METHODS TO COMPARE MEASUREMENTS AND SIMULATIONS

E. Palomo; J. Marco
Institute of Renewable Energy - CIEMAT
Avd. Complutense, 22
SP - 28040 Madrid, Spain

H. Madsen
Inst. of Mathematical Statistics and Operations Research
Building 321, The Technical University of Denmark
DK - 2800 Lyngby, Denmark

ABSTRACT

The comparisson between measurements and simulations is a very important stage in the methodology for empirical whole model validation of building energy simulation programs developed within the PASSYS project. The aim of this paper is to describe and to evaluate several statistical tools that could be used for this purpose.

INTRODUCTION

In 1986 the PASSYS project was formed by the Commission of the European Communities with the aim to increase confidence in passive solar heating systems through the development of a component testing procedure, the development of a validation methodology for building energy simulation programs, and the development of better simplified design tools.

PASSYS involves research consortia from Belgium, Denmark, France, Germany, Greece, Italy, the Netherlands, Portugal, Spain and United Kingdom. The work of PASSYS is undertaken by four specialist subgroups addressing test methodology, simulation model validation and development, simplified design tools development and test site management.

The main objective of the model validation and development subgroup (MVD) has been to approve/refine a European validation methodology and to test this by applying it to a building simulation program - the ESP energy modelling system (Clarke 1985), which by the Commission of the European Communities has been selected as the European reference simulation program.

In the first phase of PASSYS (1986-89) main emphasis has been devoted to review the theory behind the different heat transfer processes of the program, to check the corresponding source code segments, to apply analytical verification and inter-model comparison where possible, and to perform sensitivity studies.

In the second phase (1990-91), the attention has been focussed on empirical model validation. This paper deals with part of the methodology for empirical whole model validation developed by the MVD subgroup. Several statistical tools available to compare measurements and simulations are described and its use for model validation purposes discussed.

METHODOLOGY FOR EMPIRICAL WHOLE MODEL VALIDATION

The aim of performing empirical whole model validation is to detect if a model is capable of describing reality correctly. This is, however, a non-trivial task to perform, as it requires expertise in experimental design, modelling principles and simulation techniques. A methodology which ensures that one can rely on the results from the validation and that a maximun of information is obtained from the validation study is necessary. The methodology developed by the MVD subgroup comprises six stages (Ostergaard and van de Perre 1991):

1) Definition of the scope, type and nature of the physical and numerical experiment.
2) Implementation of the physical experiment on site.
3) Process the measured data.
4) Perform simulations.
5) Analyse results.
6) Document data set and validation work.

The analysis of the results is a very important stage of the methodology. Here we assess how the model related to observed data, to prior knowledge, and to its intended use. Deficient model behaviour make us to reject the model or to recommend modifications if possible, while good performance will develop a certain confidence in the model. There exist several statistical techniques for comparisson between measured and simulated values, testing the goodnes of different aspect of the model. Within the MVD subgroup *Sensitivity Analysis* and *Residual Analysis* have been applied.

As indicated before, this paper concentrates on Residual Analysis. However, in order to introduce the necessity of this kind of analysis, a short description of the capabilities and limitations of Sensitivity Analysis, which has been widely used for empirical model validation tasks, is included.

Sensitivity Analysis

Using Differential Sensitivity Analysis (DAS) or Monte Carlo Analysis (MCA), two types of sensitivity are being evaluated within PASSYS: i) individual sensitivities, which describe the influence on simulations of variations in each individual input; and ii) total sensitivities, due to the uncertainties in all the input data.

A knowledge of the individual sensitivities has several practical benefits for empirical whole model validation purposes. To identify the inputs to which the outputs are particularly sensitive (critical parameters) and those to which they are insesitive, is essential to guide the design of experiments. Useful information on which signals to measure, where and when to measure them, how to command them if possible, and which parameters must be known 'a priori' with a minimun uncertainty, can be achieved using Sensitivity Analysis techniques.

The total uncertainty in outputs due to all the inputs uncertainties enables to assess the resolution of the models. This information can be used for empirical validation studies when simulations are compared with measurements. However, it is only a first approach to the comparisson process: If measurements lead outside the total uncertainty bands, we can conclude that the model is deficient. On the contrary, if mesurements lead within these bands, judgements about the validity of the model require anhother kind of statistical tools.

Main disadvantages of Sensitivity Analysis when used

to compare measurements and simulations are: i) Results are very sensitive to the considered uncertainties in input parameters; ii) These uncertainties mainly are based on subjective judgement and experience; iii) The method does not test dynamic parts of the model, especially high dynamic parts. Anhother methods has, therefore, also been applied.

Residual Analysis

In this paper the deviations between the simulated values and the measurements are called the residuals. This is perhaps a little misleading since the term residual most often is used for the deviation between an estimated model and the measurements - the so-called prediction error. This convention has be adopted in this paper although there is a conceptual difference between simulation and prediction error.

For the prediction error case it is assumed that the residuals from an estimated model behave as white noise and are orthogonal to the model space. This means that, if the models gives a sufficient description of the measurements, then the residuals and the parameter space are independent, or, let us put it another way, the residuals do not contain any information on the parameters. Then we are able to use statistical test based on the residuals to deduce whether or not the model gives a reasonable description of the measurements. Usually, model inadecuacy can be evaluated by examining (see eg. Box and Jenkins 1976; Vandaele 1983): i) a residual plot; ii) the autocorrelation function of the residuals, and iii) the cross-correlation function between the explanatory variables of the model and the residuals.

For the simulation error case the situation is different. In this case it is most unlikely that the residuals from the simulated model are orthogonal to the model space. This means that, in the simulation situation, the residuals can be used for characterizing model inadequacies. The purpose of the following section is to describe some methods used in statistics to validate an estimated model, and to discuss how these methods can be used in the simulation case.

STATISTICAL TOOLS TO COMPARE MEASUREMENTS AND SIMULATIONS

The most simple way to compare measurements and simulations is just to depict the trace of the simulated values together with the measured values. This is important and should never be overlooked. However, there is readily a lot of questions asked: Is the mean value of the differences between the simulated and measured values zero? Are there deviations between the start of the simulation and the end of the simulation? And how do we characterize the deviations? This questions will be considered in this section.

Let us introduce a system which can be described by the nonlinear stochastic differential equation:

$$dx(t) = f(x(t),u(t),t,\theta,w(t))dt \qquad (1)$$

The vector f is a time dependent nonlinear function of the state, x(t), the input, u(t), the model parameters, θ, and the noise, w(t).

The measurements are described by the equation;

$$y(t) = g(x(t),u(t),t,e(t)) \qquad (2)$$

where the vector of measurements is time dependent nonlinear function of the state, the input and the measurements error, e(t).

The simulated system is

$$dx'(t) = f'(x'(t),u'(t),t,\theta'))dt \qquad (3)$$

with the output of the simulation:

$$y'(t) = g'(x'(t),u'(t),t) \qquad (4)$$

Thus the simulations contains no description of the noise part. There is a lot of reasons for deviations of the symbols between the true system and the simulated system. These are modelling approximations, unrecognized and undescribed input signals, noise corrupted input signals and measurement errors.

Let us define the residuals:

$$\epsilon(t) = y(t) - y'(t) \qquad (5)$$

Description of the Variation of the Residuals

Relevant characteristics of the residuals can be investigated by analyzing present non-stationarities, its persistence (autocorrelation function), and its spectral properties (autospectrum).

In a similar way, if attention is focused on second-order properties, the multivariate stochastic process defined by the residuals and the explanatory variables of the model - eg. solar irradiance, outdoor air temperature, ... - could be caracterized by examining the cross-correlation matrix of the process, its spectral matrix, and the partial coherency spectra if explanatory variables are cross-correlated.

This section includes definitions and comments on the meaning of these statistics. More information can be found in specific books about these topics - eg. (Jenkins and Watts 1968; Priesley 1981)

Stationarity

A stochastic process is said to be strictly stationary if its properties are unaffected by a change of time origin; that is, the joint distribution of any set of observations must be unaffected by shifting all the times of observations forward or backward by any integer amount. It is said to be stationary up to order m if all its joint moments up to order m are independent of the absolute time. Usually, the term stationary is applied to a stationary process up to order 2.

Different methods exist to check non-stationarity: 1) Despict residuals; 2) Estimate mean and variance over different time periods; 3) Analyze the autocovariance function.

Systematic change in the level of a time series (trend) is a typical kind of non-stationarity. Several techniques exist to remove it. However, as it is extremely difficult to tell whether a change in the level of the series is due to a deterministic or to a stochastic trend, the differencing method is one of the most effective removing the trend present in all types of series. This method consist of substracting the values of the observations from one another in same prescribed time-dependent order.

Autocorrelation Function

A stationary stochastic process $\epsilon(t)$ is simply described by its autocovariance function

$$\gamma_\epsilon(u) = E[(\epsilon(t)-\mu_\epsilon)(\epsilon(t+u)-\mu_\epsilon)] \qquad (6)$$

The displacement u ($u=t_2-t_1$) is called the lag. The acvf

shows how the dependence between adjacent values in the stochastic process changes with lag u. It is specially useful to detect "local correlations", that is whether neighboring points of the process are correlated.

Since $\gamma_\epsilon(u)$ depends on the scale of measurement of ϵ, it is convenient, in comparing two time series with possibly different scales of measurement, to define a normalized quantity called the autocorrelation function (acf)

$$\rho_\epsilon(u) = \gamma_\epsilon(u) / \gamma_\epsilon(0) \qquad (7)$$

being $\sigma_\epsilon^2 = \gamma_\epsilon(0)$ the variance of the process. Note that $\rho_\epsilon(u)$ lies between limits -1 and +1, corresponding to the complet negative and positive linear dependence.

Autospectrum

An equivalent description of a stationary stochastic process is provided by its power spectrum, which is the Fourier transform of the acvf.

$$\Gamma_\epsilon(f) = \int_{-\infty}^{\infty} \gamma_\epsilon(u)\, e^{-j2\pi/u}\, du \qquad (8)$$

The power spectrum curve shows how the variance of the stochastic process is distributed with frequency. It has an inmediate physical interpretation as an energy/frequency distribution.

As it was for the acvf, it is somtimes useful to normalize $\Gamma_\epsilon(f)$ by dividing by the variance σ^2_ϵ. The function

$$\Gamma_\epsilon(f)/\sigma^2_\epsilon \qquad (9)$$

is called the spectral density function, which is the Fourier transform of the acf.

Cross-correlation Matrix

Two stationary stochastic process $\{\epsilon(t), X_i(t)\}$ are usefully described in the time domain by their auto- and cross covariance functions. The cross-covariance function (ccvf) of lag u with $X_i(t)$ leading $\epsilon(t)$

$$\gamma_{\epsilon i}(u) = E\,[(\epsilon(t)-\mu_\epsilon)(X_i(t+u)-\mu_i)] \qquad (10)$$

measures the linear dependence existing between adjacent values in both stochastic processes - local correlation. For instance, if neighboring points in two time series are cross correlated, it would be expected that the ccvf will be large in the neirborhood of the origin and small at values distant from the origin.

To study the interactions between two processes with possibly different scales of measurement, it is necessary to define the cross-correlation function (ccf)

$$\rho_{\epsilon i}(u) = \gamma_{\epsilon i}(u) / \sigma_\epsilon\, \sigma_i \qquad (11)$$

For multivariate processes, if attention is focused on second-moment properties, the process can be characterized by the called lagged covariance matrix

$$\mathbf{V}(u) = \{\gamma_{ij}\} \; ; \qquad i,j = \epsilon,1,2,..,q \qquad (12)$$

Spectral Matrix

The theoretical cross covariance function of two stochastic processes has a Fourier transform called the cross-spectrum, $\Gamma_{\epsilon i}(f)$,

$$\Gamma_{\epsilon i}(f) = \int_{-\infty}^{\infty} \gamma_{\epsilon i}(u)\, e^{-j2\pi fu}\, du \qquad (13)$$

which can be represented as the product of the cross amplitude spectrum, $\alpha_{\epsilon i}(f)$, and the phase spectrum, $\phi_{\epsilon i}(f)$. The cross amplitude spectrum shows whether frequency components in one series are associated with large or small amplitudes at the same frequency in the other series. Similarly, the phase spectrum shows whether frequency components in one series lag or lead the components at the same frequency in the other series.

A more useful quantity than the cross amplitude spectrum, is the coherency spectrum

$$k^2_{\epsilon i}(f) = \frac{\alpha^2_{\epsilon i}(f)}{\Gamma_\epsilon(f)\,\Gamma_i(f)} \; ; \qquad (0 \le k^2_{\epsilon i}(f) \le 1) \qquad (14)$$

This statistic is useful in practice because it provides a non-dimensional measures of the correlation between two time series as a function of frequency. On an other hand, the coherency spectrum is unaltered when filetring operations are performed on the original time series. Thus is to be preferred to the cross amplitude spectrum.

For a multivaraite process, the matrix of auto- and cross spectra,

$$\mathbf{\Gamma}(f)=\{\Gamma_{ij}\} \; ; \qquad i,j = \epsilon,1,2,\ldots,q \qquad (15)$$

is called the spectral matrix.

The Squared Multiple Coherency Spectrum

For our purposes, it will be useful to consider the stochastic processes $\epsilon(t)$, $X_1(t)$, .., $X_q(t)$ related linearly by the general dynamic model

$$\epsilon(t) - \mu_\epsilon = \int_{-\infty}^{\infty} h_{\epsilon,1}(u)\,\{X_1(t-u) - \mu_1\}\, du + \ldots\ldots \qquad (16)$$

$$+ \int_{-\infty}^{\infty} h_{\epsilon,q}(u)\,\{X_q(t-u) - \mu_q\}\, du + Z(t)$$

where $X_i(t)$ (i=1,2,..,q) are the explanatory variables of the model (inputs), Z(t) is an independent and white noise, μ_i (i= 1,2,...,q,q+1) is the mean value of the i process and $h_{\epsilon,i}$ (i= 1,2,..,q) is the impulse response function from the input $X_i(t)$ to the output $\epsilon(t)$.

The spectrum of the noise process Z(t)

$$\Gamma_Z(f) = \Gamma_\epsilon(f) - \sum_{i=1}^{q} H_{\epsilon i}(f)\,\Gamma_{\epsilon i}(f) \qquad (17)$$

could be written as

$$\Gamma_Z(f) = \Gamma_\epsilon(f)\,[1 - k^2_{\epsilon 12..q}(f)] \qquad (18)$$

where

$$k^2_{\epsilon 12..q}(f) = \frac{1}{\Gamma_\epsilon(f)} \sum_{i=1}^{q} H_{\epsilon i}(f)\,\Gamma_{\epsilon i}(f) \qquad (19)$$

is called the squared multiple coherency spectrum of the output process and the q input processes. The multiple coherency spectrum measures the proportion of the output spectrum which can be predicted from the inputs at different frequencies.

The Squared Partial Coherency Spectra.

In the case of multivariate time series, it is useful to be able to measure the cross spectrum between

the residulas (output) and one of the inputs processes after allowance is made for the effect of the other input processes. This leads to the partial cross spectrum, which is the frequency domain analog of the partial correlation.

Predicting the output $\epsilon(t)$ from past values of all the inputs except one of them (eg. $X_k(t)$), leads to residuals that will be represented by $\epsilon_\epsilon(t)$. Similarly, predicting the input $X_k(t)$ from all the other inputs leads to the residuals $\epsilon_k(t)$. The *partial cross covariance function* between $X_k(t)$ and $\epsilon(t)$, after allowing for $X_j(t)$, j<>i, is defined as the cross-covariance function between the residuals $\epsilon_k(t)$ and $\epsilon_\epsilon(t)$. The *partial cross spectrum* is the Fourier transform of the partial cross covariance function. And the *partial cross spectral density* (or partial cross spectrum) is defined as the quotient between the partial cross spectrum and the squared root of the autospectra of the residuals $\epsilon_k(t)$ and $\epsilon_{q+1}(t)$.

For q inputs, the partial cross spectrum between the output $\epsilon(t)$ and the input $X_k(t)$ is given by

$$k_{k\epsilon|K}(f) = - \frac{\pi_{\epsilon k}(f)}{\{\pi_{\epsilon\epsilon}(f)\ \pi_{kk}(f)\}^{1/2}} \qquad (20)$$

where π_{lm} is the minor of the element Γ_{lm} in the spectral matrix of all (q+1) variables. The *squared partial coherency*, $k^2_{k\epsilon|K}$, is the squared modulus of (20) and the *partial phase spectrum*, $\Phi_{k\epsilon|K}$, its argument. The first statistic measures the squared covariance "at frequency f" between the processes $\epsilon(t)$ and $X_k(t)$ when allowance is made for the influence of the other inputs. The second one, measures the "direct" phase difference between both processes after allowing for the phase differences between $X_k(t)$ and the other inputs and between $\epsilon(t)$ and the inputs $X_j(t)$ with j<>k.

Some Tests for White Noise

A white noise sequence $\epsilon(t)$ is a sequence where the autocovariance function is zero for lags other than zero,

$$\mathrm{Cov}[\epsilon(t),\epsilon(t+k)] = 0 \quad \text{for} \quad k <> 0 \qquad (21)$$

In this section some rather simple tests for white noise are outlined. These tests are widely used in time series analysis for validating an estimated model, since the fundamental assumption on e.g. ARMAX-models is that the stochastic process is the output of a linear transfer function with white noise input.

A Sign Test

For a white noise sequence with zero mean we will expect that the probability for a change in the sign of the residual from time to time is 1/2, i.e.

$$P\{\text{change in sign from } t-1 \text{ to } t\} = 1/2 .$$

On the assumption that the residual is white noise, the individual changes in sign will be independent, hence

$$\text{Total number of change in sign} \approx B(N-1,\ 1/2) ,$$

where N is the total number of residuals. For large values of N the Binomial distribution can be approximated by the normal distribution, since

$$B\ (N-1,\ 1/2) \approx N\ ((N-1)/2,\ (N-1)/4) \qquad (22)$$

for large values of N. Hence the total number of changes signs can be compared with the above normal distribution, and if the number is outside some resonable confidence limits then the assumption of white noise must be rejected.

Test Based on the Autocorrelation Function

If the residuals is white noise then

$$\hat{\rho}_\epsilon(k) \in_{\text{approx.}} N\ (0,\ 1/N) \qquad (23)$$

Furthermore, the individual values are approximately independent.

The above fact can be used to test individual values in the autocorrelation function. 95% confidence limits for single autocorrelation value is thus $\pm 1.96\ \sqrt{(1/N)}$.

A test based on the important part of the total autocorrelation function is the *Portmanteau lack of fit test*:

$$Q^2 = N\ (\hat{\rho}_{\epsilon(\theta)}(1)^2 + \hat{\rho}_{\epsilon(\theta)}(2)^2 + \ldots + \hat{\rho}_{\epsilon(\theta)}(m)^2)\ \in_{\text{approx.}} X^2(m) \qquad (24)$$

where a reasonable value of m is about 15 - 30. If k parameters is estimated then the degree of freedom is (m-k) instead of m.

The above test quantity is simply based on the fundamental fact that a sum of m squared independent normal distributed variables becomes X^2-distributed with m degrees of freedom.

As described the normality of the estimated values of the autocorrelation of the residuals is only approximative. A more precise calculation leads to the *Ljung-Box statistics*:

$$Q^2 = N(N-2) \sum_{k=1}^{m} \hat{\rho}_{\epsilon(\theta)}(k)^2 / (N-k)\ \in_{\text{approx.}} X^2(m) \qquad (25)$$

Test in the Frequency Domain

For the frequencies $f_i = i/N$; i=0,1, ... , (N/2), the *periodogram* for the residuals is calculated as

$$\hat{I}(f_i) = 1/N\ [(\sum_{t=1}^{N} \epsilon_t \cos 2\pi f_i t)^2 + (\sum_{t=1}^{N} \epsilon_t \sin 2\pi f_i t)^2] \qquad (26)$$

which is a description in the frequency domaine of the variation of the residuals, where $I(f_i)$ tells how much of the variation of the residuals which is at the frequency f_i.

The *Cumulated Periodogram* is then defined as:

$$\hat{C}(f_j) = [\sum_{i=1}^{j} \hat{I}(f_i)] \ / \ [\sum_{i=1}^{N/2} \hat{I}(f_i)] \qquad (27)$$

which is a non-decreasing function of the frequencies.

For white noise the variation is equally distributed on all frequences (hence the term 'white'), and the total variation of N observations is equal to $N\sigma_\epsilon^2$, hence the theoretical periodogram for white noise is

$$I(f_i) = 2\sigma_\epsilon^2 \qquad (28)$$

and the theoretical cumulated periodogram is thus a straight line from (0,0) to (0.5, 1). On the assumption

that the residuals is white noise, then the estimated cumulated periodogram will be close to that straight line. The closseness is evaluated by a Kolmogorov-Smirnov test, which are known in tests for given distribution functions.

If the estimated cumulated periodogram falls outside the lines given by the confidence limits, then the white noise assumption must be rejected. Furthermore, the frequency where the estimated line falls outside indicates that the residuals still contains important correlation in time for that frequency.

Some tests for non-orthogonality

To test whether two time series are correlated or not, several procedures exist. Provided both time series have been prefiltered to convert them to white noise, the sample cross-correlation function (ccf) of the filterd series can be used to test whether they are correlated. However, the ccf is useful only in detecting local correlation. If there is a tendency for the ccf to contain periodoc components, these may not be detected using the ccf. Hence it is also necessary to use frequency-domain tests.

Test based on the cross-correlation function.

If the residuals $\epsilon(t)$ and the explanatory variable $X_i(t)$ are uncorrelated processes then

$$\hat{\rho}_{\epsilon i}(k) \in_{\text{approx.}} N(0, 1/N) \qquad (29)$$

The above fact can be used to test individual values in the cross-correlation function. 95% confidence limits for single correlation value is thus $\pm 1.96\sqrt{(1/N)}$.

As indicated before, the normality of the estimated values of the cross-correlation between the residuals and the explanatory variables is only approximative. A more precise calculation leads, as in the case of the acf, to the *Ljung-Box statistics*:

$$Q^2 = N(N-2)\sum_{k=1}^{m} \hat{\rho}_{\epsilon i(\theta)}(k)^2/(N-k) \in_{\text{approx.}} X^2(m) \qquad (30)$$

Test in the frequency domain.

If no correlation exist between the residuals and the set of explanatory variables of the model, then in (17)

$$H_{\epsilon k}=0 \; ; \qquad k=1,2,\ldots,q \qquad (31)$$

and, consequently,

$$k^2_{\epsilon 12..q}(f) = 0 \qquad (32)$$

Hence, a check for non-zero squared multiple coherency would be useful to test the hypothesis that non zero-correlation exists between the whole set of inputs and the residuals.

Such a test is based in the fact that the random variable

$$\frac{\hat{k}^2_{\epsilon 12..q}(f)}{1-\hat{k}^2_{\epsilon 12..q}(f)} \cdot \frac{v-2q}{q} \qquad (33)$$

is distributed as $F_{2q,v-2q}$. q is the number of inputs and v the variance ratio of the spectral window used to estimate the squared multiple coherency.

If the hypothesis of non-zero squared multiple coherency is unacceptable, other interesting test could be use in order to know how each input contributes to explain the residuals. Such a test is based on the squared partial coherency. This statistic is related with the squared multiple coherency by

$$1 - k^2_{k\epsilon|K}(f) = \frac{1 - k^2_{\epsilon 12...q}(f)}{1 - k^2_{\epsilon K}(f)} \qquad (34)$$

where $k^2_{\epsilon K}(f)$ is the squared multiple coherency calculated without considering the input $X_k(t)$.

If $\epsilon(t)$ and $X_k(t)$ are uncorrelated processes, then

$$1 - k^2_{\epsilon 12...q}(f) = 1 - k^2_{\epsilon K}(f) \qquad (35)$$

and

$$k^2_{k\epsilon|K}(f) = 0 \qquad (36)$$

Hence, the spectral estimator of the squared partial coherency could be used as a test for non-zero cross correlation between residuals and each input in the frequency domain. Zero values mean that no correlation exists between the considered input and the residuals, unity values mean that the residuals could be completly recovered from the input, and values between 0 and 1 correspond to situations where the residuals can be partially predicted from the input.

Some Suggestions for Model Validation

This part of the paper includes some suggestions on the use of the preceeding concepts and tests for empirical whole model validation purposes:

1) Depict simulations and measurements, or, alternatively, the residuals.
2) Characterize the residuals by mean, variance, autocovariance function and spectrum. The spectrum tells us about hig frequency versus low frequency variation.
3) Apply some simple test for white noise on the residuals. Perhaps only the sign test. The quantities given by the other tests are rather useless for model validation, since we do not expect the residuals to be orthogonal to the model.
4) Characterize the multivariate process {residuals, explanatory variables of the model}. As we are looking for causal relations between residuals and explanatory variables, the most interesting characterization of the process could be done by examining the croos-correlation functions between residuals and inputs, the squared multiple coherency of the whole process and the squared partial coherency corresponding to each input. Last option is better than to anlyze the squared coherency because inputs are frequently cross correlated.
5) If no clear information could be achieved simply examining the above mentioned statistics, apply some of the proposed tests for non-zero correlation between residulas and inputs.

AN EXAMPLE ON THE USE OF THE PROPOSED METHODOLOGY

This section includes an example of using the above described statistical tools.

The System

The system is a PASSYS test cell. The cell's structure is based on a "hard casing" concept, where the supports and the insulation are separated. The cell is prefabricated with a rugged steel frame shelved with mineral wool. The outer surfaces are covered with chipboard and, finally, shielded with stainless steel. The inner surfaces are insulated by styrofoam to give a constructional U-value of < 0.1 W/m^2K. Internally the

cell has two zones, a service room and a test room, divided by a well insulated partition containing a sealed connecting door. The possibility exists for attaching different south wall components. During the analyzed experiment, an opaque calibration wall was incorporated as south component.

The Data

The analyzed data sets were supplied by ABACUS, University of Strachclyde, UK. From 7th February to 10th March 1989, an experiment was carried out on two test cells with calibration wall. One cell free-floated acted as control, the other underwent the following schedule: a preconditioning period, a four-hour rectangular radiant heat pulse of 2kW followed by a period of free-floating until the cell was within 0.5 C of the control cell, a two hour rectangular convective heat pulse of 2kW followed by a similar cooling period, five days of constant 30 C internal temperature heating, and a final free-floating decay. Service room temperatures were constant troughout, at about 20 C. The measurement taken were standard climate parameters, internal air temperature and energy input in the heating periods.

Sensitivity Analysis

ESP simulations were run for the period 7th February to 9th March 1989, during the period of the experiment, and therefore measured climatic data were used in the simulations. The output parameter of interest in this case was the internal air temperature. The assumed geometry for the test cell, its thermo-physical and optical properties, and the definition of its surroundings and its operation modes, can be found in (Pinney and Strachan 1989). This document also includes the list of parameters we are varying in the simulations and their assumed uncertainties.

Figure 1 shows the resulting base case and associated uncertainty bands estimated by DSA techniques, together with the measured internal air temperature and the prevailing external temperature. Measurements are inside the uncertainty bands except for the initial period. No more than 10% of the measured values for the indoor air temperatute leads outside these bands. Dangerous conclusions could be achieved from this observation if no more analysis were performed comparing measurements and simulations. It will be necessary, at least, to perform simply visual comparissons between measurements and base case simulations, which are assumed as the most probable response of the system.

Residual Analysis

When measurements leads inside the uncertainty bands, judgements about the validity of the model require another kind of analysis. Using the above described statistical tools, an analysis on the residuals was performed.

Characterizing Residuals

Looking at figures 1 and 2, several relevant characteristics of the residuals can be derived:

i) Residuals show a clear trend. From time 0 to time 150, they are changing from +2C to -2C values. Then, they fluctuate around 1.5C.
ii) Residuals show two clear peaks. One in the radiant heat pulse period (+7C) and the other in the convective heat pulse period (-5C).
iii) Measured temperatures are generally lower than the simulated ones. It seems that radiant heat input leads to underestimations and convective heat input to overestimations of the indoor air temperature.
iv) The measured decay is faster than the simulated one.

For the whole period of measurements/simulations mean and variance values for the residuals were -1.4C and $1.3C^2$, respectively. A mean value different from zero means that the model, with the set of selected values for its parameters, can not represent adequately the stationary behaviour of the test cell. The variance is a measure of the fluctuations around the mean value.

Persistency in the residuals is analyzed by examining its sample autocorrelations estimated (Fig. 3). As residuals are non-stationary, a first oder differncing process has been applied on the residuals before to estimate them. The resulting autocorrelations show a spike at lag 4, and they start to be zero from lags greather than 6. This two observations are enough to conclude that the residuals do not behave as white noise. No other statistical test for white noise are necessary.

How the variance is distributed over frequency is analyzed through the estimated density power spectrum of the residuals. Most part of the variance is concentrated at low frequency. After differencing the residuals - high-pass filter -, the resulting spectrum (Fig. 4) show two wide bands where the variance is concentrated: one around 0.12 h^{-1} and the other around 0.35 h^{-1}. The model, with the set of selected values for its parameters, not only fails when reproducing stationary or very low dynamic regimes, but also when simulating intermediate and high dynamic regimes.

The Multivariate Process: Residuals/Inputs

More information about what may cause deviations between measurements and simulations could be derived from the analysis of the multivariate process residuals/inputs.

A first analysis could be performed on the sample cross correlations estimated. Figures 5 and 6, show how residuals are correlated with service room air temperatures (fig.4) and with power heating (fig.6). Before to estimate their cross-correlations, these time series have been differentiated. From lag -10 to 10, both curves show correlations values outside the 95% confidence bands and pronounced patterns, especially in the case of the power heating. No more cross-correlations curves are shown - eg. solar irradiance, outdoor air temperature, wind velocity, wind direction and relative humidity - because no significant correlation has been detected between residuals and these other inputs.

In the frequency domain, two kind of analysis have been carried out examining the squared multiple coherency and the squared partial coherencies.

Figures 7 includes the squared multiple coherency and the squared partial coherency for the power heating and for the service room air temperature estimated after differencing the original time series. Figure 8 contents the same information but for the non-differenciated time series. Because the analyzed set of time series is not stationary, this second procedure is not applicable in a strict sense, but it could be useful as a first approach to investigate the model behaviour at very low frequency.

Main conclusion when examining the estimated squared multiple coherency is that a great proportion of the residuals spectrum can be predicted from the explanatory variables of the model. When examining the estimated squared partial coherencies, the conclusions are: i) The

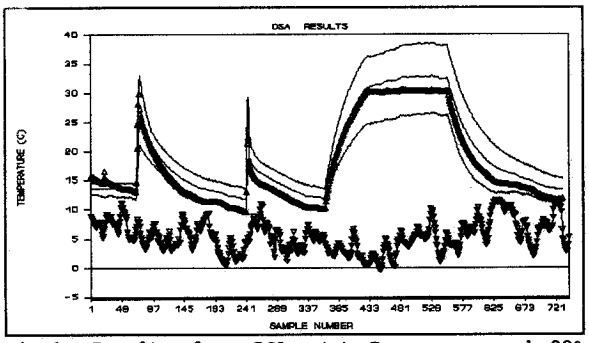

Fig.1. Results from DSA. (-) Base case and 99% uncertainty bands. ( ) Measurements. ( ) Outdoor air temperature.

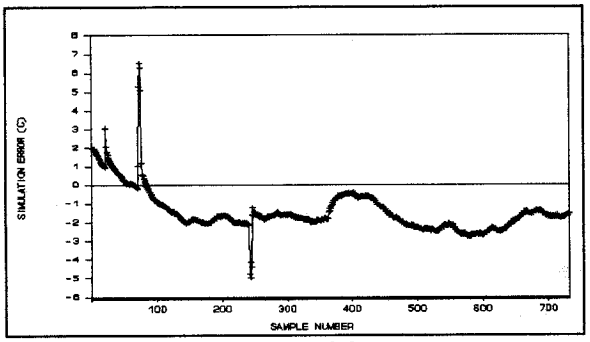

Fig.2. Residuals from ESP model.

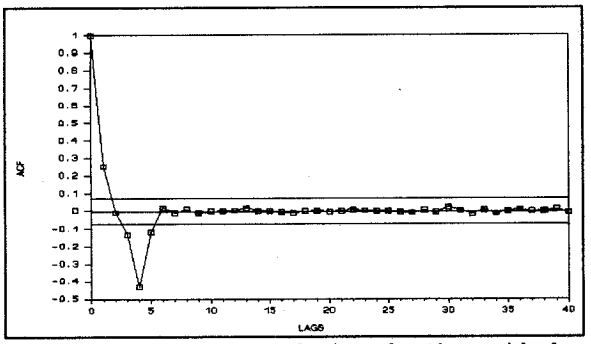

Fig.3. Estimated autocorrelations for the residuals.

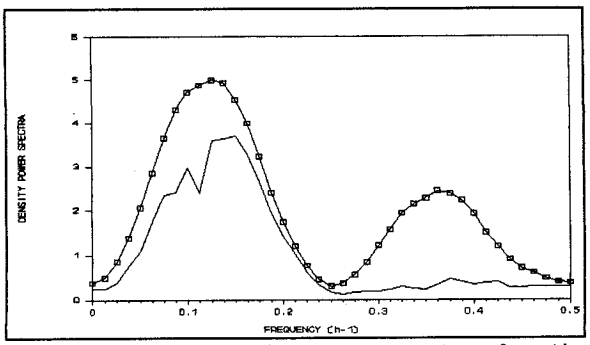

Fig.4. Estimated density power spectra for the residuals and for th noise process.

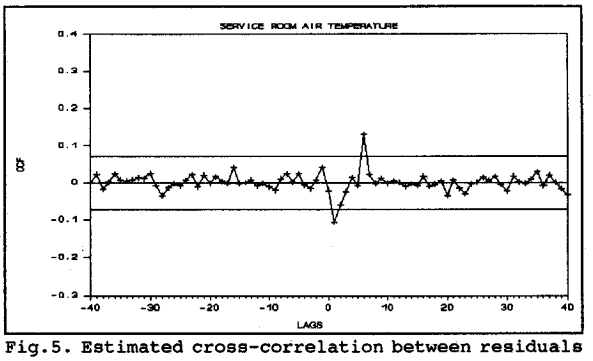

Fig.5. Estimated cross-correlation between residuals and service room air temperature.

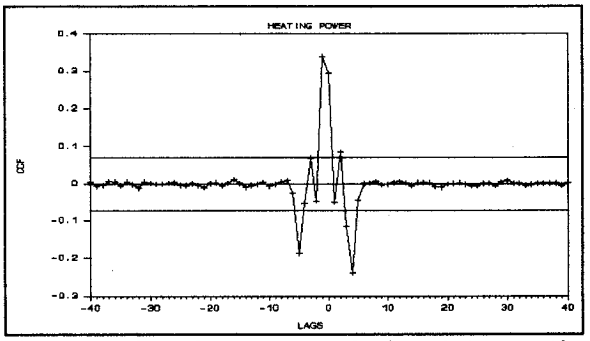

Fig.6. Estimated croos-correlation between the residuals and the power heating.

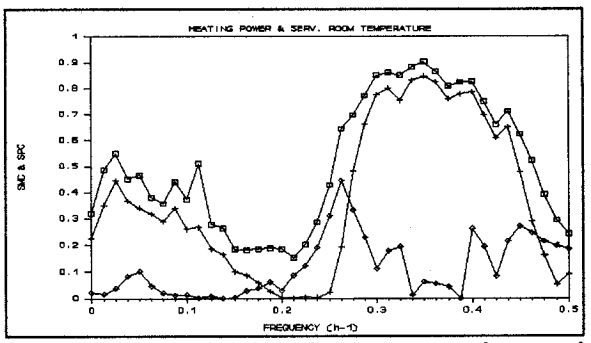

Fig.7. Squared multiple coherency ( ) and squared partial coherency for the power heating ( ) and for the service room air temperature ( ).

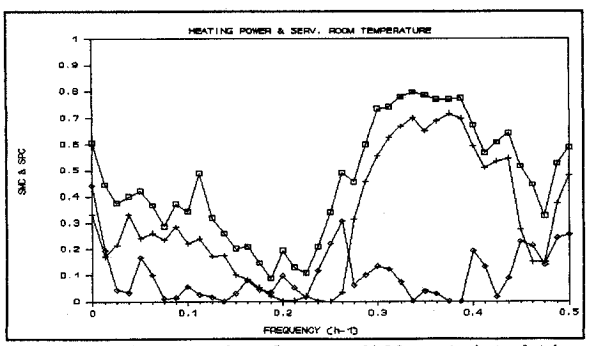

Fig.8. As figure 7, but for non-differentaiated time series.

most relevant input to explain part of the residuals at intermediate and high frequencies (fig.7) is the power heating, and, in a second position, the service room air temperature. ii) At very low frequencies (fig.8), not only these inputs present a non-irrelevant squared partial coherency, but also the outdoor air temperature ($k^2 \approx 0.3$) and the solar irradiance ($k^2 \approx 0.2$).

Conclusions

Under the imposed experimental conditions, the model, with the set of selected values for its different parameters, does not describe adequately the stationary nor the dynamic behaviour of the test cell.

Main problems detected at intermediate and high frequencies are related to the power heating. May be hypothesis about convective coefficients, air stratification, optical properties of the indoor surfaces or about the convective/radiative split of the heat input must be revised.

At low frequencies, the detected problems are not only related to the above mentioned quantities, but also to the outdoor air temperature and to the solar irradiance. Modifications on parameters conecting the two cell's zones, on the hypothesis about infiltrations and on the hypothesis of unidimensional heat conduction trough walls - edge effects are very important in a test cell - could be necessary.

The decorrelation detected between residuals and other inputs different than the power heating or the service room air temperature, does not mean that the corresponding physical processes or parameters linked to them are well represented in the model. Note that the main forcing functions for the test cell during the experiment were the power heating and the service room temperature - test cell with calibration wall. Information about the adequacy of the ESP when representing physical processes involving solar irradiance, outdoor air temperature, relative humidity, wind velocity or wind direction will required another kind of commponent or experiment.

SUMMURY AND CONCLUSIONS

Several statistical tools have been described and its use for empirical whole model validation discussed. Main conclusion of this paper could be:

i) Sensitivity analysis is a useful tool to guide experiments, but it is not very efficient when trying to detect model inadequacy. Usually uncertainty bands are too large to draw any firm conclusions about the model's validity. On the other hand, it cannot by itself show where errors are ocurring in the model. So there is a need of another kind of validation techniques.

ii) Residual analysis appears as a promising alternative. It gives useful information not only to determine how the model relates to reality, but also about what could be the causes of a deficient model behaviour.

Model's validity to describe the stationary regime of a system could be analyzed through the mean value of the residuals.

Variance is a second measure of the model's ability to described the observed data. How the variance is distributed over frequency can be analyzed using the estimated spectrum of the residuals. Dynamic parts of the model could be checked from this kind of analysis.

What proportion of the residual's spectrum can be predicted from the explanatory variables of the model is estimated by the squared multiple coherency. The effect of unrecognized and undescribed input signals, noise corrupting inputs and measurement errors remain in the called noise spectrum.

Where errors are ocurring in the model can be investigated using the estimated squared partial coherencies, which show what inputs can recover part of the spectrum of the model. Hence, an approach to to knowledge of what physical process or input parameters are no adequately represented in the model is achieved.

REFERENCES

BOOK:
Box, G.E.P. and J.M. Jenkins. 1976. Time series analysis: Forecasting and control. Holden-Day, San Francisco, California.

BOOK:
Clarke, J.A. 1985. Energy Simulation in Building Design. Adam Hilger Ltd, Bristol and Boston.

BOOK:
Jenkins, G.M. and D.G. Watts. 1968. Spectral analysis and its applications. Holden-Day Inc., San Francisco, California.

SPECIAL PUBLICATION:
Pinney, A. and P. Strachan. 1989. Whole Model Validation. Chapter 15 of The PASSYS Project Phase 1. Subgroup Model Validation snd Development - Final Report. 1986-1989. Commision of the European Communities, DGXII.

IN A CONFERENCE PROCEEDINGS:
Ostergaard Jensen, S. and R. van de Perre. 1991. Tools for Whole Model Validation of Building Simulation Programs. Experience from the CEC Concertated Action PASSYS. Accepted in the IBPSA - Building Simulation '91. (Sophia Antipolis, Nice, FR, August 20-22)

RESEARCH REPORT:
Palomo, E. and F.M. Téllez. 1991-c. Non-parametric tests on the residuals for model validation purposes. (PASSYS Ref. xxx-91-PASSYS-MVD-WD-zzz)

BOOK:
Priesley, M.B. 1981. Spectral analysis and time series. Academic Press.

BOOK:
Vandaele, W. 1983. Applied time series and Box-Jenkins models. Academic Press, Inc.