinki, IAHS ...*
- Matott, L., Babendreier, J., and Purucker, S. (2009a). “Evaluating Uncertainty in Integrated Environmental Models: A Review of Concepts and Tools, Supporting Information.”



- Matott, L., Babendreier, J., and Purucker, S. (2009b). "Evaluating uncertainty in integrated environmental models: A review of concepts and tools." *Water Resources Research*, 45(6), 1–14.
- Metropolis, N., and Rosenbluth, A. (1953). "Equation of state calculations by fast computing machines." *The journal of chemical ...*
- Montoya, J. (2008). "Desarrollo de un modelo conceptual de producción, transporte y depósito de sedimentos." *PhD Tesis. Universidad Politécnica de Valencia*.
- Moradkhani, H., and Sorooshian, S. (2008). "General review of rainfall-runoff modeling: Model calibration, data assimilation, and uncertainty analysis." *Hydrological Modelling and the Water Cycle*, 1–24.
- Omlin, M., and Reichert, P. (1999). "A comparison of techniques for the estimation of model prediction uncertainty." *Ecological Modelling*, 115(1), 45–59.
- Popper, K. (1957). "Philosophy of science: A personal report." *British philosophy in the mid-century*.
- Puricelli, M. (2003). "Estimación y Distribución de parámetros del suelo para la modelación hidrológica." *PhD Tesis. Universidad Politécnica de Valencia*, 434.
- Raftery, A., and Lewis, S. (1992). "How many iterations in the Gibbs sampler." *Bayesian statistics*.
- Ramirez, L., Francés, F., and Bussi, G. (2012). "Análisis de la implementación de un modelo hidrológico distribuido con información estándar en España." *MSc Tesis. Universidad Politécnica de Valencia*.
- Refsgaard, J., Van der Sluijs, J., Højberg, A., and Vanrolleghem, P. (2007). "Uncertainty in the environmental modelling process – A framework and guidance." *Environmental Modelling & Software*, 22(11), 1543–1556.
- Reichert, P., and Omlin, M. (1997). "On the usefulness of overparameterized ecological models." *Ecological Modelling*, 95(2-3), 289–299.
- Ríos, S., and Ríos, D. (2001). "La arquitectura del conocimiento complejo, el sentido común y la probabilidad."
- Saltelli, A., Ratto, M., Tarantola, S., and Campolongo, F. (2006). "Sensitivity analysis practices: Strategies for model-based inference." *Reliability Engineering & System Safety*, 91(10-11), 1109–1125.
- Schoups, G., and Vrugt, J. (2010). "A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors." *Water Resources Research*, 46(10), 1–17.





- Silva, L. (2001). “El enfoque bayesiano: otra manera de inferir.” *Gac Sanit*, 15(4), 341–346.
- Soetaert, K. (2010a). “R Package FME : Inverse Modelling , Sensitivity , Monte Carlo – Applied to a Nonlinear Model.”
- Soetaert, K. (2010b). “R Package FME: Tests of the Markov Chain Monte Carlo Implementation.”
- Soetaert, K. (2010c). “R Package FME : Inverse Modelling , Sensitivity , Monte Carlo – Applied to a Dynamic Simulation Model.”
- Soetaert, K., and Petzoldt, T. (2010). “Inverse modelling, sensitivity and monte carlo analysis in R using package FME.” *Journal of Statistical Software*, 33(3).
- Soetaert, K., and Petzoldt, T. (2012). *Package “ FME ” Manual*.
- Uusitalo, L. (2007). “Advantages and challenges of Bayesian networks in environmental modelling.” *Ecological Modelling*, 203(3-4), 312–318.
- Vrugt, J., and Braak, C. (2011). “DREAM(D): an adaptive Markov Chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems.” *Hydrology and Earth System Sciences*, 15(12), 3701–3713.
- Vrugt, J., Braak, C., Diks, C., Robinson, B., Hyman, J., and Higdon, D. (2009). “Accelerating Markov Chain Monte Carlo Simulation by Differential Evolution with Self-Adaptive Randomized Subspace Sampling.” *International Journal of Nonlinear Sciences and Numerical Simulation*, 10(3), 273–290.
- Vrugt, J., Braak, C., and Gupta, H. (2009). “Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?” *research and risk*, 23(7), 1011–1026.
- Vrugt, J., Braak, C., Gupta, H., and Robinson, B. (2008). “Response to comment by Keith Beven on ‘Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?’” *Stochastic Environmental Research and Risk Assessment*, 23(7), 1061–1062.
- Vrugt, J., Gupta, H., and Bouten, W. (2003). “A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters.” *Water Resources Research*, 39(8), 10.
- Vélez, I. (2001). “Desarrollo de un modelo hidrológico conceptual y distribuido orientado a la simulación de las crecidas.” *PhD Tesis. Universidad Politécnica de Valencia*.



- Wu, Y., and Liu, S. (2012). “Automating calibration, sensitivity and uncertainty analysis of complex models using the R package Flexible Modeling Environment (FME): SWAT as an example.” *Environmental Modelling & Software*, Elsevier Ltd.
- Yapo, P., and Gupta, H. (1996). “Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data.” *Journal of Hydrology*, Elsevier, 181(1-4), 23–48.
- Zhang, X., Liang, F., Yu, B., and Zong, Z. (2011). “Explicitly integrating parameter, input, and structure uncertainties into Bayesian Neural Networks for probabilistic hydrologic forecasting.” *Journal of Hydrology*, Elsevier B.V., 409(3-4), 696–709.