NILU
TEKNISK NOTAT 8/79
REFERANSE: 10003/79
DATO: APRIL 1979

IDENTIFICATION OF MODELS FOR SOME TIME SERIES OF ATMOSPHERIC ORIGIN WITH AKAIKE'S INFORMATION CRITERION

AV

KARL J. EIDSVIK

NORWEGIAN INSTITUTE FOR AIR RESEARCH
P.O. BOX 130, 2001 LILLESTRØM
NORWAY

ISBN- 82-7247-108-6

SUMMARY

Akaike's method for model identification has been used to identify "Markov" chain models for simple transformations of daily precipitation at three locations in southeast Norway and wind force and wave height at one location in the Norwegian Sea. Attempts at identification of the horizontal wind vector as an autoregressive process have also been made. The estimated order of a model appears to increase with the sample size. It may also have a significant uncertainty. The analytical complexity of identified models may appear to be unnecessarily large for some purposes.

TABLE OF CONTENTS

IDENTIFICATION OF MODELS FOR SOME TIME SERIES OF ATMOSPHERIC ORIGIN WITH AKAIKE'S INFORMATION CRITERION

1 INTRODUCTION

The most subjective, time consuming and difficult aspect in the analysis of stochastic time series is usually the identification of a convenient model. The real, unknown statistical properties of a time series may be extremely complicated and beyond our reach. The identification must then be restricted to a search for a parsimonious model that is sufficiently complete compared to the purpose at hand and the available information. Different models of the same time series may even turn out to be useful according to the purpose of the analysis. The problems of model identification are illustrated in, for instance, the classical book of Box and Jenkins (1).

Akaike (2) has suggested an objective method of model identification. He argues that the Kullback-Leibler mean information (3) establishes a reasonable cost function that allow an efficient search of a model for general application. The only information about the time series used for identification is the available data. Akaike's method has given reasonable results in some applications. We will apply it for the purpose of an exploratory identification of time series of atmospheric origin.

2 THEORY

Since Akaike's identification method involves a measure of quality, its justification must be based on usefulness in applications rather than on rigorous theoretical considerations.

We will therefore only give a short, lighthearted outline of the theory, which is more fully described by Akaike (2,4,5,6) and Tong (7).

2.1 Akaike's information criterion

Let the true, unknown distribution of the stationary vector u be g(u). One of the parametric families suggested to approximate this distribution is $f(u/\theta)$, where θ is the parameter vector. Akaike argues that a most sensitive criterion for discriminating between deviations of $f(u/\theta)$ from g(u) is the Kullback-Leibler mean information (3) (see also Shannon and Weaver (8)).

$$I\ (g;\ f(\cdot/\theta)) = \int g(u) \ln g(u)\ du - \int g(u)\ \ln f(u/\theta) du \qquad (2.1)$$

As both g and f are probability distribution functions, $I(g;f(\cdot/\theta)) \geq 0$, with equality only if $f(u/\theta) = g(u)$. A simplification of equation (2.1) is obtained by assuming that $g(u) = f(u/{}_M\theta)$ and that $\Delta\theta = \theta - {}_M\theta$ is close to zero. A Taylor expansion of $\ln f(u/\theta)$ in equation (2.1) then gives

$$\begin{aligned} I({}_M\theta;\theta) &\simeq \tfrac{1}{2}\Delta\theta_i \Delta\theta_j \int g(u)\ \frac{\partial \ln f}{\partial \theta_i}\ \frac{\partial \ln f}{\partial \theta_j}\ du \\ &= \tfrac{1}{2}\Delta\theta_i \Delta\theta_j\ J_{ij} \\ &= \tfrac{1}{2}||\ \Delta\theta|| \end{aligned} \qquad (2.2)$$

The estimate of ${}_M\theta$, ${}_k\hat{\theta}$, must be restricted to a given number, k, of independent parameters. That is, the components of ${}_k\hat{\theta}$ span a space R^k of smaller dimension than R^M, the space spanned by the components of ${}_M\theta$. When ${}_k\theta = E_k\hat{\theta}$ (the expected value of ${}_k\hat{\theta}$) and

$$\Delta\theta = {}_k\hat{\theta} - {}_M\theta = ({}_k\hat{\theta} - {}_k\theta) + ({}_k\theta - {}_M\theta) \qquad (2.3)$$

are introduced into equation (2.2) we have

$$2I({}_M\theta;\ {}_k\hat{\theta}) = ||\ {}_k\theta - {}_M\theta|| + ||{}_k\hat{\theta} - {}_k\theta|| + \text{const}\ ({}_k\hat{\theta} - {}_k\theta) \qquad (2.4)$$

The mean of the last term vanish. As $\sqrt{N}({}_k\hat{\theta}-{}_k\theta)$ is asymptotically Gaussian with zero mean and variance matrix J^{-1} (2), the quadratic form $N||{}_k\hat{\theta}-{}_k\theta||$ is chi-squared distributed with k degrees of freedom. Equation (2.4) then gives

$$2N\cdot E\{I({}_M\theta;\ {}_k\hat{\theta})\} \simeq N||{}_k\theta-{}_M\theta|| + k \tag{2.5}$$

When $I({}_M\theta;{}_k\hat{\theta})$ is adopted as the risk function in the model building, the associated cost function (expected risk) is given by equation (2.5). It remains to <u>estimate</u> the cost, or more precisely, to estimate to estimate

$$||{}_k\theta-{}_M\theta|| \simeq -2\int f(u/{}_k\theta)\ \ln\frac{f(u/{}_k\theta)}{f(u/{}_M\theta)}\ du \tag{2.8}$$

It may be shown (4,7) that a likelihood ratio statistics, which for independent observations reads,

$${}_k\eta_M = -2\sum_{i=1}^{N}\ \ln\ \frac{f(u_i/{}_k\hat{\theta})}{f(u_i/{}_M\hat{\theta})} \tag{2.9}$$

is asymptotically noncentral chi-squared distributed both for independent and dependent observations. The noncentrallity parameter is $N||{}_k\theta-{}_M\theta||$ and the degree of freedom is (M-k). An unbiased estimate for the cost function $2E\{I({}_M\theta;{}_k\hat{\theta})\}$ is therefore

$$AIC(k) = N^{-1}\left[{}_k\eta_M + 2k - M\right] \tag{2.10a}$$

The central variance of AIC(k) is proportional to that of ${}_k\eta_M$

$$\text{Var}\ ({}_k\eta_M) = 4N||{}_k\theta-{}_M\theta|| + 2(M-k) \tag{2.11}$$

which may be large when $||{}_k\theta-{}_M\theta|| \neq 0$ and $N\to\infty$

The statistics (2.10a) is the estimate of the cost function suggested by Akaike. The constant, unknown M must be chosen "reasonably" large. With fixed M, some terms of equation (2.10a) become constant, so that various versions of AIC are used.

$$AIC(k) \propto {}_k\eta_M - 2(M - k) \qquad (2.10b)$$

$$AIC(k) \propto -2 \ln (\text{max likelihood}) + 2k \qquad (2.10c)$$

As the number of independent parameters, k, in a model increases, the fit to the data increases and ${}_k\eta_M$ decreases. However, the uncertainty of the model, characterized by 2k-M, will then increase. The best approximating model is the one which achieves the most satisfactory compromise between fit and uncertainty, i.e., the minimum AIC model.

The above measure should be applicable to model identification quite generally. However, as the likelihood or likelihood ratio must be computed, the search for model is restricted to a parametric family.

2.2 Markov chain

Suppose that $u(t) = \{1,2,...s\}$ is a scalar variable in discrete time and discrete, finite state space. It is assumed that u(t) is a chain of order p:

$$P\{u(t)/....., u(t-1)\} = P\{u(t)/u(t-p),..,u(t-1)\} \qquad (2.12)$$

It is convenient to call this a "Markov" chain even if p>1. The identification consists in deciding on the value of p. Bartlett (9), Hoel (10) and Good (11) have designed statistical tests for this decision. Tong (7) has derived the AIC for this case.

Under the hypothesis H_r: the chain is r-dependent, the likelihood function is:

$$L(r) \propto \prod_{t=1}^{N-r} P\{u(t+r) = l/u(t) = i, \ldots u(t+r-1) = k\}$$

$$\propto \prod_{t=1}^{N-r} P_{ij..kl} \qquad (2.13)$$

The observed number of transitions over the states $i \to j \ldots k \to l$ is $n_{ij..kl}$. That is

$$L(r) \propto \prod_{i,j..k,l} P_{ij..kl}^{n_{ij..kl}} \qquad (2.14)$$

Maximum likelihood estimates for the parameters of the process, $P_{ij..kl}$, are

$$\hat{P}_{ij..kl} = \frac{n_{ij..kl}}{n_{ij..k}} \quad ; \quad n_{ij..k} = \sum_l n_{ij..kl} \qquad (2.15)$$

The maximum likelihood function is thus

$$L(r,\hat{P}) \propto \prod_{i,j..k,l} \hat{P}_{ij..kl}^{n_{ij..kl}} \qquad (2.16)$$

Under the hypothesis H_{r-1}: the chain is (r-1)-dependent, corresponding expressions are obtained. With

$$\hat{P}'_{ij..kl} = \hat{P}_{j..kl} \qquad (2.17)$$

the likelihood ratio for testing H_{r-1} within H_r is thus

$$r\text{-}1\lambda_r = \frac{L(r-1,\hat{P})}{L(r,\hat{P})} = \prod_{ij..kl} \left(\frac{\hat{P}'_{ij..kl}}{\hat{P}_{ij..kl}} \right)^{n_{ij..kl}} \qquad (2.18)$$

By repeated use of equation (2.18) it is observed that

$$ {}_p\lambda_M = {}_p\lambda_{p+1} \cdot {}_{p+1}\lambda_{p+2} \cdots {}_{M-1}\lambda_M \qquad (2.19) $$

$$ {}_p\eta_M = -2 \ln {}_p\lambda_{p+1} - 2 \ln {}_{p+1}\lambda_{p+2} - \cdots - 2 \ln {}_{M-1}\lambda_M \qquad (2.20) $$

which may be computed from the frequency counts $n_{ij..kl}$. It is known (10) that ${}_p\eta_M$ is, under H_p, asymptotically chi-squared distributed with degrees of freedom $\nu=(s-1)(s^M-s^P)$. The AIC measure is then obtained from equation 2.10b as

$$ AIC(p) = {}_p\eta_M - 2(s-1)(s^M-s^P) \qquad (2.21) $$

In order to judge how well defined the minimum of AIC is, we consider a measure for the random variation of neighbouring AIC-s. From equations (2.20) and (2.21) we have

$$ AIC(p-1) - AIC(p) = {}_{p-1}\eta_p - 2s^{p-1}(s-1)^2 \qquad (2.22) $$

Under the p-1 hypothesis, ${}_{p-1}\eta_p$ is a chi-squared variable with $s^{p-1}(s-1)^2$ degrees of freedom (9). The standard deviation of AIC(p-1)-AIC(p), taken as the approximate measure of random variations, is then

$$ \{2s^{p-1}(s-1)^2\}^{\frac{1}{2}} \qquad (2.23) $$

which is usually much smaller than $\sqrt{2\nu}$, a result of the dependence between neighbouring AIC-s.

Although it may be illogical to use classic hypothesis testing when none of the suggested hypothesis are probably correct (2), it is comforting if an identified model passes commonly used tests at a reasonable significance level. In terms of the likelihood ratio test, H_r is rejected within H_M at significance level α if ${}_p\eta_M > y_*(\alpha)$. As $\nu \geq 30$, the chi-squared distribution of ${}_p\eta_M$ approaches a Gaussian with mean ν and standard deviation $\sqrt{2\nu}$.

That is:

$$\frac{p^{\eta_M}-\nu}{\sqrt{2\nu}} \to n(0,1) \text{ as } \nu \geq 30 \qquad (2.24a)$$

so that H_p is accepted at α if

$$p^{\eta_M} - (\nu + y_*(\alpha)(\sqrt{2\nu}) < 1 \text{ for } \nu \geq 30 \qquad (2.24b)$$

Here $y_*(\alpha) \in \{\frac{1}{2}, 2\}$ for reasonable choices for α.

The Markov chain model is convenient when it is possible to divide the state space into a few classes. Few assumptions are necessary. However, as the size of the state space increases above two, the number of parameters increases tremendously with increasing order.

2.3 Autoregressive process

A more effective choice of parameters may be possible if u(t) is assumed to be a continuous, nearly Gaussian vector of dimension d in descrete time. It is assumed that u(t) is a linear autoregressive process

$$\sum_{i}^{p} A(i)\, u(t-i) = a(t)$$

$$A_{ij}(n)\, u_j(t-n) = a_i(t) \qquad (2.25)$$

Here A(i) are dxd matrixes with A(0) = I; a(t) is the d-dimensional, white, one step ahead prediction error; E a(t) = 0, $E\{a(t)\, a'(t)\} = G$. The number of independent A(i)-components and G-components is $d^2 \cdot p$ and d^2, respectively. The likelihood function is

$$L(a;\, {}_k\theta) = ((2\Pi)^d |G|)^{-\frac{1}{2}N} \exp\{-\tfrac{1}{2} \sum_{t=1}^{N} a'(t) G^{-1} a(t)\} \qquad (2.26)$$

which gives

$$-2\ln L(a;{}_k\theta) \propto N \ln |G| + \sum_{t=1}^{N} a'(t)\, G^{-1} a(t) \qquad (2.27)$$

Akaike (12) shows the maximum likelihood estimate of G to be

$$\hat{G} = \frac{1}{N} \sum_{t=1}^{N} a(t)\, a'(t) \qquad (2.28)$$

At this value of G the -2 log likelihood function becomes

$$-2\ln L(a;{}_k\hat{\theta}) \propto N \ln |\hat{G}| \qquad (2.29)$$

The AIC measure (2.10c) may now be written

$$\begin{aligned} AIC(p) &= N \ln |\hat{G}| + 2\,(p+1)d^2 \\ &\propto N \ln |\hat{G}| + 2\,p \cdot d^2 \end{aligned} \qquad (2.30)$$

This identified model is the same as the one identified with the minimum one step ahead prediction error (2)

$$FPE(k) \simeq |\hat{G}|\left(1 + \frac{2k}{N}\right) \qquad (2.31)$$

$$\begin{aligned} \ln FPE(k) &\simeq \ln |\hat{G}| + \frac{2k}{N} \\ &\simeq AIC(k) \cdot N^{-1} \end{aligned}$$

When $|\hat{G}|$ is unbiased there is consequently a tendency for FPE(k) and AIC(k) to decrease with N. The decrease is highest for the largest k, so that there will be a tendency to identify a higher order model when N increases.

It is necessary to estimate the coefficients A, before G (and a) can be estimated. For this purpose Akaike (13) uses the Yul-Walker equations, derived by minimizing the error:

$$E\, a_i^2 = A_{ij}(n) A_{ik}(m)\; E\{u_j(t-n) u_k(t-m)\}$$

$$= A_{ij}(n)\; A_{ik}(m)\; Q_{jk}(n-m) \qquad (2.32)$$

The coefficients that minimize this error are found by differenciating with respect to the coefficients $A_{rs}(t)$; $r,s,t > 0$:

$$\frac{\partial E a_i^2}{\partial A_{rs}(t)} = Q_{jk}(n-m)\; \{ \frac{\partial A_{ij}(n)}{\partial A_{rs}(t)}\; A_{ik}(m) + A_{ij}(n)\; \frac{\partial A_{ik}(m)}{\partial A_{rs}(t)} \}$$

$$= Q_{jk}(n-m)\; \{ \delta_{ir} \cdot \delta_{js} \cdot \delta_{nt}\; A_{ik}(m) + \delta_{ir} \cdot \delta_{ks} \cdot \delta_{mt}\; A_{ij}(n) \}$$

$$= 2\, A_{rj}(n)\; Q_{js}(n-t)$$

$$= 2 \sum_{n=0}^{p} A(n)\; Q(n-t) \qquad (2.33)$$

At the minimum the derivatives are zero so that

$$\sum_{n=0}^{p} A(n)\; Q(n-t) = 0 \qquad \text{for } t = 1,2 \ldots p \qquad (2.34)$$

The matrix G is found from

$$G_{ij} \simeq E a_i a_j = A_{ik}(n)\; A_{jp}(m)\; Q_{kp}(n-m)$$

$$= A_{ik}(t)\; A_{jp}(n)\; Q_{pk}(n-t)$$

$$= A_{ik}(t) \sum_{n=0}^{p} A(n)\; Q(n-t) \qquad (2.35)$$

By equation (2.33) the last sum is zero unless t=0 so that

$$G_{ij} = \delta_{ik} A_{jp}(n) Q_{pk}(n)$$

$$= A_{jp}(n) Q_{pi}(n) \qquad (2.36)$$

With the covariance matrix estimated from the data

$$\hat{Q}_{ij}(m) = \frac{1}{N} \sum_{t=1}^{N-m} u_i(t) u_j(t+m) \qquad (2.37)$$

equation (2.34) is used to estimate A; $\hat{G}_{ij}$ is then obtained from equation (2.36).

As illustrated by, for instance Eidsvik (14,15), the Yul-Walker equations may be ill-conditioned. That is, the solution, A(i), varies considerably due to small variations of the estimated covariance.

3 APPLICATIONS

Akaike's theory of model identification is applicable for a large sample and a stationary process. For geophysical time series, containing seasonal and possibly climatic variations, these are conflicting requirements. Subjective judgement is needed in order to reach a compromise so that the sample size is maximized and effects from seasonal variations are minimized.

3.1 Daily precipitation, Southeast Norway

3.1.1 Occurrence and nonoccurrence

The occurrence and nonoccurrence of daily precipitation is a process that has traditionally been discussed in terms of Markov chain models (Gabriel and Neuman (16), Nordø (17), Katz (18), Gates and Tong (19) and Chin (20)). The observation period that has been available for these studies is less than

30 years of data. Available to us are data from ca 80 years at the stations Hedrum and Nordodal in SE Norway, and from ca 60 years at Røldal in SW Norway. When using Akaike's identification method, a long time series is valuable because the theory is only asymptotically valid, and we do not know how a "small sample estimate" converges toward the "true" value.

To avoid seasonal variations as much as reasonable, models are identified for each month of the year. Figure 3.1 shows histograms of the estimated order at the three stations. There was apparently no systematic variation of the order with the time of the year. The differences between the histograms are probably not significant so that the order is identified as 2 or 3 for all months and all stations.

To obtain an impression of the randomness of the AIC estimates, the likelihood ratio statistic, ${}_p\eta_M$, for each month and station, is plotted in Figure 3.2. Asymptotically ${}_p\eta_M$ is $\chi^2(0,\nu)$ when H_p is true. If it is assumed that all chains are of the same order, and that the sample size is large enough, the scatter at the accepted p-value in Figure 3.2 should be approximately given by a $\chi^2(0,\nu)$ distribution. The data indicate that the sample mean value is significantly higher than ν, especially for small p. Therefore the chains are probably not of the same order, or the sample size is not large enough. Yet, if the limiting distribution and classical hypothesis tests had been used, many of the tested models would have been accepted at a high significance level. Approximately half of the identified models (order 2 or 3) would be accepted at a higher significance level than 15%.

The only way to increase the sample size is to include more than one month in the analysis. This is done at the risk of introducing significant seasonal variations. To minimize this effect, the attention is restricted to 2 or 3 winter months and 2 or 3 summer months only. Figure 3.3 shows the estimated order for different definition of the summer and winter season.

The estimated order shows a tendency to increase with the length of the interval defining the season. The reason is either seasonal variations and/or a methodological tendency to underestimate the order when the sample size is small.

Figure 3.4 illustrates the variation of AIC with the number of years used for analysis. The last 30 years tend to give much the same AIC curves as the largest samples size. However, there also seems to be a tendency for a lower order model to be identified in the 30 year sample. The last 10 years sample may give significantly different AIC curves. As the AIC minima for the 10 years data are low, use of these data would indicated that chains of the order 1 or 2 should be accepted at a high significance level. A change of model during the last 30 years suggests a climatic change during this time interval, which is considered to be unlikely. The above results do therefore indicate that a large sample size is needed to give stable estimate of the order of the Markov chain for the occurrence and nonoccurrence of daily precipitation. With 25 years of precipitation data at many localities, Chin (20) found the estimated order to be relatively stable after approximately 10 years of data. Our estimates indicate that not even 30 years of data may be enough. Figure 3.3 also indicates that the minima of AIC does not tend to be well defined compared to the measure (2.18) so that the estimated order has a significant uncertainly. It follows that the estimated order is probably not suited as a variable to describe nonstationary or nonhomogeneous effects as discussed in (19) and (20).

Although the statistical properties (transition probabilities) of the identified chains are of interest, a discussion of this is not considered to be a topic for this report. However, the estimated transition probability matrixes are shown in Table A.1 of the Appendix. It is also noticed that the occurrence and nonoccurrence of daily precipitation may sometimes be predicted with remarkable accuracy with only information on the precipitation history of the last few days.

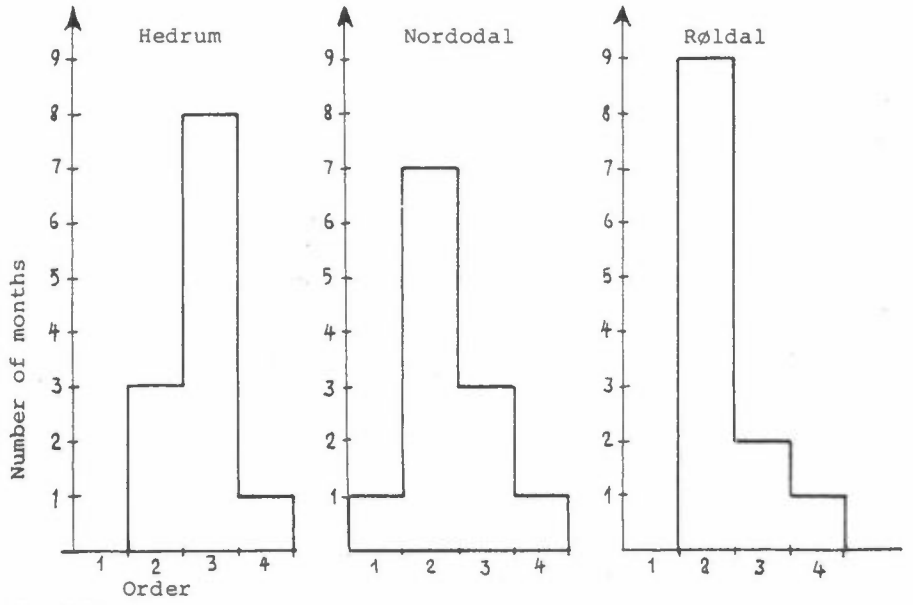

Figure 3.1: Histograms of identified order for each month at the tree stations. Occurrence and nonoccurrence of daily precipitation.

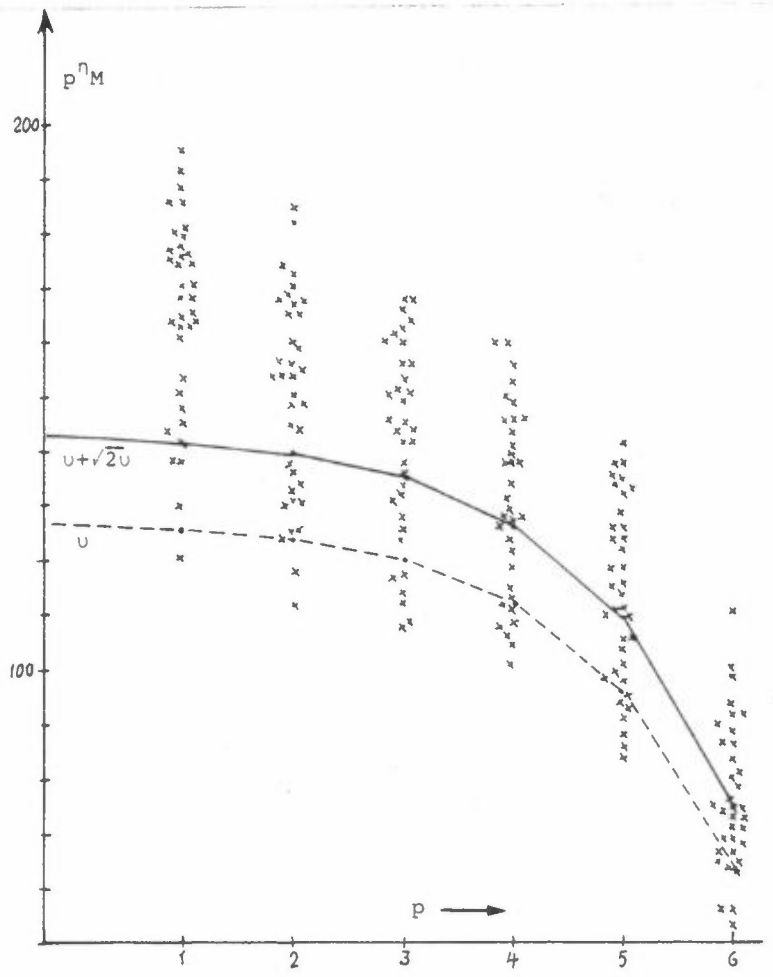

Figure 3.2: Likelihood ratio statistics. Occurrence and non-occurrence of daily precipitation. Each month and station.

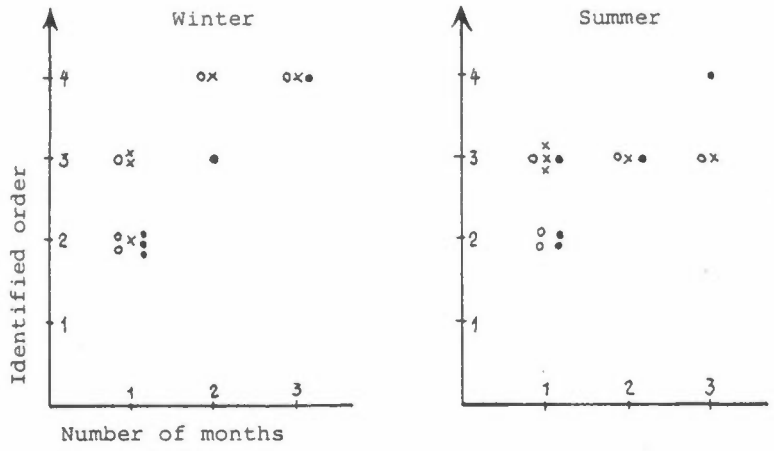

Figure 3.3: Identified order for different definition of the season. Occurrence and nonoccurrence of daily precipitation. x: Hedrum, o: Nordodal, ●: Røldal.

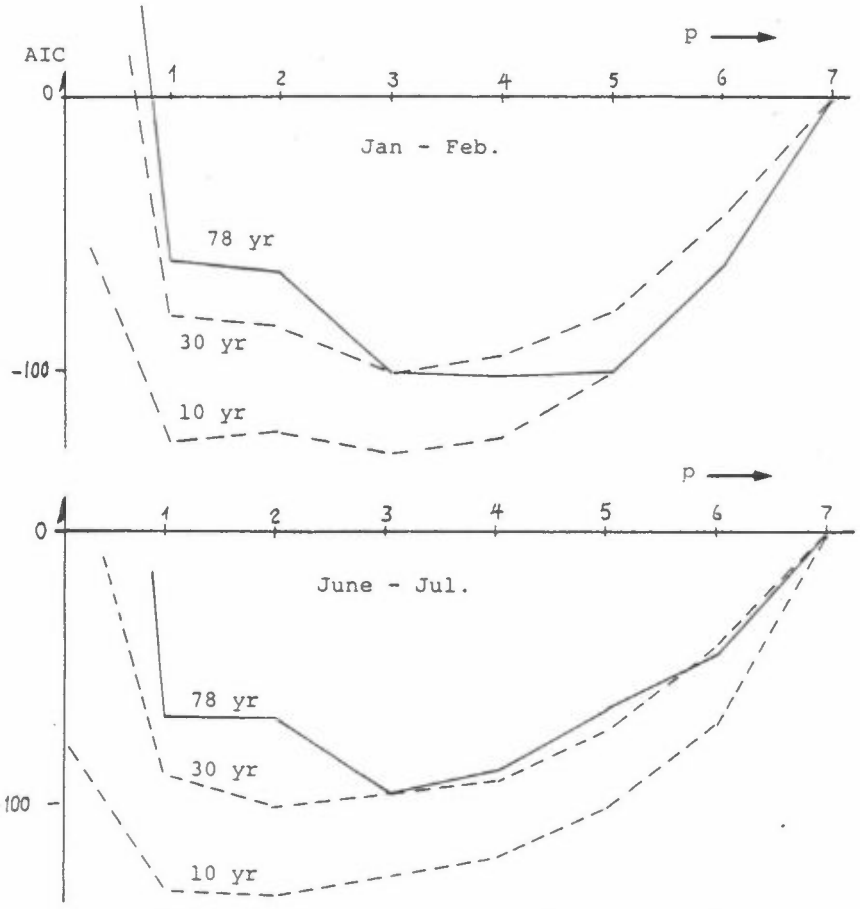

Figure 3.4a: Sample size variations of the AIC-estimates. Occurrence and nonoccurrence of daily precipitation. Hedrum

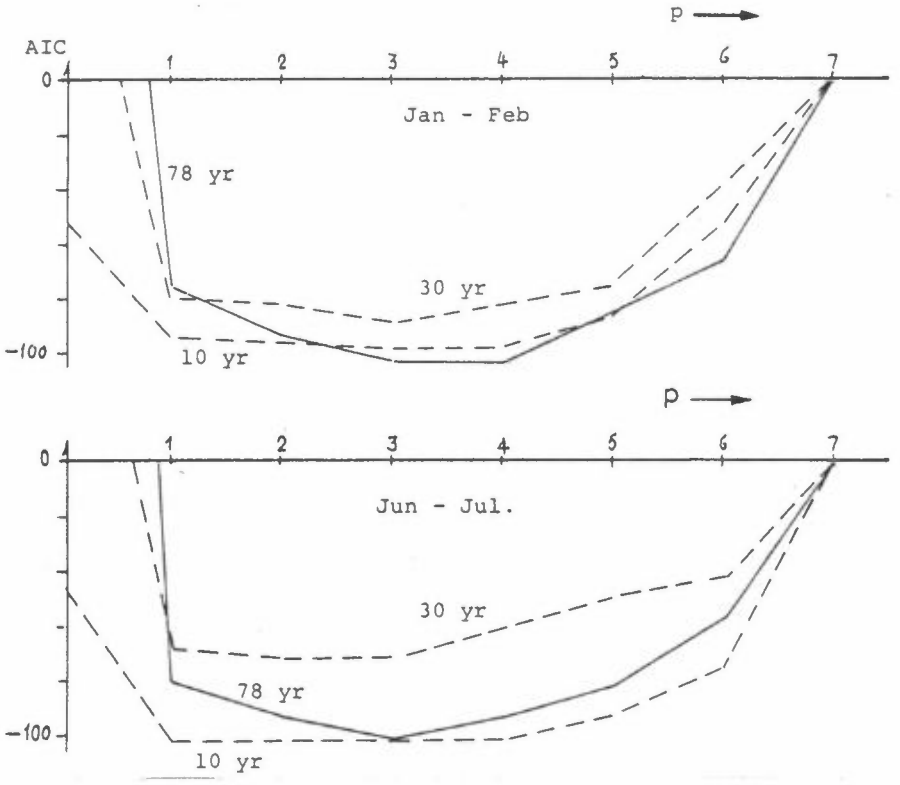

Figure 3.4b: Sample size variations of the AIC-estimates. Occurrence and nonoccurrence of daily precipitation. Nordodal.

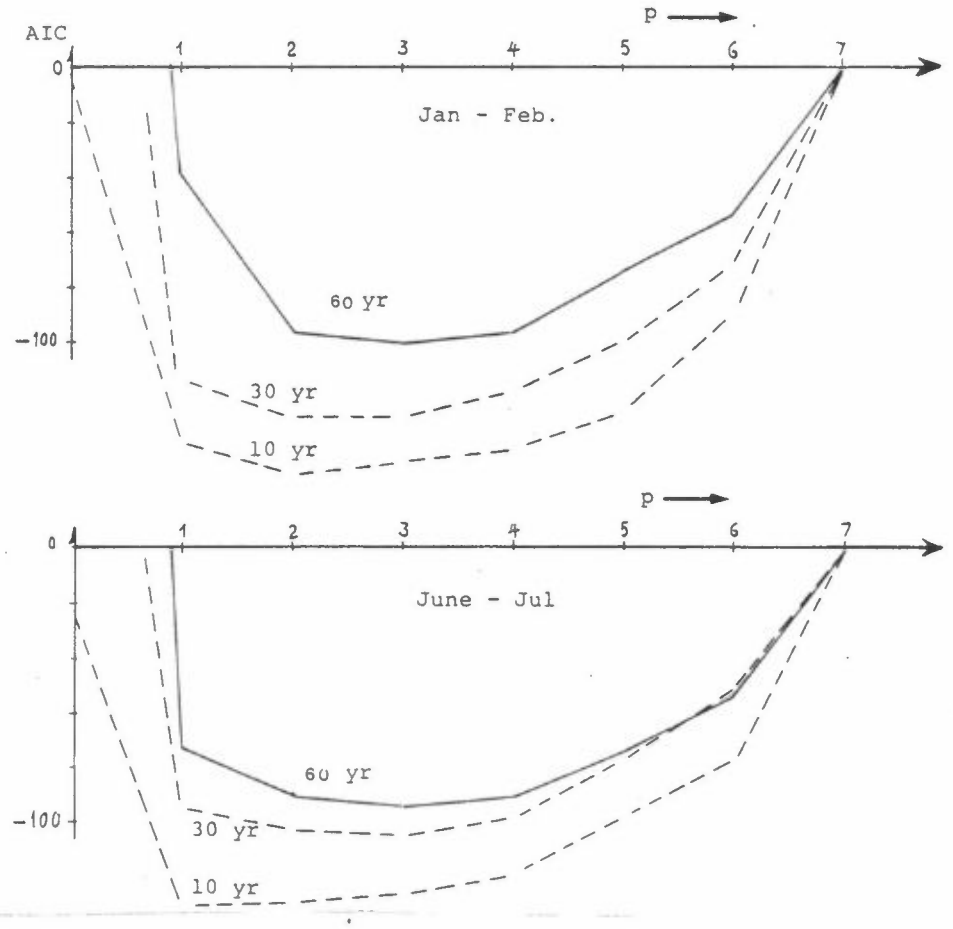

Figure 3.4c: Sample size variations of the AIC-estimates. Occurrence and nonoccurrence of daily precipitation. Røldal.

3.1.2 Nonoccurring, precipitation less and more than conditional mean

To obtain a rough quantitative modelling of the amount of daily precipitation, the state space is increased to s = 3 such that u = 1 is "no precipitation", u = 2 is "precipitation less than" -, and u = 3 is "precipitation larger than" the estimated conditional mean precipitation given the occurrence of precipitation.

Figure 3.5a illustrates the variation of the identification statistics for the winter season. There is a fairly well defined minimum of AIC at p = 2 for all stations. However, only at one of the stations is this model accepted at a reasonable high significance level. At the other stations no model is accepted at a reasonable level. Figure 3.5b illustrates the results for the summer season. For the two stations, where a low order Markow chain is most easily accepted in terms of significance level, the minimum of AIC is located at 1, 2 or 3.

With use of only the 30 last years of data, 1 or 2 order models were accepted at a large significance level at all stations. The large difference between the estimates of $_k\eta_M$ based on 30 years and ca 80 years of data could indicate that $_k\eta_M$ has not reached its asymptotic distribution with 30 years of data.

The estimated transision probability matrixes are presented in Table 3.2. It is indicated that the amount of daily precipitation can not usually be accurately predicted by information on the history of the precipitation only.

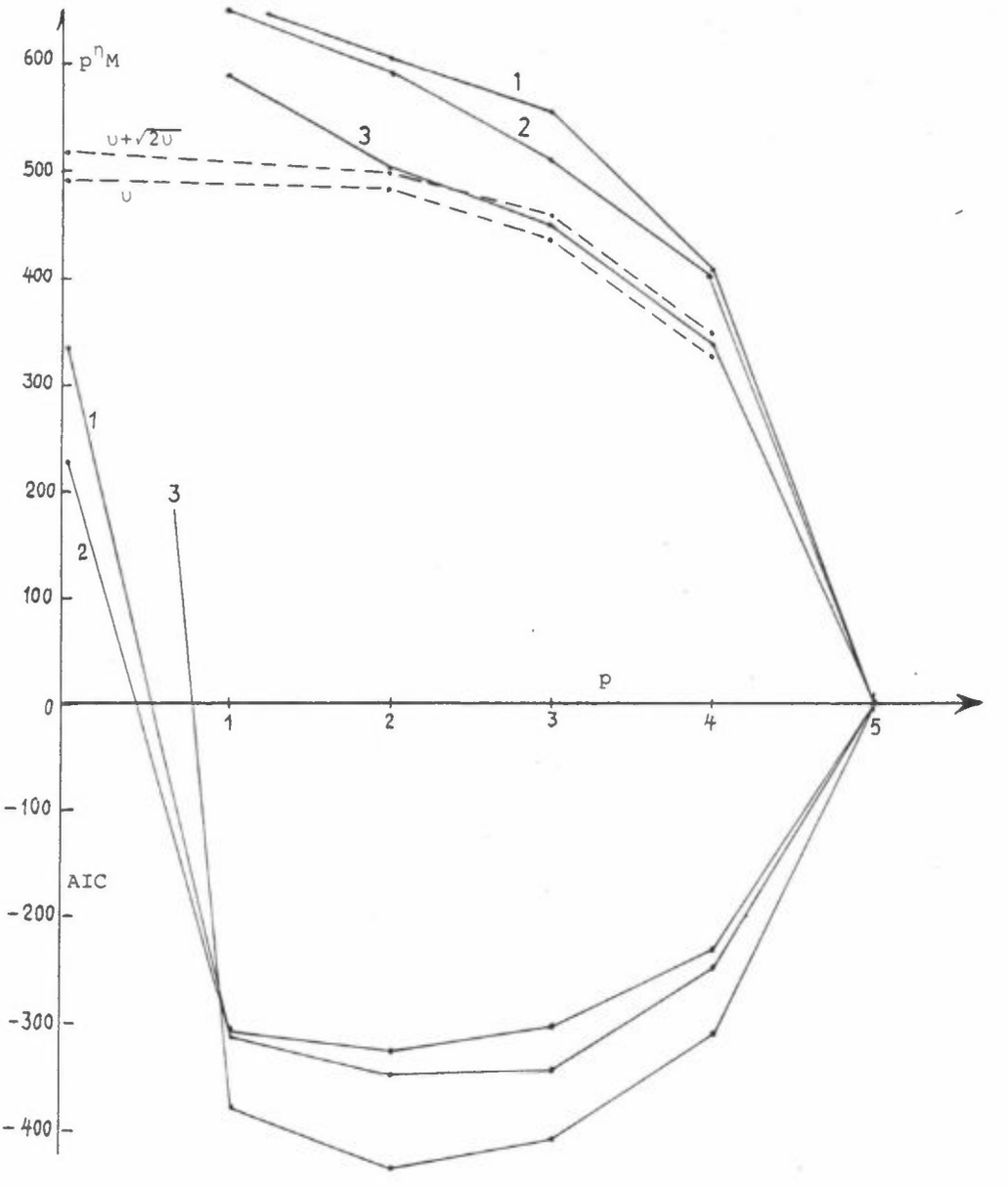

Figure 3.5a: Identification variables $_p\eta_M$ and AIC (p), January and February. Nonoccurrent, precipitation less and more than conditional mean.
1: Hedrum, 2: Nordodal, 3: Røldal.

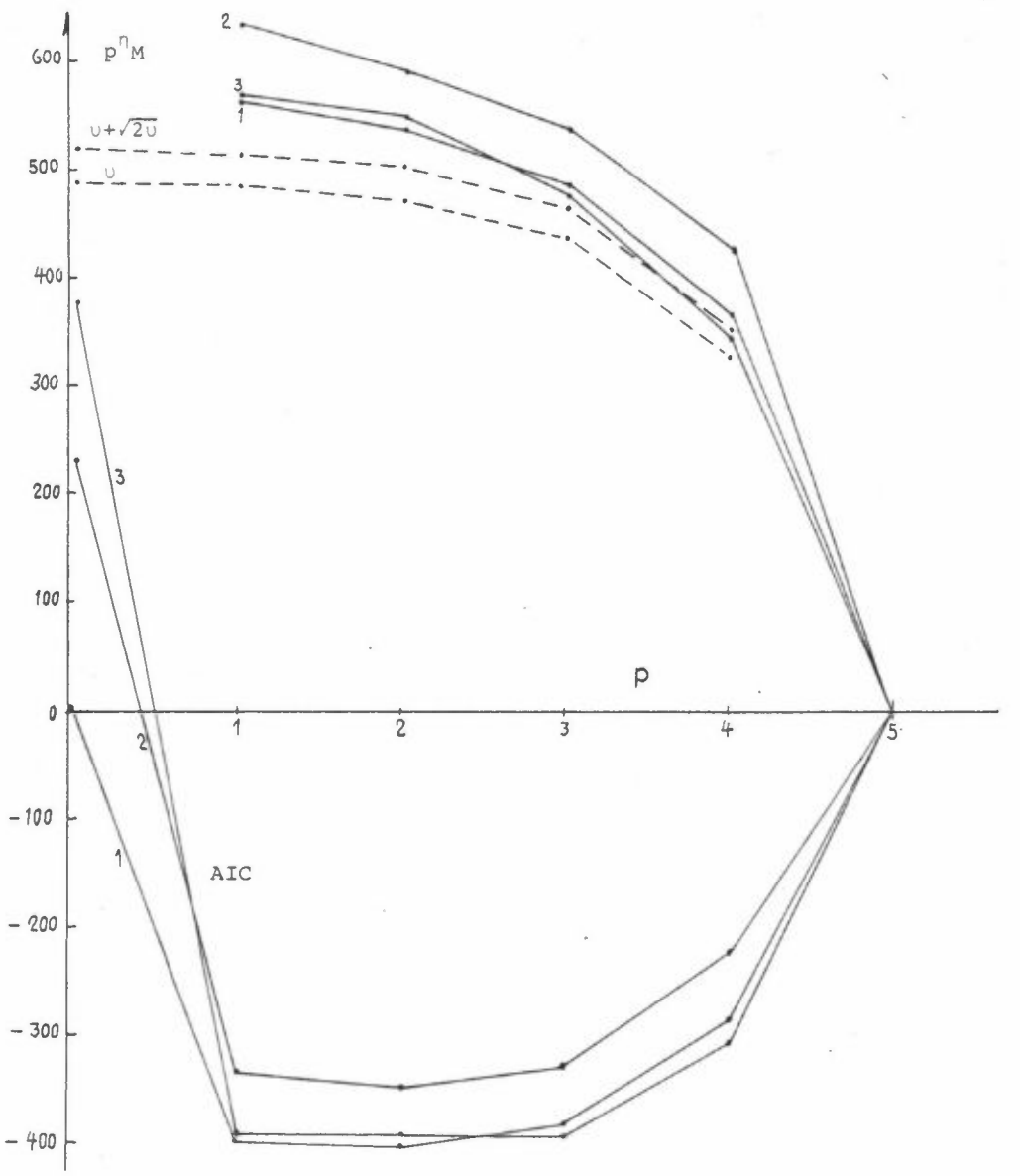

Figure 3.5b: Identification variables $_{p}{}^{\eta}{}_{M}$ and AIC (p). June and July. Nonoccurrent, precipitation less and more than conditional mean.
1: Hedrum, 2: Nordodal, 3: Røldal.

3.2 Wind force and wave height, Norwegian Sea

In the last years, new oil industry systems are being constructed to operate in the Norwegian Sea. For "optimal" design and operation of such a system it is desirable to know how it responds to its environments. A reasonably complete analysis of this problem is normally complicated even for the simplest system. We shall not embark on an analysis like that. However, it is useful to have preliminary ideas about the structure of some critical environmental conditions. It is suggested that the wind force and the (resulting) wave height impose restrictions of the very rough characteristics shown in Table 3.1 (Håland (21)). A rough idea of the stochastic structure of the "general" operation conditions may then be obtained by considering the "operational conditions" to be a Markov chain with s = 3.

Table A.3: Classifications of wind velocity and wave height (not necessarily consistent) of some relevance to operations at the sea.

Wind velocity (m/s)	Wave height (m)	Operational characteristics
$0 < u < 8$	$0 < u < 1.3$	Few difficulties
$8 \leq u < 14$	$1.3 \leq u < 4$	Difficulties to some systems
$14 \leq u$	$4 \leq u$	Difficulties to many systems

Separate analyses are done for wind velocity and wave height. The data used are from the weather ship "Polarfront". Sampling interval is 3 hrs. Twenty-eight years of data are available.

The identification statistics for wind are shown in Figure 3.6. AIC show a well defined minimum at $p \simeq 3$ both for the summer and winter season. This model is also accepted at a high significance level. The estimated transision probability matrixes are presented in Table A.3. For the case when one class is very unlikely, there may be few observations or none of sequences involving this state. This explains the nonphysical zero- and one-probabilities.

The identification statistics for wave height are shown in Figure 3.7. A second and third order model is clearly accepted

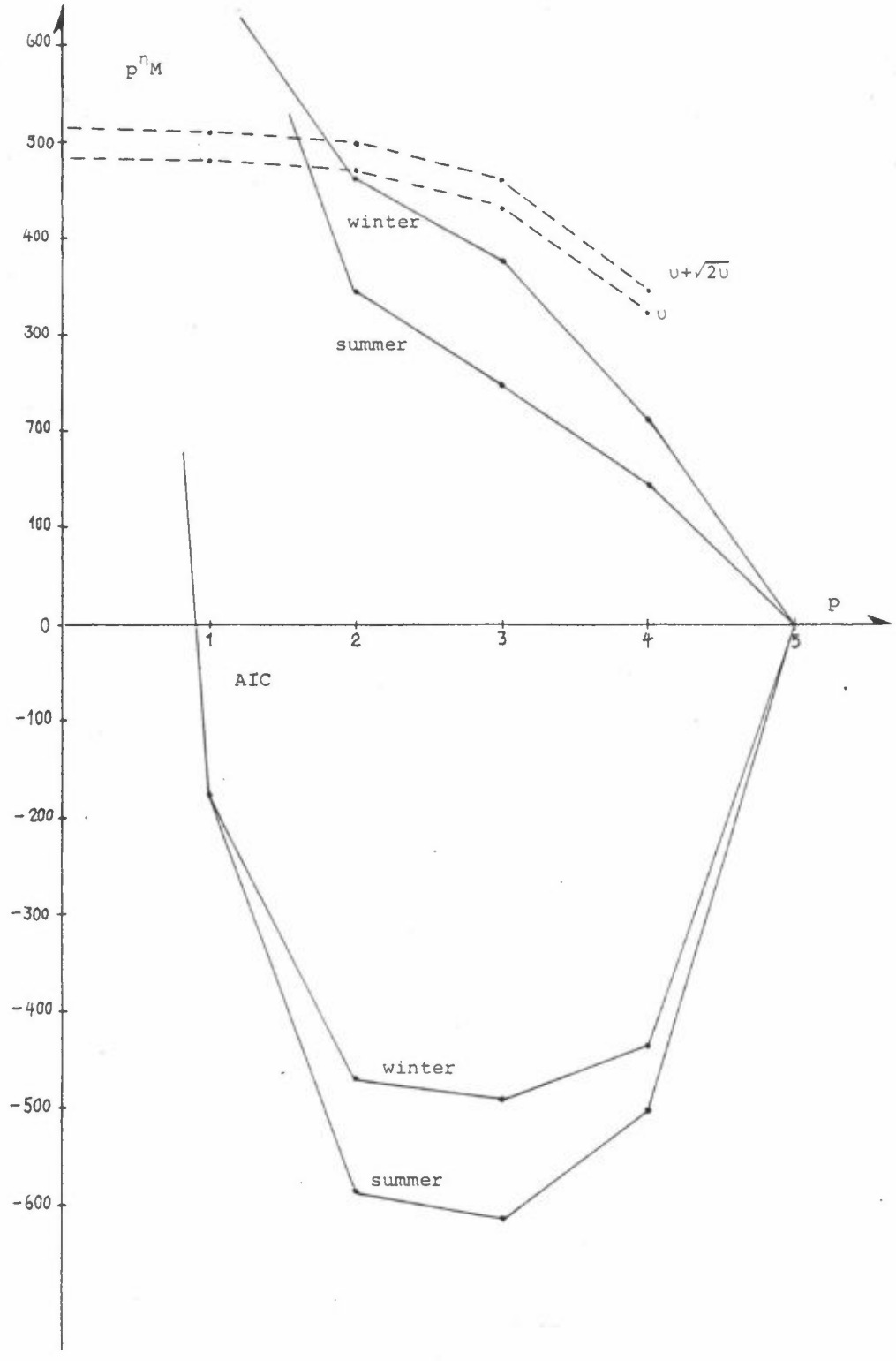

Figure 3.6: Identification variables $p^{\eta}M$ and AIC (p). Wind velocity. Polarfront.

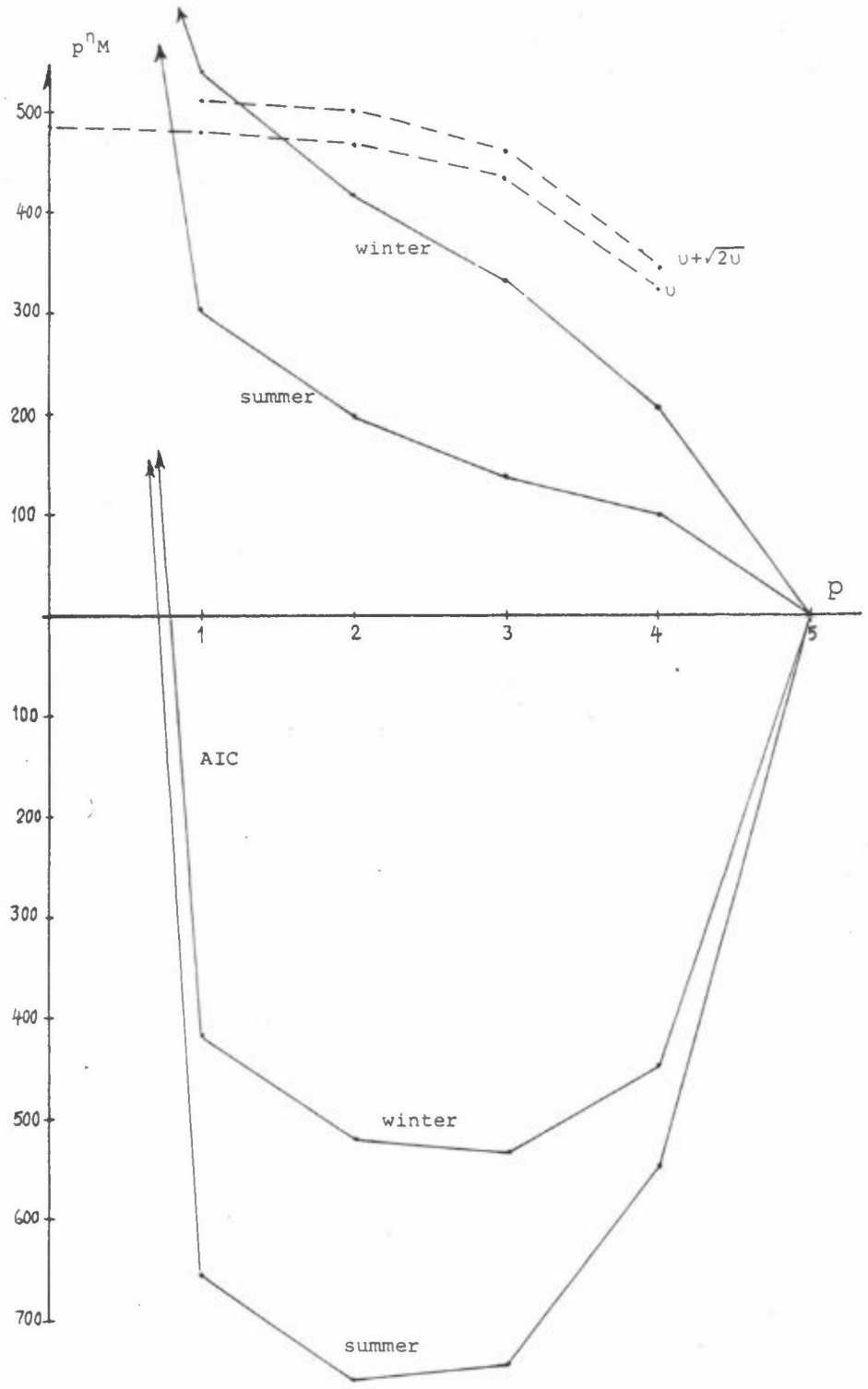

Figure 3.7: Identification variables $p^{\eta}M$ and AIC (p). Wave height. Polarfront.

for the summer and winter data, respectively. The estimated transision probability matrixes are presented in Table A.4.

Table A.3 indicates that the wind force is a nonlinear process. Once it has stayed small (or large) for some time, this state has a tendency to last. Nevertheless, wind components are usually considered to be nearly Gaussian (and linear). The covariance structure of the horizontal wind vector is shown in Figure 3.8. It is observed that the lag has to be larger than 12 before significant deviations from the exponential autocovariance law are encountered. This suggest that an AR(1) process is a reasonable approximation for each velocity component. The figure also illustrates that the small, maximum cross covariance occur at lag 7. The AR-model will have to be of a large order to describe this small effect properly.

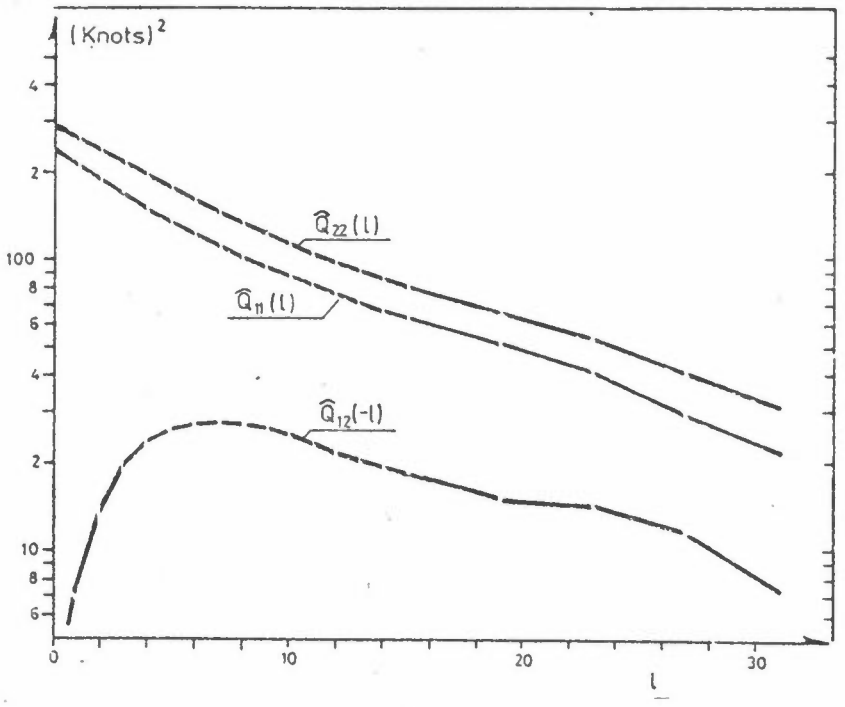

Figure 3.8: Covariance of the horizontal velocity components. Polarfront (Dec, Jan, Feb). Mean value over 28 years.

Figure 3.9 shows that the minimum of AIC(p) has not been reached before p=11. The tendency of local minima for smaller p do even indicate that AIC(p) would continue to decrease for p>12. As illustrated by Table 3.2, the decrease of the one step ahead prediction error from the nonstationary random walk model to the stationary AIC-identified model of the order 11 is very small.

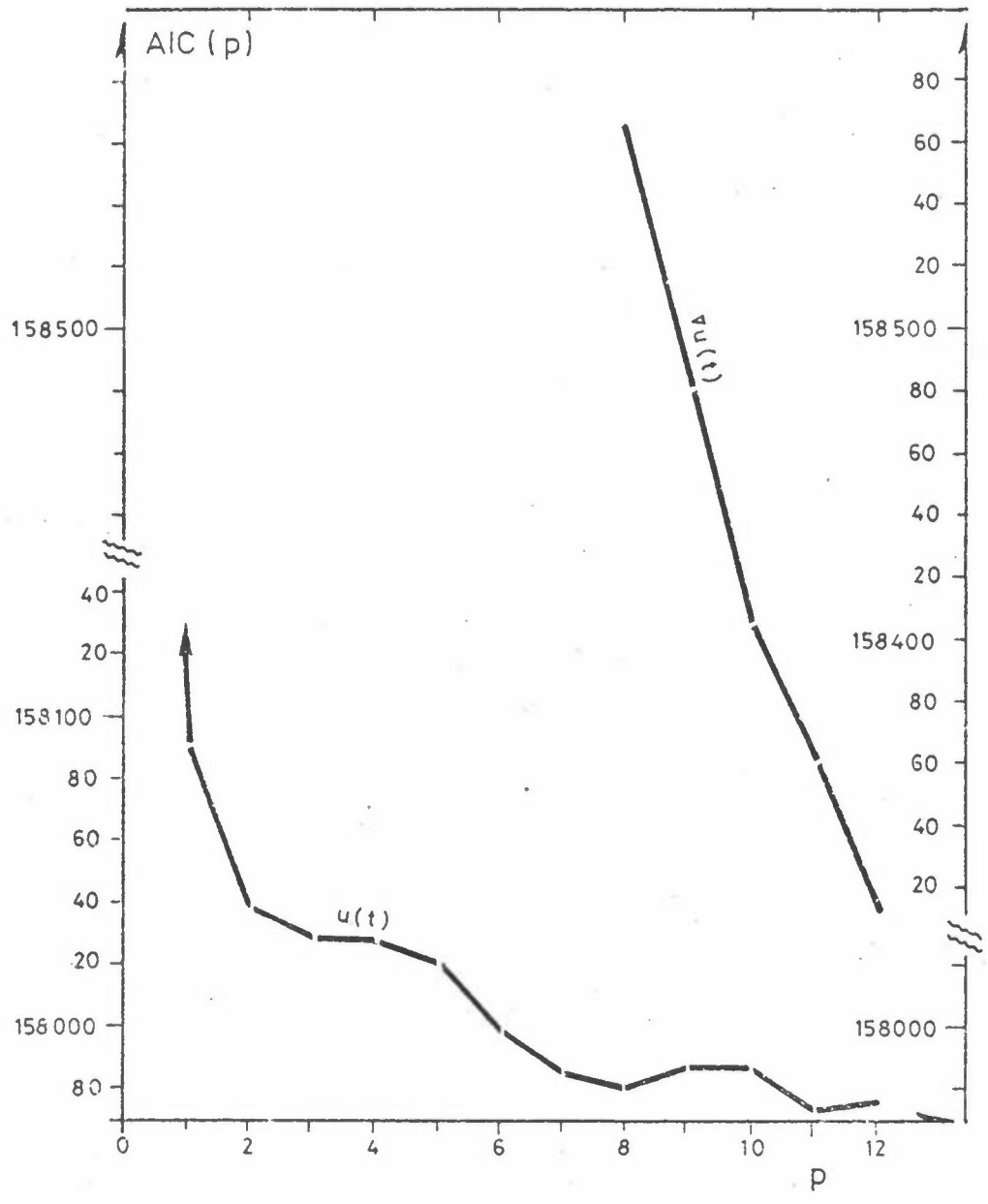

Figure 3.9: AIC(p) for autogressive and autoressive first order integrated models of horizontal wind vector at Polarfront (Des, Jan, Feb).

Table 3.2: Estimated prediction variance, G_{ij}, for three linear prediction schemes. Horizontal wind vector Polarfront. (Des, Jan, Feb.)

	$\hat{u}(t)$=climatic mean (knots^2)		$\hat{u}(t)=u(t-1)$ (knots^2)		$\hat{u}(t)=-\sum_{i=1}^{11} A(i)u(t-i)$ (knots^2)	
Variable no.	1	2	1	2	1	2
1	240.0	0.58	55.0	-3.6	52.0	-3.3
2	0.58	286.0	-3.6	50.7	-3.3	48.8

The minimum (and random walk) 3 hour prediction error of Table 3.2 is typically 3.5 m/s. As the combined effect of measurement error and small scale turbulence is most probably of the order 1-2 m/s, this is judged to be an accurate prediction. Therefore, for some purposes of prediction over lead times a few hours, it appears as the random walk model, with its simplicity, may be the model to prefer, in spite of Akaike's advice. However, as the prediction variance increases linearly with the lead time for this model, Table 3.2 shows that it is useless when the lead time is larger than approximately $5\ \Delta t = 15$ hours.

4 CONCLUDING REMARKS

To decide on the usefulness of Akaike's identification method, diagnostic checking should have been done. Unfortunately, time was not available for this.

It is indicated that the order of a "Markov" chain for the occurrence and nonoccurrence of precipitation should be larger than the traditionally accepted first order. The identified model order for this process increase with the sample size. With large sample size, the minima of the AIC curves are not well defined so that the estimated orders appear to have a significant uncertainty. The sample size normally available for this process (30 years) may be two small for stable estimates. The best approximating "Markov" chain for the process: nonoccurrent, precipitation less and larger than the conditional mean, has been estimated as a second order chain. The likelihood statistics indicate that any low order Markov chain model of this process could be inaccurate.

The wind velocity and wave height at Polarfront have been transformed to discrete variables in such a way as to be of some relevance to operation of systems on the sea. The transformed

variables have been modelled by "Markov" chains. Both for wind velocity and wave height well defined orders are identified. The identified models appear to be accurate. Representation of the horizontal wind vector as an autoregressive process gives a very slow decrease of Akaike's cost function with the order.

Although a finite order "Markov" chain and an autoregressive process can, by proper transformations, be represented as first order (Markov) processes, there is, for some purposes, a great increase in analytical complexity with increasing order. It appears as the increase of model accuracy with the order may, for some purposes, not be worth the associated increase of analytical difficulty. This is one factor not taken into account in Akaike's method. But then, it would also be peculiar if one simple measure could relieve us of all the problems of model identification.

Acknowledgement

This work was done at the Norwegian Meteorological Institute. E. Olsen at the Environmental Data Center provided the Polarfront data and adopted Akaike's TIMESAC programs for identification of vector ARIMA models to our computer.

5 REFERENCES

(1) Box, G.E.P. Jenkins, G.M. — Time series analysis, forecasting and control. San Fransisco, California. Holden Day, 1971.

(2) Akaike, H. — A new look at the statistical model identification. *IEEE Trans. on automatic control.* AC-19 716-723 (1974).

(3) Kullback, S. — Information theory and statistics. N.Y., Wiley, 1959.

(4) Akaike, H. — Information theory and an extension of the maximum likelihood prinsiple. In: *2nd. International symposium on information theory*. Eds. B.N. Petrov and F. Csaki. Budapest, Akademiai Kiado, 1973.

(5) Akaike, H. — Canonical correlation analysis of time series and the use of an information criterion. In: *System identification. Advances and case studies*. Eds. R.K. Mehra and D.G. Lainotis. N.Y., Academic Press, 1976.

(6) Akaike, H. — On the likelihood of a time series model. In: *Institute of Statisticians 1978 conference on time series analysis (and forecasting)*. Cambridge University 1978.

(7) Tong, H. — Determination of the order of a Markov chain by Akaikes information criterion. *J. Appl. Prob.* 12, 488-497 (1975).

(8) Shannon, C.E. Weaver, W. — The mathematical theory of communication. Urbana, Univ. of Illinois Press, 1949.

(9) Bartlett, M.S.
The frequency goodness of fit test for probability chains. *Proc. Camb. Phil. Soc.* 47, 86-95, (1951).

(10) Hoel, P.G.
A test for Markoff chains. *Biometrika* 41, 430-433 (1954).

(11) Good, I.J.
The likelihood ratio test for Markoff chains. *Biometrika* 42, 531 (1955).

(12) Akaike, H.
Maximum likelihood. Identification of Gaussian Autoregressive moving. Average Models. *Biometrika* 60, 255 (1973).

(13) Akaike, H.
Autoregressive Model fitting for control. *Ann.Inst. Statist. Math.* 23, 163-180 (1971).

(14) Eidsvik, K.J.
Ekman Layer fluctuations modelled as autoregressive integrated moving average stochastic processes. Kjeller 1977. (Intern rapport FFI VM-54.)

(15) Eidsvik, K.J.
On near optimal interpolation and extrapolation of atmospheric variables using a few measurement stations. Kjeller 1978. (Teknisk notat FFI VM-295.)

(16) Gabriel, K.R.
Neumann, J.
A Markow chain model for daily rainfall occurrence at Tel Aviv. *Quart. J. Roy. Met. Soc.* 88, 90-95 (1962).

(17) Nordø, J.
Some applications of Markov chains. In: *Fourth conference on probability and statistics in Atmospheric Sciences.* Tallahassee, Florida, 1975. Boston, Am. Met. Soc., 1975, pp. 125-130.

(18) Katz. R.W.
Precipitation as a chain-dependent process.
J. Appl. Met., 16, 671-676 (1977).

(19) Gates, P.
Tong, H.
On Markov chain modelling to some weather data.
J. Appl. Met., 15, 1145-1151 (1976).

(20) Chin, E.H.
Modelling daily precipitation occurrence process with Markov chain.
Water Resour. Res. 13, 949-956 (1977).

(21) Håland, L.
Bidrag til beskrivelse av klimaet på kontinentalsokkelen.
Oslo, Met. Inst., 1978
(Scientific report No. 18.)

APPENDIX A

Table A.1: Transition probability matrixes for the AIC-identified models for occurrence and nonoccurrence of daily precipitation.

Initial Chain \ Final State	January - February						June - July					
	Hedrum		Nordodal		Røldal		Hedrum		Nordodal		Røldal	
	1	2	1	2	1	2	1	2	1	2	1	2
1 1 1 1	83	17	78	22	81	19	78	22	77	23	76	24
1 1 1 2	45	55	49	51	30	70	52	48	42	58	34	66
1 1 2 1	75	25	75	25	69	31	69	31	67	33	61	39
1 1 2 2	42	58	41	52	23	77	43	57	39	61	29	71
1 2 1 1	73	27	73	27	74	26	69	31	69	31	67	33
1 2 1 2	51	49	41	49	25	75	42	58	33	67	34	66
1 2 2 1	68	32	62	38	57	43	71	29	66	34	68	32
1 2 2 2	41	59	40	60	22	78	44	56	38	62	27	73
2 1 1 1	75	25	78	22								
2 1 1 2	39	61	41	59								
2 1 2 1	65	35	61	38								
2 1 2 2	45	55	36	64								
2 2 1 1	71	29	70	30								
2 2 1 2	43	57	39	61								
2 2 2 1	67	33	69	31								
2 2 2 2	33	67	38	62								
Marginal	64	36	62	38	47	53	65	35	59	41	51	49

Table A.2: Transision probility matrixes for the AIC-identified models for nonoccurence, less and larger daily precipitation than the conditional mean.

Initial Chain \ Final State	January - February									June - July								
	Hedrum			Nordodal			Røldal			Hedrum			Nordodal			Røldal		
	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3
1 1	80	15	6	76	17	7	79	17	4	76	16	7	75	18	7	72	20	8
1 2	47	37	15	48	36	16	32	51	17	56	28	17	43	36	20	36	43	21
1 3	36	30	34	36	33	31	13	42	45	34	42	24	30	45	25	17	46	37
2 1	70	19	11	70	22	8	62	31	7	71	19	11	67	21	12			
2 2	42	43	15	44	39	17	31	48	22	44	38	18	47	35	18			
2 3	29	38	33	29	35	36	8	47	45	37	40	23	31	42	27			
3 1	65	24	10	64	20	15	48	34	18	69	22	9	65	19	16			
3 2	44	32	24	51	26	23	32	43	26	51	35	13	37	41	22			
3 3	35	31	34	35	30	35	11	38	51	39	37	25	27	38	34			
Marginal	64	23	13	62	24	14	47	34	18	65	23	12	59	27	14	51	32	17
Cond. mean	6.4			4.0			9.1			7.0			5.6			5.7		

Table A.3: Transition probability matrixes for the AIC-identified models of wind velocity at Polarfront.

1 = (u < 8 m/s), 2 = (8 m/s ≤ u < 14 m/s), 3 = (14 m/s ≤ u).

Initial Chain \ Final State	Winter			Summer		
	1	2	3	1	2	3
1 1 1	81	19	0	91	9	0
1 1 2	20	69	10	29	70	1
1 1 3	3	26	71	0	25	75
1 2 1	68	28	3	70	30	0
1 2 2	13	73	14	19	79	2
1 2 3	3	24	72	6	38	56
1 3 1	50	50	0	33	33	33
1 3 2	5	68	26	25	75	0
1 3 3	0	16	84	0	14	86
2 1 1	74	24	3	84	15	0
2 1 2	22	68	9	31	67	1
2 1 3	3	25	72	0	67	33
2 2 1	67	31	2	75	24	0
2 2 2	14	74	12	15	82	3
2 2 3	1	31	68	1	38	60
2 3 1	33	50	17	100	0	0
2 3 2	8	70	22	6	87	7
2 3 3	1	25	74	0	42	58
3 1 1	65	30	4	50	50	0
3 1 2	9	82	9	33	33	33
3 1 3	0	14	86	33	33	33
3 2 1	67	30	3	92	8	0
3 2 2	13	71	16	16	76	8
3 2 3	0	33	67	0	63	38
3 3 1	47	39	13	100	0	0
3 3 2	10	67	23	6	85	9
3 3 3	1	18	81	1	29	69
Marginal	27	44	28	60	37	3

Table A.4: Transition probability matrixes for the AIC-identified models of wave height at Polarfront.

1 = (u < 1.3 m), 2 = (1.3 m ≤ u < 4 m), 3 = (4 m ≤ u)

Initial Chain \ Final State	Winter			Summer		
	1	2	3	1	2	3
1 1 1	82	17	0	90	10	0
1 1 2	12	85	3	17	83	1
1 1 3	0	29	71	0	0	100
1 2 1	75	24	1	83	17	0
1 2 2	10	85	5	9	90	1
1 2 3	0	7	93	0	28	72
1 3 1	33	33	33	100	0	0
1 3 2	0	100	0	5	90	5
1 3 3	0	13	87	1	27	72
2 1 1	78	21	0			
2 1 2	13	81	6			
2 1 3	0	0	100			
2 2 1	76	24	1			
2 2 2	5	90	5			
2 2 3	1	19	81			
2 3 1	67	33	0			
2 3 2	5	87	7			
2 3 3	1	19	80			
3 1 1	71	29	0			
3 1 2	8	75	17			
3 1 3	0	0	100			
3 2 1	54	38	8			
3 2 2	8	84	8			
3 2 3	0	21	79			
3 3 1	61	33	6			
3 3 2	7	80	13			
3 3 3	1	15	84			
Marginal	19	62	19	48	50	2