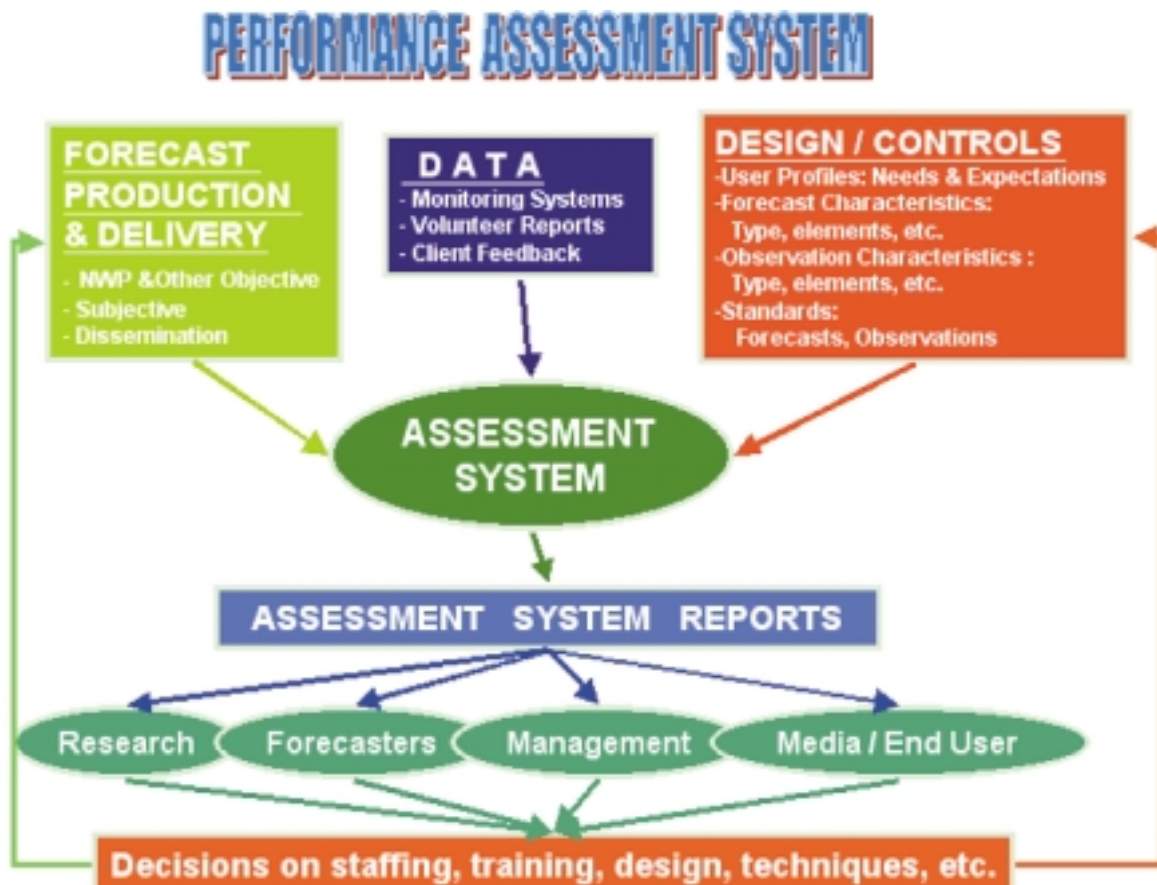# World Meteorological Organization

GUIDELINES ON PERFORMANCE ASSESSMENT OF PUBLIC WEATHER SERVICES

WMO/TD No. 1023
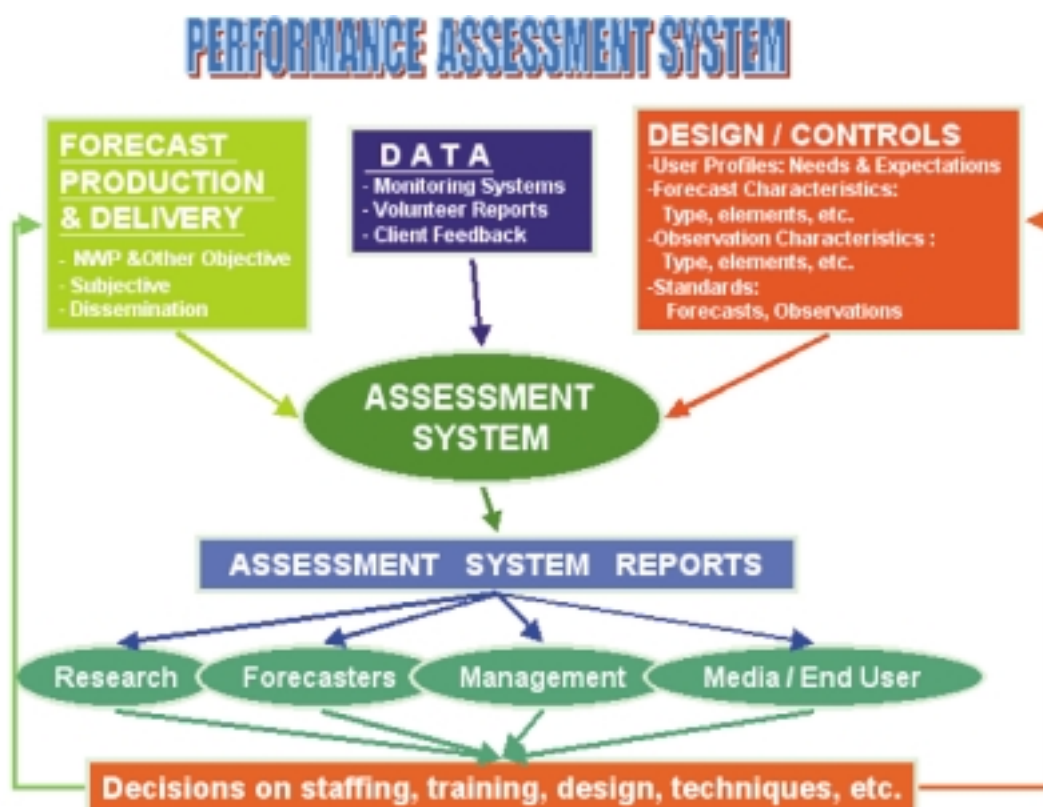
World Meteorological Organization

GUIDELINES
ON PERFORMANCE ASSESSMENT
OF PUBLIC WEATHER SERVICES

Geneva, Switzerland
2000

Text by Neil Gordon and Joseph Shaykewich

Cover design by Irma Morimoto
Graphic provided by Joseph Shaykewich

NOTE

CONTENTS

Chapter 1
INTRODUCTION

Weather services delivered to the public are one of the most visible returns for the taxpayers' investment in meteorological services. It is difficult to quantify this particular Return On Investment in financial terms. It is both possible, and essential, to carry out ongoing performance assessment of public weather services to ensure that they are efficiently and effectively meeting the public's needs.

There are many technical papers and publications on the narrow topic of forecast verification, including numerous accuracy and skill scores. There is less material available by way of guidance on why and how verifications should be carried out, and on the more general topic of assessing whether user needs are being met, rather than just whether forecasts are accurate. Forecast accuracy is irrelevant if the forecast products are not available to the public at a time and in a form that is useful.

The purpose of this Technical Document is to provide broader guidance on performance assessment of public weather services, with something of an emphasis on forecasts and warnings. An assessment programme can be seen in the context of a quality system, where it is important to ensure that the information gathered and processed is focussed on user requirements, to be used in making decisions and taking actions to improve performance, rather than just being gathered for the sake of it. In essence, the object of the exercise is to ensure a sustainable and cost-effective system delivering quality public weather services.

The guidelines are based on an outline developed at a meeting of the WMO Public Weather Services Expert Team on Product Development And Verification And Service Evaluation, in Hong Kong, China in November 1999. Two of the terms of reference of this team were to "Prepare recommendations on standardised verification techniques for public warnings and forecasts", and to "Prepare guidelines on technical and user-oriented verification mechanisms including measures of overall satisfaction with the service". This guidance addresses both terms of reference in the context of overall performance measurement, but does not provide hard and fast rules on standardised verification techniques.

Some of the basic guidelines about performance assessment include:
- Know why you are carrying it out (what new information do you want to discover?)
- Do not just collect and process information and then file it away
- Be prepared to take actions based on the results
- Gather information designed to help a National Meteorological Service (NMS) make strategic decisions about all aspects of public weather services

- Favour simplicity where possible, rather than overly complicated schemes
- Be very careful about the statistical significance of results based on small samples or short records
- Provide regular reports to stakeholders
- Make relevant, interpreted, information available to the public.

There are two major methods available for gathering information in an assessment programme – *Verification*, and *User-Based Assessment*. Neither can stand alone. It is important to do both, in a balanced fashion. The amount of effort spent on each will depend on the country, the nature of the services, and the user community. The worst thing would be not to do either of them!

The overall purpose of *Verification* of forecasts is to ensure that products such as warnings and forecasts are accurate, skilful and reliable from a technical point of view. As far as possible, forecast verifications are produced in an objective fashion, free of human interpretation. The results tend to be numbers and statistics, which can be manipulated and interpreted using statistical theory. There is no guarantee that verification results will match people's perceptions of how good the forecasts are. Nonetheless, information gathered through verification can be very useful for improving the accuracy of forecasts.

On the other hand, *User-Based Assessment* should give a true reflection of the user perception of products and services provided by the NMS, as well as qualitative information on desired products and services. It is almost completely subjective information, subject to human perception and interpretation.

In carrying out an assessment programme combining both methods, there are some commonalities. Although verifications may typically provide objective numbers, they should still be based around numbers which are relevant to users. It should be possible to match user-based assessment results (e.g., of perceptions of forecast accuracy) with corresponding technical verification results, and seek common trends and patterns. In both methods, there is no single score or method that can give "The Answer". Various scores and assessment methods have their particular uses.

In Chapter 2 of this Technical Document, the three key purposes for performance assessment will be discussed.

Services can only improve if actions are taken – the six main areas are dealt with in Chapter 3.

Chapter 4 considers in detail how to carry out Verifications, and Chapter 5 is on User-Based Assessment. The final chapter reviews why and how to carry out an assessment programme, and provides some guidance on an "entry-level" programme.

Chapter 2
KEY PURPOSES

There are three key purposes for carrying out an assessment programme for public weather services. They are:

(1) Ensuring that public weather services are responding to user requirements
(2) Ensuring the effectiveness and efficiency of the overall public weather services system
(3) Ensuring the overall credibility and proven value of public weather services.

Another way of looking at this, is that the three purposes are about:

(1) Making sure that you are providing the right products
(2) Making sure that you have a good system for making them
(3) Building stakeholder support for the NMS.

2.1   ENSURING THAT USER REQUIREMENTS ARE MET

There are a wide variety of end-users of public weather services. These include individual members of the general public, emergency management agencies, and paying customers for specialised services.

In order to make sure that user requirements are being met, first of all it is necessary to know what they are – and what better way than asking the users? This topic is covered extensively in Chapter 5.

The definition of the needs in the particular case of weather forecasts can encompass what weather elements are most important, when and how forecasts should be delivered, in what format, and with what accuracy.

Knowing what the needs are, it is necessary to find out whether they are being met, and take actions to improve where possible.

This may be as simple as checking and then changing the issue time of forecasts to make sure that they are available when they are most useful. It can also involve keeping score on how many forecasts are issued late, and changing management practices and schedules to ensure that forecasts are issued on time.

Verifying the accuracy of forecasts is, of course, another aspect. But it needs to be done in ways that are relevant to the user, who has probably never heard of a "Brier Score".

2.2   ENSURING THE EFFECTIVENESS OF THE PUBLIC WEATHER SERVICES SYSTEM

It is one thing to provide public weather service that meet user needs – and quite another to do it effectively and efficiently, from an overall point of view. This purpose is not about *what* is delivered and *how*. Rather, it is about the organization, management and planning of the overall public weather services system that delivers the services.

A performance assessment programme can gather information that can be used to make strategic decisions about the future delivery of services, about staffing, about training, research and development, and about the best mix of information from computer models and from human value adding.

2.3   ENSURE CREDIBILITY OF AND SUPPORT FOR THE PUBLIC WEATHER SERVICES SYSTEM

Even if public weather services have been designed and delivered to meet user needs, there may be a perception problem over how good they are. This can be serious, and life threatening. For example, if the public has a poor perception of the accuracy of topical cyclone forecasts, they may disregard warnings, resulting in major loss of life and property. In the best of all possible worlds weather forecasts will never be perfect, so this can be a vicious circle, with public credibility declining every time there is the inevitable poor forecast.

An assessment programme can assist in two ways – by finding out what the public perceptions are, and by gathering and publicising facts about performance to improve the public perception and credibility of the services. Those occasions that forecasts do go wrong can be used as opportunities to publicise the role of the NMS and to draw attention yet again to the *fact* (gained from the assessment programme) that, say, forecasts are usually 85% accurate.

Similar information on performance can be incredibly useful for gaining the support of other stakeholders, including government ministers responsible for the NMS. The NMS will be in a much stronger position for sustaining and building funding if it can demonstrate such things as its level of performance, public satisfaction with its services, and the impacts of previous investment and research and development programmes.

Chapter 3
AREAS THAT ACTIONS ARE REQUIRED TO MEET THE KEY PURPOSES

There is no point in gathering information through an assessment programme without using it. Using it means taking actions. This chapter is about the six main areas where actions need to be taken – mostly through changing what is being done now (unless it is perfect, which is unlikely!) or making plans for future changes.

(1) Improve the products to be provided
(2) Improve how the products are delivered
(3) Improve the production system
(4) Carry out needed research and development
(5) Train and develop staff
(6) Communicate information.

All of these action areas should involve feedback loops. Information is gathered on user requirements and on performance levels. Actions are taken to improve matters. The final step of "closing the loop" is also important – checking what the actual impact was of those actions, in order to learn how to do better next time.

Of course, there is also an assumption here that the NMS has the resources and staff to take such actions. There may well be a gap between the measured performance and expectations, but no ability to improve it because of lack of resources, or because there are no people available to carry out training.

The fundamental management issue here, which is beyond the scope of this Technical Document, is how best to allocate limited resources (and they are always limited) to best effect, to improve the situation, based on the information gathered from the assessment programme.

3.1   PRODUCT DEFINITION

The product definition is assumed to include what information is included in the product, and how it is formatted and expressed. This may include, for example, criteria for warnings.

The techniques discussed fully in Chapter 5, such as surveys, focus groups, and direct visits and discussions, can be used to identify user requirements for products. Naturally, this will not be done in a vacuum, since many products will already exist. It is crucial to ensure that the information gathered can be used to make decisions and take actions on product definition. This should always be borne in mind when designing the survey – know why you are asking the questions, and have some idea about what you are likely to do, depending on the answers.

3.2   DELIVERY MECHANISMS

Part of the user requirement is how the product is delivered, and when. Similar methods to those in the previous section need to be used to check with the users on what capabilities they have for accessing and receiving products, and then to improve the delivery system to better meet those needs.

3.3   PRODUCTION SYSTEM

There are many aspects of the production system that may need to be changed as a result of information gathered in an assessment programme. Just a few of the numerous possible changes are:

- Re-configuration of data networks to gather new data required for products and services, possibly at the expense of data which may no longer be required
- Obtaining new sources of local or global NWP model information on which to base new products and services
- Revising shift schedules to accommodate new, or modified (or discontinued!) products
- Revising shift schedules to accommodate new delivery times
- Installing systems (e.g., fax machines, or a web server) for new means of delivery of products
- Using more automated products (e.g., for maximum temperature forecasts) if verifications prove that these satisfy accuracy requirements and they can be cost-effectively produced
- Devoting more forecaster shift time to producing critical warnings which have proven not to be accurate enough
- Centralising forecasting, or de-centralising forecasting.

3.4   RESEARCH AND DEVELOPMENT

Information gathered through verifications, and user-based assessment, can be used to determine the priorities for research and development, and to reshape what R&D needs to be carried out. Some typical examples of the actions that may take place as a result are:

- Research and document case studies of weather situations which have been shown through verifications to be poorly handled (e.g., heavy rain situations)
- Basic research into phenomena where improvements are demonstrably needed (e.g., tropical cyclone development)
- Development of forecast techniques for new services (e.g., prediction of road surface icing)
- Development and improvement of local or regional NWP models in support of many products
- Development of statistical post-processing of NWP model output for new products (e.g., precipitation probabilities) or to improve existing products.

The most important aspect of all these examples is that they are driven by the knowledge of user requirements, and of existing performance levels – gained from the assessment programme.

3.5 STAFF TRAINING AND DEVELOPMENT

Once again, there are many actions that may take place as a result of information from a performance assessment programme. A few examples are:

- Recruiting and training more forecasters based on projected shift requirements from planned introduction of new products and services
- Training staff to make use of new numerical guidance information
- Training staff on the scientific basis of a new product, and operational procedures for producing it
- Re-training staff on the fundamental meteorology of a weather phenomenon which verifications show is being poorly forecast
- Training staff on how to write forecasts in a new and more "user-friendly" style (which surveys have shown the public would find more useful)
- Training staff on how to reduce a known bias of over-forecasting precipitation occurrence.

3.6 COMMUNICATION

One of the most important actions that must be taken is to communicate the results and information gathered from a performance assessment programme. Information is only of value if people know about it. It must be in a form that is understandable to the audience, and tailored to their likely use of it.

Firstly, information gathered must be made available to the staff of the NMS. Managers need information to guide them in decision making. Forecasters need information by way of feedback on their performance, particularly in relation to systematic errors that may need to be corrected. Researchers need information on performance of the system, and on likely new products so they can plan and prioritise R&D. All staff need information on the technical accuracy of the services delivered, and on public expectations, perceptions and needs. **All staff should have a sense of ownership, accountability, and pride in what is being delivered to the users.**

Secondly, relevant and appropriate information must proactively be made available to stakeholders in general. This may be a formal requirement of some kind of "Service Charter" or agreement with the government or community at large on services to be provided. Communicating such information is particularly important in relation to the third key purpose of "Ensuring the overall credibility and proven value of public weather services". If there is a vacuum of information, particularly on demonstrated performance, public perceptions will be based on anecdotal evidence. People tend to remember the last time a forecast went wrong – not how well forecasts do overall.

The most important stakeholder is the source of funds for the NMS – the government on behalf of the taxpayers. Information from a performance assessment programme must be communicated to demonstrate performance, to demonstrate the beneficial impacts of previous investment in the NMS, and in support of future plans for the development of the NMS.

Finally, and often in reaction to events, information must be communicated to the public via the media when opportunities present themselves. A good example is when there has been a severe weather event. Whether or not this was well forecast, the public interest in severe weather is heightened, and this is a good opportunity to include information on overall performance of the public weather services as part of the "weather story", to build public support and credibility.

Chapter 4
VERIFICATION

4.1   INTRODUCTION

4.1.1   Overall Purpose

The overall purpose of verification is to ensure that products such as warnings and forecasts are accurate, skilful and reliable from a technical point of view. This is distinct from whether the products are actually meeting user needs, which is covered separately in the next chapter. Nonetheless, the technical assessments should be in terms of measures that are *relevant* to user needs.

There are many dimensions and techniques of forecast verification. This Technical Document is not intended to cover all possibilities, but to provide sufficient general information on the possibilities. An extensive survey of verification techniques was carried out by Stanski *et al.* (1989) and published by WMO. The work by the late Allan Murphy (1997) is also worth reviewing for his philosophy on verification, and for the list of references.

4.1.2   Accuracy, Skill and Reliability

In concept, forecast verification is simple. You just need to compare the forecast weather with the observed weather actually occurred. The accuracy[1] of a forecast is some measure of how close to the actual weather the forecast was. The skill of a forecast is taken against some benchmark forecast, usually by comparing the accuracy of the issued forecast with the accuracy of the benchmark. A benchmark forecast can be something simple such as climatology, chance, or persistence, or it could be a partly or completely automated product. The skill measure should give some meaningful information about what value has been added in the forecast process, compared to the usually much simpler or cheaper benchmark forecast.

There is a great deal of theory and practice about measures of forecast accuracy, involving sometimes-complex formulas for comparing frequency distributions of forecast versus observed weather. Usually, an accuracy measure gives information on the spread of differences between forecast and observed weather. A typical example is a Root-Mean-Square-Error (RMSE) – the square root of the mean of the squared difference between forecast and observed.

Reliability is another aspect of forecast accuracy (it does not involve comparison with a control forecast). Literally, this means the extent to which the forecast can be "trusted" on average. One measure of reliability would be the average bias in a maximum temperature forecast – the average of the forecast values minus the average of the observed values.

Reliability measures are also used to assess how closely forecasts expressed in probability terms match reality. For example, suppose you were verifying a set of many forecasts of the probability of occurrence of rain. Suppose also that there were 100 occasions when the forecast probability was around 30 % (e.g., between 25% and 35%), but it only rained on 10 of those occasions. The implication is that the forecasts of a probability of 30% chance of rain were not very reliable, since it really only rained 10% of the time on average.

4.1.3   Objective and Subjective Verifications

There are two main ways of verifying forecasts – *objective* and *subjective.*

Objective verification is based on purely objective comparisons of forecast and observed weather elements. There is no element of human interpretation of either the forecast or observation.[2] The results can be replicated. Objective methods should be based on sound statistical theory – essentially the comparison of observed and forecast numbers.

Subjective methods involve some human assessment of forecasts and/or observations. They are a result of human perception, and the results are not always consistent and cannot necessarily be replicated. However, these perceptions are a true reflection of the value of the forecast to the individual or user who does the assessment.

4.2   GUIDING PRINCIPLES

Unless careful planning is done, there is a risk that a verification programme will never get off the ground, or that it will be engulfed in an avalanche of numbers that are never used. The purpose of this section is to suggest guiding principles on the *Why*, *How* and *What Next* of Verification.

4.2.1   Principles Related to Why to Verify

There are four main reasons for verifying forecasts:
(1) We must know the quality of our products
(2) We need information to aid decision-making
(3) We need information to feed back into process improvement
(4) We need appropriate information for reporting to users and other stakeholders.

---

[1] There is sometimes confusion between *accuracy* and *precision*. The *precision* of a forecast is how much detail is put into it in time, space, weather elements, and numbers of significant digits in numerical values. For example, a forecast maximum temperature of 23.42963°C would be very *precise*, but that does not make it *accurate*!

[2] The observed weather element may of course have been made by a human observer as part of a routine weather observing programme – this can be distinguished from subjective assessment of observations such as estimating precipitation that occurred in a spot in a data-sparse region.

Knowing the Quality of the Products

It is essential for any service provider to know the quality of the products and services they provide.

However, historically, because of some of the perceived difficulties of verifying weather forecasts, and the work involved, NMSs have probably not done this as much as they should have.

That time of not knowing is now over. In an era of shrinking budgets for NMSs, increased demands for accountability for expenses and investments, and competition, NMSs *must* know how well they are doing. *Assumptions* about how well they are doing are no longer good enough.

Furthermore, the information gathered on forecast quality can be extraordinarily valuable, provided that it is carefully gathered and analysed, and appropriately used. Information on forecast quality is like having a medical check-up – it can help you work out what parts of your forecast production system are working and what are not. It can provide *facts* rather than *assumptions* for discussions with customers, and the media, and the government.

Information to Aid Decision Making

NMSs are continually making decisions that involve allocation of resources, staffing, training, research and development, and large expenditures. It is vital to make sure that sufficient information is available on the quality of the final output products to support these decisions.

Measuring and quantifying forecast performance allows you to compare forecasters, and forecast systems, and perform "what if" scenarios on how different systems might perform.

Many examples of where actions can be taken, and decisions made, can be found in Chapter 3 of this Technical Document.

Feed Back into Process Improvement

Verification results should provide information that is of value in ongoing process improvement in forecast operations. Just one simple example would be recognition that rain is forecast far too often. Verification information can be analysed further to see what the weather conditions are like when the forecast was wrong, and to look for trends. You might find that there are particular weather conditions when the over-forecasting is raking place. Forecasters can use this information to improve their own performance, and it can be used to drive research and development projects.

Since verification involves a comparison between forecasts and observations, it can be used to pick up quality problems in either. If the forecasts are being passed through some automatic decoder program that is having problems, this may indicate that some forecasters are using the wrong syntax for writing their forecasts. (This can be fixed by training the forecasters to do better, or by putting new systems in place that do not allow forecasts to be written the wrong way to start with.) Large,

systematic differences between the forecast and observation may turn out to be a problem in the observation, not the forecast!

Appropriate Information for Reporting

Much of the information from a verification programme can be used internally.

However, there is also an increasing, and perfectly understandable, demand from users and other stakeholders for information on the quality of products and services. Providing such information can be very useful for an NMS.

Users sometimes have an incorrect perception of the quality of forecasts, which can be corrected by sharing appropriate verification information with them. Of course, the verification information may also validate their perceptions or poor forecast – there is no point in hiding this, but there will be value in discussing the issue with users and working together on how the forecasting can be improved to better meet their needs.

Government ministers like to have proof of "value for money" expended on NMSs, and particularly like to see evidence of improvements over time, as a payback for money that they have committed to the NMS budget.

Verification information can be useful in dealings with the media, particularly when countering any negative publicity on a particular forecast that may have gone wrong.

A key word here, of course, is "appropriate". Information for reporting purposes needs to be carefully selected, simple, and relevant for reporting purposes. Complicated and hard to understand scores will not enhance the image of the NMS.

4.2.2   Principles Related to How to Verify

When considering how to conduct verification, it is vital to refer back to the principles in the previous Chapters on why verification is being done. If the "how" of verification is not answering questions or providing information needed under "why", then it may not be needed.

There are four key principles on how to verify forecasts:
(1) There Should Be an Overall Plan
(2) Measures Must be Relevant to the Users (internal and external)
(3) Keep It Simple
(4) Use Consistent Elements, Locations, Methods and Scores.

Overall Plan

Before embarking on a verification programme, it is very worthwhile to take some time to develop an overall plan. This should cover many of the issues addressed in this Technical Document, focussing on particular issues for your country. Those staff who will be producing and using the results need to be involved in the development of the plan, to ensure ownership, a commitment to success, and broad understanding of the purposes.
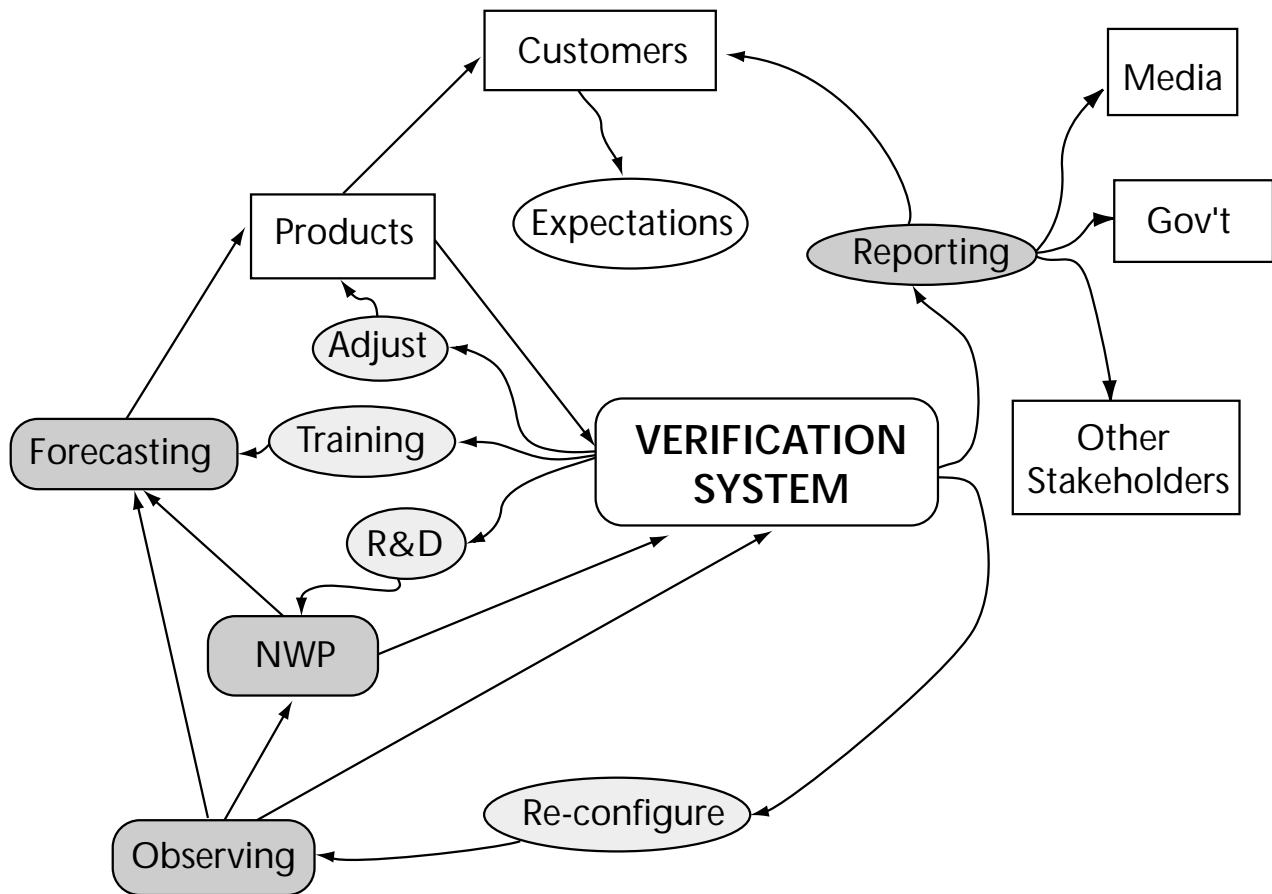
The plan needs to take into account why the measures are being produced.

The diagram above illustrates the overall information flows in an operational verification system. Meteorological information and product flows are shown with straight lines. Observations are used in NWP and by forecasters, who then produce products, which go to users. The observations, NWP information, and products also feed into the verification system. This system employs user expectations, to produce reports for the paying customers, and for the media and government and other stakeholders. Information from the verification system may also be analysed and used to make decisions about re-configuring of the observing system, what research and development may be done to improve NWP and to feed into training to improve forecaster performance, and also to adjust the definition and format of products.

User-relevant Measures

Information should be relevant to the needs of the users. There is little point in producing scores that are complex and satisfying theoretically, and have all the right attributes of proper[3] scores, if no one can understand or use them. For example, scores which give "percent correct" accuracy are not always favoured by the theoreticians, but they are easily understood by the public.

It is important that the verification scheme truly reflects the perception of the public or users on the accuracy of the forecast. Surveys may show that the public believe that a temperature forecast is "correct" if it is within 3°C, and verifications can then be made in those terms. However, a higher level of accuracy may be needed by an electricity supplier wanting to forecasting power demand, for whom the temperature forecast may need to be within 1°C.

It is also important that the system captures how good performance is for the times when the forecast most needs to be right – the relevant and critical times. For example, in a place that rarely gets frosts, a constant forecast of "no frost" may be right 99% of the time, but is clearly of no value, since it always says the same thing.

Depending on the climate of the region and the time of year, some weather elements are more important than others. For example, there may be little value in verifying maximum temperatures in a region where they always vary little from day to day.

You may also take into account the needs of internal users of the information for decision making. For example, some particular skill measures may be useful when making decision on the value of numerical guidance and the value added by forecasters.

Keep it Simple

Embarking on a verification programme can be a daunting prospect for an NMS with little experience in this area. It is better to use simple, easy to understand measures, than to

---

[3] A "proper" score is one that encourages a forecaster to forecast what he or she truly believes, rather than biasing (or hedging) the forecast one way or another in the hope of producing a better score.

implement very complex schemes. It is also better to concentrate on verifying for just a few key places, rather than trying to verify many weather elements for many places. Keeping the number of verifications down avoids being buried in numbers that are never analysed, and keeps costs down.

Consistency

One of the most useful aspects of verification information is that the results can be tracked with time to see how performance is (one hopes) improving. But performance cannot be tracked if the weather elements, locations, methods and scores keep changing. And tracking performance in a statistically significant way may take a long time series of information. For example, at least four years of data will be needed to analyse seasonal differences in performance in a meaningful fashion.

It is, therefore, important to ensure consistency in an ongoing verification programme. You should be consistent by using the same weather elements, from the same locations, for the same times, and using the same accuracy and skill measures. Then results can be tracked in time, rather than trying to work out whether change in skill were due to using a new score, or to verifying for a different location after a couple of years.

However, it can also be very useful to save the raw data used for the verifications so that if some new verification method is introduced it may be possible to go back and recompute the verifications results from the beginning.

4.2.3   Principles Related to What to Do with Results

The ultimate benefit of a verification programme will only come about when the results are used, in support of the four reasons we are actually doing verifications (see Section 4.2.1). The key principles are quite simple, really:
(1)   Use the results
(2)   Do not misuse the results.

Using the Results

Communicate them: In general, the results should be communicated appropriately and promptly, rather than just being filed away. This will facilitate general use of the information. Communication includes reporting to users and stakeholders, and providing direct, immediate feedback to forecasters. Forecasters are usually very interested in the results of verification. They want to know if they have systematic errors in their forecast so that they can correct them.

Analyse them: The results should be analysed to assist in decision making.

If the verification results are not acceptable, then decisions may need to be made on the end-to-end forecasting process in order to improve matters. This could include improved data gathering, better numerical guidance, research and development targeted at the weather elements being verified, training programmes, improved procedures, processes and tools in the forecast room, staffing levels.

Analysis of the results should be ongoing to ensure that benefits are coming from these improvements.

If the results are acceptable, this information can be used to validate previous decisions, and to assess the likely future impact of new decisions to be taken.

Use Them for Process Improvement: On a shorter timescale, verification results should provide information that is of value in ongoing process improvement in forecast operations. Just one simple example would be recognition that maximum temperature forecasts for a city tend to have a warm bias (say, of 1.5°C) – forecasters can use this information to improve their own performance.

Not Misusing the Results

Verification results based on small sample sizes, or of rare events, may have very large margins of error. It is a good idea, where possible, to compute error bars on verification results. Care is needed in interpreting information that has poor statistical validity. This includes being too proud of very good results (which may not last!) or too concerned about very poor results (which hopefully also won't last!).

You should be careful to double check the results if they are either very good or very bad – there may have been a problem with the data or with the computer programs.

Care must also be used in trying to compare results between regions with different climates, which may not be meaningful, even if the verification methods were exactly the same.

4.3   PERFORMANCE MEASURES

There are many scientific papers and documents on various measures of performance that can be used for verification. See, for example, Stanski *et al.* (1989), and Murphy (1997). The intent of this Technical Document is not to duplicate such material, but to give a sample of the simplest and most common measures that can be used, together with some brief examples of their application.

There are two fundamentally different types of variables, which can be forecast in two fundamentally different ways.

The two types of variables are *continuous* (numbers), and *categorical* (e.g., rain or no-rain, or a category of precipitation amount).

They can be forecast either *deterministically*, by giving just a single value or category, or *probabilistically*, through giving some information on the probability distribution of the continuous number, or the individual probabilities for the possible categories which could occur.

A forecast expressed in probability terms is more useful for making decisions than a forecast that explicitly states what will occur. The user can choose to take one or other decision based on the probabilities, and their particular knowledge of the costs of taking decisions, and rewards or losses depending on the weather that actually occurs. In the final analysis, the value of a probabilistic forecast comes down literally to the value that such a sophisticated user can extract by making decisions based on the forecast rather than some benchmark assumptions.

In this section typical performance measures for the most common types of forecast will be discussed.

4.3.1 Deterministic Forecasts of Values of Continuous Weather Variables

The most common forecasts are of actual values of weather elements, as real numbers (as distinct from probabilistic forecasts of numbers). Examples of such weather elements are:
- Temperature
- Wind speed
- Wind-chill
- Humidity
- Precipitation amount.

The following simple example of a set of twenty maximum temperature forecasts will be used in this section to illustrate the scores. Both the forecasts and the observations have been rounded to the nearest whole degree Celsius, since this is how the public usually see or hear them. In real life, twenty forecasts would be far too small a sample to draw any conclusions from. This example is purely intended to explain the various scores and how they can be interpreted.

The table includes other columns of information, which will be explained later.

MAX TEMP (°C)					
Forecast (F)	Observed (O)	F-O	ABS(F-O)	(F-O)^2	Within ±2°C
17	17	0	0	0	1
24	20	4	4	16	0
28	29	-1	1	1	1
22	25	-3	3	9	0
14	16	-2	2	4	1
16	17	-1	1	1	1
17	17	0	0	0	1
16	16	0	0	0	1
15	14	1	1	1	1
19	18	1	1	1	1
22	19	3	3	9	0
21	17	4	4	16	0
16	18	-2	2	4	1
20	18	2	2	4	1
27	31	-4	4	16	0
21	20	1	1	1	1
15	14	1	1	1	1
22	28	-6	6	36	0
20	23	-3	3	9	0
15	18	-3	3	9	0
Average: 19.4	19.8	-0.4	2.1	6.9	60%
		Bias	MAE	MSE	% correct

Reliability

Suppose there are *N* forecasts f_i and corresponding observations o_i for $i = 1...N$

A gross measure of reliability is the mean bias. It is simply the average of the forecast value minus the average observed value, or

$$bias = \frac{1}{N}\sum_{i=1}^{N}\left(f_i - o_i\right)$$

For our simple example, *N* is 20, the average forecast maximum is 19.4°C and the average actual maximum is

19.8°C, so there is a slight bias of -0.4°C – on average the forecast maxima were 0.4°C colder than the actual maxima.

Other more complicated reliability measures can be computed. For example, the bias could be considered separately for forecasts of colder than 20°C, compared to forecasts of 20°C or more, to see whether the bias depends on the forecast. It might be that forecasters tend to underdo the maximum temperatures more when they expect it to be colder.

Before carrying out calculations of more detailed bias information such as this, it is important to think about what reason there might be for variations.

Another way of looking for bias is also simply to plot the forecast versus observed values. This is easily done these days using standard spreadsheet software. The following graph shows the forecast versus observed maximums, together with the line representing a "perfect forecast". While this is *far* too small a sample to draw any definitive conclusions from, there is a hint here that both the coldest forecasts and the warmest forecasts tend to be too cold.
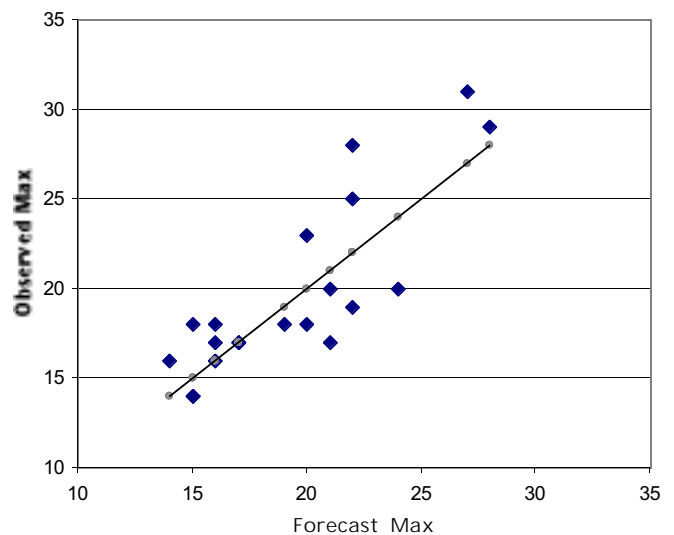
Accuracy

Various accuracy measures are shown in the previous table for this example.

In terms of accuracy, the *Mean Absolute Error* or *MAE* is:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left(\mid f_i - o_i \mid\right)$$

For the example, this is 2.1°C. The *MAE* is a very simple measure of accuracy to use and to explain to users – "it's the average difference between the forecast and observed temperature". However, people are often more concerned about the large errors, and this measure does not take these into account as much as ….

The *Mean-Square Error* or *MSE* is

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(f_i - o_i\right)^2$$

For the example, this is 6.9. The *MSE* is affected more by large errors, and has the nice statistical property of

being a "proper" score – forecasters will do best if they always forecast the average of what they truly believe the maximum temperature is likely to be. It is also the quantity that is minimised with classical linear regression equations that try and relate some predictor variables to the variable being predicted (the predictand).

However, the *MSE* has unfriendly units of °C squared. So, instead, what is usually used is its square root ….

The *Root-Mean-Square Error* or *RMSE* is

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f_i - o_i)^2}$$

This has units of °C, and for the example the *RMSE* is 2.6°C.

Another measure that is commonly used for weather elements such as temperature, is the "percent correct" of forecasts that are within some allowable range, e.g., within ±2°C or ±3°C. This is shown in the above table by putting a 1 when the forecast was within ±2°C of the observed maximum, and 0 otherwise, then averaging the values. The result for this example is that 60% of the forecasts are within ±2°C.

It is obviously crucial for this measure to know what the public or specialised user considers to be a "correct" forecast. But this measure of accuracy is a very simple and useful one to explain to the public once this has been decided.

Skill

Skill is measured against some benchmark forecast – typically climatology, persistence, or perhaps a numerical guidance forecast.

Continuing with the same example, suppose that the benchmark forecast is taken to be the climatological maximum temperature for this period of 20°C. Then the corresponding table for this benchmark forecast is:

MAX TEMP (°C)					
Benchmark Forecast (F)	Observed (O)	F-O	ABS(F-O)	(F-O)^2	Within ±2°C
20	17	3	3	9	0
20	20	0	0	0	1
20	29	-9	9	81	0
20	25	-5	5	25	0
20	16	4	4	16	0
20	17	3	3	9	0
20	17	3	3	9	0
20	16	4	4	16	0
20	14	6	6	36	0
20	18	2	2	4	1
20	19	1	1	1	1
20	17	3	3	9	0
20	18	2	2	4	1
20	18	2	2	4	1
20	31	-11	11	121	0
20	20	0	0	0	1
20	14	6	6	36	0
20	28	-8	8	64	0
20	23	-3	3	9	0
20	18	2	2	4	1
Average: 20.0	19.8	0.3	3.9	22.9	35%
		Bias	MAE	MSE	% correct

For example, if MAE_f is the *Mean Absolute Error* of the forecast, and MAE_b is the *Mean Absolute Error* of the benchmark, then one skill measure is

$$\frac{MAE_b - MAE_f}{MAE_b} = 1 - \frac{MAE_f}{MAE_b}$$

which will be zero when the forecast has the same accuracy as the benchmark, and 1 when the forecast is perfect. This is typical for a skill measure. Note, however, that since forecasts are (almost) never perfect, the practical upper limit of a skill measure may be much smaller than 1.

For this particular example, the skill measure based on *MAE* is:

$$1 - \frac{MAE_f}{MAE_b} = 1 - \frac{2.1}{3.9} = 0.45$$

If MAE_f is the *Mean Squared Error* of the forecast, and MAE_b is the *Mean Squared Error* of the benchmark, another skill measure is effectively the reduction of variance, or

$$1 - \frac{MSE_f}{MSE_b}$$

For the example of 20 maximum temperature forecasts this is:

$$1 - \frac{6.9}{22.9} = 0.70$$

If the accuracy measure being used is the percent correct (of forecasts that are within an acceptable range of the observations), then another skill measure is:

$$\frac{PC_f - PC_b}{100\% - PC_b}$$

And for the example this is

$$\frac{0.60 - 0.35}{1 - 0.35} = 0.38$$

where the value of 0.38 means that the percent correct for the actual forecasts has gone 0.38 of the distance between the benchmark value of 35% and a perfect score of 100%.

4.3.2   Deterministic Forecast for Two Categories

Typical two category forecasts are:
- Yes or No for occurrence of precipitation
- Yes or No for occurrence of severe weather
- Rain versus snow.

As can be seen, such a forecast can usually be expressed as yes or no for an event. These are sometimes called forecasts of a dichotomous variable. The combination of forecasts and observations for a set of forecasts being verified can be put into a contingency table such as:

		Observed	
		Yes	No
Forecast	Yes	A	B
	No	C	D

To illustrate the use of this, suppose there has been a set of forecasts of whether or not there will be measurable precipitation "today". These could be spot forecasts that there would be greater than 0.1mm rain between 6 am and 6 pm during the daytime, together with observations from that spot on whether or not precipitation was measured.

The following table shows the results for this example, for a month's worth of data (31 days). Again, there are not many numbers here, but the purpose is to show the use of various scores. The numbers come from an example, which is shown in Appendix 1, together with all the reliability, accuracy and skill measures, which will now be described, and a few more.

		Observed	
		Yes	No
Forecast	Yes	*19*	*4*
	No	*2*	*6*

Reliability

The simplest bias measure is the ratio of the number of times the event was forecast over the number of times it was observed:

$$Bias = \frac{A + B}{A + C}$$

For the example this is

$$Bias = \frac{19 + 4}{19 + 2} = \frac{23}{21} = 1.10$$

so, for this particular case, precipitation is forecast 10% more often than it occurs. This may not necessarily be a major problem, particularly in forecasts of rare and severe events. Because the benefits of taking precautions against such events can be much higher than the cost of protecting against them, over-forecasting may in fact be a good thing. But for typical, ordinary events, it would be better if the Bias was around one.

Accuracy

The simplest accuracy measure is the percent correct for all the forecasts:

$$PC = \frac{A + D}{A + B + C + D}$$

For the example this is:

$$PC = \frac{19 + 6}{19 + 4 + 2 + 6} = \frac{25}{31} = 81\%$$

This particular measure may have quite high values for events that are either very rare or very common, so it needs to be interpreted with some care. However, for events such as this example, where precipitation occurred on 20 out of 31 days, it is quite a good measure.

If the event is a significant or a rare one, there may not actually be any count of the times when the event was neither forecast nor occurred. This could be the case, for example, with warnings of heavy rainfall. The numerous times when a warnings was not issued, and when heavy rain didn't occur, may not actually be counted.

In this case, it is common to use three measures of accuracy – *POD, FAR* and *CSI.*

The Probability of Detection (*POD*) is the proportion of times the event occurred that it was correctly forecast:

$$POD = \frac{A}{A + C}$$

For the example of rainfall forecasts this is:

$$POD = \frac{19}{19 + 2} = 0.90$$

The False Alarm Ratio (*FAR*) is the proportion of forecasts of the event that turned out to be false alarms:

$$FAR = \frac{B}{A + B}$$

The *FAR* for the example is:

$$FAR = \frac{4}{19 + 4} = 0.17$$

The Critical Success Index (*CSI*) is the ratio of the correct "yes" forecasts of the event to the sum of the correct forecasts, the false alarms, and the misses.

$$CSI = \frac{A}{A + B + C}$$

The *CSI* for the example is:

$$CSI = \frac{19}{19 + 4 + 2} = 0.76$$

Skill

It is possible to produce skill scores using the above measures of accuracy applied to both the forecasts as issued, and to some benchmark forecast. For example if there were some numerical guidance forecasts for which the *Critical Success Index* was CSI_b and the *CSI* for the issued forecasts was CSI_f then a possible skill score to use is:

$$\frac{CSI_f - CSI_b}{1 - CSI_b}$$

However, for this two-category case, one simple benchmark forecast is to use the sample frequency of events for the sample of forecasts being evaluated.

The sample frequency of "yes" events for the example is:

$$\frac{A + C}{A + B + C + D} = \frac{19 + 2}{19 + 4 + 2 + 6} = \frac{21}{31} = 0.68$$

If there was no relationship at all between a "yes" forecast and whether the event occurred (this is surely a benchmark forecast with no skill) then one would expect for each "yes" forecast that 68% of the time rain would happen,

and 32% of the time it would not. The same would apply for the "no" forecasts.

Thus, for this benchmark forecast, by pure chance one would expect the value of A in the contingency table to be the number of "yes" forecasts by the frequency of "yes" events:

$$CHA = (A + B) \times \frac{A + C}{A + B + C + D}$$

$$= (19 + 4) \times \frac{19 + 2}{19 + 4 + 2 + 6} = 23 \times \frac{21}{31} = 15.6$$

A common skill measure – the Heidke Skill Score – can then be computed as:

$$\text{Heidke Skill Score} = \frac{A - CHA}{A + B - CHA}$$

$$= \frac{19 - 15.6}{19 + 4 - 15.6} = \frac{3.4}{7.4} = 0.46$$

The Equitable Threat Score is a correction of the *CSI* to take into account *CHA*, and is defined as:

$$\text{Equitable Threat Score} = \frac{A - CHA}{A + B + C - CHA}$$

$$= \frac{19 - 15.6}{19 + 4 + 2 - 15.6} = \frac{3.4}{9.4} = 0.36$$

Finally, the often-used Hanssen and Kuipers (1965) score can be given as:

$$HKS = \frac{A}{A + C} + \frac{D}{B + D} - 1 = \frac{19}{19 + 2} + \frac{6}{4 + 6} - 1$$

$$= 0.90 + 0.60 - 1 = 0.50$$

This skill score also does not make *explicit* use of a benchmark forecast. However, a naïve forecast of always forecasting "yes", or always forecasting "no", will give a score of zero. Similarly, a naïve forecast with a random choice each time between yes and no will also have an expected score of zero. Positive values of the *HKS* therefore represent skill over these naïve forecasts, with a score of 1 for perfect forecasting.

4.3.3   Probabilistic Forecast for Two Categories

A probabilistic forecast for two categories can be treated as the probability that the first of them will occur, since the probability of the second category is one minus the probability of the first.

Suppose there are N probability forecasts p_i and corresponding observations for o_i for $i = 1...N$

Each forecast p_i will be in the range from 0 to 1, expressing the probability of a "yes". Each observation will be 0 if that event (the first category) did not occur, and 1 if the event did occur.

Data from the following table will be used as an example – it has just twenty probability forecasts, which is not enough to draw any conclusions, but it can be used to illustrate the various scores.

PROB FORECASTS		
Prob (p)	Obs (o)	(p-o)^2
0.25	0	0.06
0.95	1	0.00
1.00	1	0.00
0.85	1	0.02
0.05	0	0.00
0.15	0	0.02
0.25	0	0.06
0.15	0	0.02
0.10	0	0.01
0.50	0	0.25
0.85	1	0.02
0.75	0	0.56
0.15	0	0.02
0.65	0	0.42
1.00	1	0.00
0.75	1	0.06
0.10	0	0.01
0.85	1	0.02
0.65	1	0.12
0.10	0	0.01
Average: 0.51	0.40	0.09
		Brier Score

Reliability

A simple measure of reliability is the overall bias – the average of the forecast probabilities, divided by the frequency of occurrence:

$$Bias = \frac{\frac{1}{N} \sum_{i=1}^{N} p_i}{\frac{1}{N} \sum_{i=1}^{N} o_i}$$

For the example, the average forecast is 0.51 and the average observation is 0.40 so there is a *Bias* of 1.28 – over-forecasting of the probabilities.

Other reliability measures can be generated by dividing the forecast probabilities up into various ranges and seeing for each range what the actual frequency of occurrence was. For example, Reliability diagrams can be produced showing this information (see, for example, Wilks, 1995).

Accuracy

The most common accuracy measure for these kinds of forecasts is the Brier Score, (Brier, 1950) which is just the *Mean Squared Error* (*MSE*) for these particular forecasts and observations:

$$\text{Brier Score} = \frac{1}{N} \sum_{i=1}^{N} (p_i - o_i)^2$$

For this case, the Brier Score is 0.09.

Skill

If BS_f is the Brier Score for the forecast, and BS_b is the Brier Score for the benchmark forecast (in this case, climatology), then the Brier Skill Score can be expressed as:

$$\text{Brier Skill Score} \ = \ \frac{BS_b - BS_f}{BS_b} = 1 - \frac{BS_f}{BS_b}$$

Hence, this is like a reduction in variance (RV). It is in the form of a percentage improvement over the climatological benchmark, with a skill score of 1.0 for perfect forecasting.

In this case, BS_f is 0.09 and BS_b for a climatological probability of 0.40 is 0.24, so the Brier Skill Score is 0.63.

4.3.4 Deterministic Forecast for Multiple Categories

There are two different kinds of forecasts for multiple categories. One is where they are not ranked – there is no particular order to the categories. An example of this is where there may be a number of categories of precipitation type – for example, rain, snow, mixed precipitation, freezing rain. More commonly, the categories are ranked, and do have some kind of order. Examples include wind speeds in terms of Beaufort force rather than values, visibility categories, and precipitation in categories of increasing amounts.

To illustrate how this might work, suppose forecasts of rain are being made for a tropical location, where typically the weather might be in three categories – "dry", "showers", or "wet" (widespread showers or rain) for a 12 hour period from 6 am to 6 pm.

An observation of "dry" might correspond to no rain observed at the station; of "showers" if no rain was recorded at the station, but rain was reported in the area or thunder was heard; and of "wet" if rain was recorded at the station.

In the case of two categories (See Section 4.3.2) all the information about the verifications of a set of forecasts was obtained using a 2 by 2 contingency table. For m multiple categories, an m by m contingency table can also be used. (In the example of three categories of dry, showers and wet, m would be 3).

Such a table will be used for the remainder of this section. The elements of the contingency table will be taken as n_{ij}, which is the number of times that the observed category was i and the forecast category was j, where i and j are both in the range 1 to m.

The notation n_{*j} will be used for the total number of times that category j was forecast, no matter what was observed:

$$n_{*j} = \sum_{i=1}^{m} n_{ij}$$

and n_{i*} for the total number of times that category i was observed, no matter what category was forecast:

$$n_{i*} = \sum_{j=1}^{m} n_{ij}$$

Similarly, the total number N of forecasts being verified can also be given by n_{**} where:

$$n_{**} = \sum_{i=1}^{m} \sum_{j=1}^{m} n_{ij}$$

By way of example, for the three-category example of dry, showers and wet:

		Observed		
	Dry (1)	Showers (2)	Wet (3)	Sum of Forecasts
Dry (1)	n_{11}	n_{21}	n_{31}	n_{*1}
Showers (2)	n_{12}	n_{22}	n_{32}	n_{*2}
Wet (3)	n_{13}	n_{23}	n_{33}	n_{*3}
Sum of Observations	n_{1*}	n_{2*}	n_{3*}	n_{**}

(Forecast is the row label)

An example of some numbers in this 3 by 3 contingency table, which will be used for the scores, is:

		Observed		
	Dry (1)	Showers (2)	Wet (3)	Sum of Forecasts
Dry (1)	63	13	8	84
Showers (2)	15	45	30	90
Wet (3)	7	22	38	67
Sum of Observations	85	80	76	241

(Forecast is the row label)

Reliability

It is hard to have one overall number expressing reliability for the multiple category case. Instead, it is better to compare the number of times that each category was forecast with the number of times that it occurred.

The bias for forecast category j is then n_{*j}/n_{j*}.

In the three-category example, the bias for category 1 ("dry") is very close to one at 84/85. The "showers" category is slightly over-forecast, with a bias of 90/80, or 1.13. On the other hand, the "wet" category is slightly under-forecast, with a bias of 67/76, or 0.88.

Accuracy

The most commonly used accuracy measure for multiple categories is just the proportion correct – the sum of the diagonal

elements of the contingency table divided by the total number of forecasts. This is usually expressed as a percentage.

$$\frac{\sum\limits_{i=1}^{m} n_{ii}}{N}$$

Note that this accuracy score is equivalent to giving a mark of 1 for each of the exactly correct forecasts, zero for the ones where the correct category was not forecast, and then taking the overall accuracy score to be the average mark.

For the example, the sum of the diagonal elements is 146, and the total is 241, so the percent correct is 146/241 or 61%.

Other accuracy scores make use of the assumption that some credit should be given for a "near-miss" by one category, though the mark for being out by more than one category might be zero. Gordon (1982) developed a general methodology for these kinds of scores for accuracy and skill.

Skill

The simplest skill measures will involve a comparison between the accuracy of the actual forecasts and of some benchmark. Typical benchmark forecasts would be always to forecast the climatologically most likely category, or to randomly forecast a category based on the climatological frequency of the categories. Again, the climatology may be based on the sample itself. If PC_f is the percent correct for the forecasts, and PC_b the percent correct for the benchmark, then the skill is just:

$$\frac{PC_f - PC_b}{1 - PC_b}$$

For the example, suppose the benchmark forecast is to always forecast "showers", since this is the most common observed category. The result would be that the forecast was correct 80 times (the number of times the "showers" category was observed) and the percent correct for the benchmark is 80/241 or 33%.

For this case, the skill would then be:

$$\frac{0.61 - 0.33}{1 - 0.33} = 0.42$$

The skill scores proposed by Gordon (1982) provide a more direct and theoretically satisfying means of assessing skill, including confidence intervals on the score, though they may be less readily explained to the user community.

4.3.5   Probabilistic Forecast for Multiple Categories

For completeness, a description will now be given of probabilistic forecasts for more than multiple categories. However, the details and technicalities involved are beyond the primary purpose of this Technical Document, so the reader should refer to Stanski and Burrows (1989) for more details on these kinds of scores.

Suppose there are N forecasts, each of which has probabilities for m categories, p_{ij} for $i = 1$ and $j = 1...m$. The corresponding observations will be called o_{ij}, although in each case this will take on a value of 1 for the observed category and 0 for the other categories.

Reliability

As in the case of deterministic forecasts, reliability needs to be measured for each of the forecast categories, and can be done so using a bias for each category – the average of the forecasts probabilities for that category, divided by the frequency of occurrence:

$$Bias_j = \frac{\frac{1}{N}\sum\limits_{i=1}^{N} p_{ij}}{\frac{1}{N}\sum\limits_{i=1}^{N} o_{ij}}$$

More complex information on reliability can be assessed using reliability diagrams for each of the forecast categories, or by looking at the information in terms of observed categories.

Accuracy

Usually the categories are ranked, and the most common accuracy measure is the Ranked Probability Score (*RPS*) originally devised by Epstein (1969).

Using the above notation, the *RPS* for the individual forecast i is:

$$1 - \frac{1}{m-1}\left[\sum\limits_{j=1}^{m}\left(\sum\limits_{k=1}^{j} p_{ik} - \sum\limits_{k=1}^{j} o_{ik}\right)^2\right]$$

This has a range of 0 (bad) to 1 (a perfect forecast).

Skill

A skill score against a benchmark can be computed in the usual way, by comparing the Ranked Probability Score RPS_f for the forecast with RPS_b for the benchmark.

$$\frac{RPS_f - RPS_b}{1 - RPS_b}$$

4.3.6   Forecasts of Timing of Events

Discussion so far has concentrated on weather variables and categories. However, there is also increasing interest in the timing of events, rather than just whether or not they will occur.

It can be useful to collect and assess statistics on the forecast and observed time to:
- Start of precipitation
- End of precipitation
- Time of change of precipitation type (e.g., rain to snow, or snow to rain)
- Start of a severe event
- End of a severe event.

This verification information can be treated using the assessment measures for continuous weather variables (see Section 4.3.1). For example, if precipitation is forecast to start at 1500 and actually starts at 1100 then this can be treated as an error in the forecast of +4 hours, or 4 hours late. The

information can also be categorised – for example, by turning the timing error into categories of:

- Too early by 6 hours or more
- Too early by 2 to 6 hours
- About right (within 2 hours)
- Too late by 2 to 6 hours
- Too late by more than 6 hours.

Note that in order to accumulate statistics on timing errors, it is a given that the event was actually forecast and did actually occur. Use of categories may enable two more categories to be analysed – "forecast but not observed", and "observed but not forecast".

Another, different timing statistic is the lead-time for the occurrence of severe weather events. This is probably best just summarised, with statistics such as the average lead-time and distribution of times produced. Reliability would come separately from timing error calculations for the start of the event.

For example, the lead time for the start of gale forecast winds could be analysed. Skill could be assessed by comparing the lead-time for warnings as issued by the forecast office, with the lead-time based on an NWP model forecast.

4.3.7 Forecasts of the Location of Events

The discussion until now in this Chapter has assumed that the comparison is between forecast and observed weather variables for a place, or at most for some small region.

However, there are some types of forecast which explicitly predict the areal coverage and extent of an event. One example would be a warning of severe weather where the coverage is drawn on a map, or stated to occur over a number of counties. This forecast could be verified by dealing with each small region individually. It can also be verified as a whole, and the usual statistics produced.

The following diagram illustrates a typical situation where a forecast area of severe weather overlaps, but does not exactly match the area where severe weather was observed:
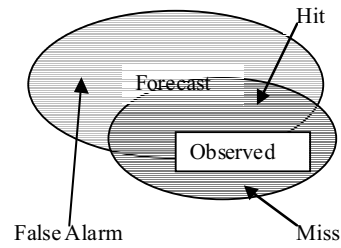
Reliability

The reliability for such a forecast can be assessed by comparing the average areal coverage of the forecast with the average areal coverage of the observed event.

Accuracy

The accuracy can be assessed by computing the Threat Score for each forecast or event, and then averaging for the verification period. The Threat Score is analogous to the Critical Success Index (see 4.3.2) for a series of two-category forecasts. In the diagram above, the area of overlap where the event was both observed and forecast can be considered to be a "hit". The area where the event was forecast but not observed can be considered to be a "false alarm". The area where the event was observed but not forecast can be considered to be a "miss".

$$TS = \frac{Area(Hit)}{Area(Hit) + Area(FalseAlarm) + Area(Miss)}$$

Skill

Skill can also be assessed analogously with the CSI, including use of the Equitable Threat Score (see 4.3.2). This needs some definition of what the "hit" area would be by pure chance. For example, if a country was divided up into 30 regions, and a particular severe weather event affected 10 of them, and a warning was for 15 regions to be affected by the event, then through pure chance one would have expected hits in 15*(10/30) regions, or 5 regions of the country.

Chapter 5
USER-BASED ASSESSMENT

5.1   INTRODUCTION

As stated in the introductory chapter to this Technical Document it is important to carry out ongoing performance assessment of public weather services to ensure that they are efficiently and effectively meeting the public's needs and contribute to longer term societal objectives. Managers need relevant information to appropriately lead and manage information, products, services and policy development. Most NMSs are now routinely required to report annually to central agencies and to meet these requirements they have to systematically collect and analyse performance information. This activity needs to be undertaken in the context of established measurement strategies and defined performance targets.

More recently some NMSs have developed "Service Charters" which detail their pledge of performance to their user communities – specifically, their country's citizens. These service charters provide a brief overview of the services provided, a commitment of performance against specific targets (both purely verification oriented and user-based), and a commitment to consult and identify a means by which the citizens may register their concerns. These service charters can be perceived as being the NMS's contract with the citizen. As such they have become an important component of the performance measurement strategy of the Services that have adopted them. They represent a public commitment to measure performance and to report on it according to publicised commitments and targets.

To facilitate management's approach to adopting a results-based, integrated strategy for performance measurement, review and reporting, it is useful to develop a performance framework and a system that is comprehensive and timely, and that balances expectations with the NMS's capacities. The framework should also reflect the ability to manipulate, and the dependent relationship between, the various dimensions of the NMS's capacities and the resultant consequences of doing so. It can serve as an evolving descriptive management tool to meet the needs of the management team. A performance measurement framework attempts to define the linkages between decisions on resource utilisation and results.

The basic logic model is illustrated in Figure 1. The NMS achieves its objectives through managing its programmes and determining its priorities. The NMS does this by manipulating the mix in its capacities according to some optimal balance. The activities and outputs reach a target client group (community of interest) either directly or with the aid of co-delivery partners and stakeholders as determined by the service delivery capacity mix chosen. As a result of the activities and outputs, the community of interest group exhibits a behavioural response of some sort, and immediate impacts can occur. Over the longer term, a modified behavioural pattern can emerge which can lead to more extensive and consequential impacts that, if the program is performing well, may be causally linked to the NMS's long term objectives. In theory, indicators are developed to measure performance in each of these areas and sources of information are identified.
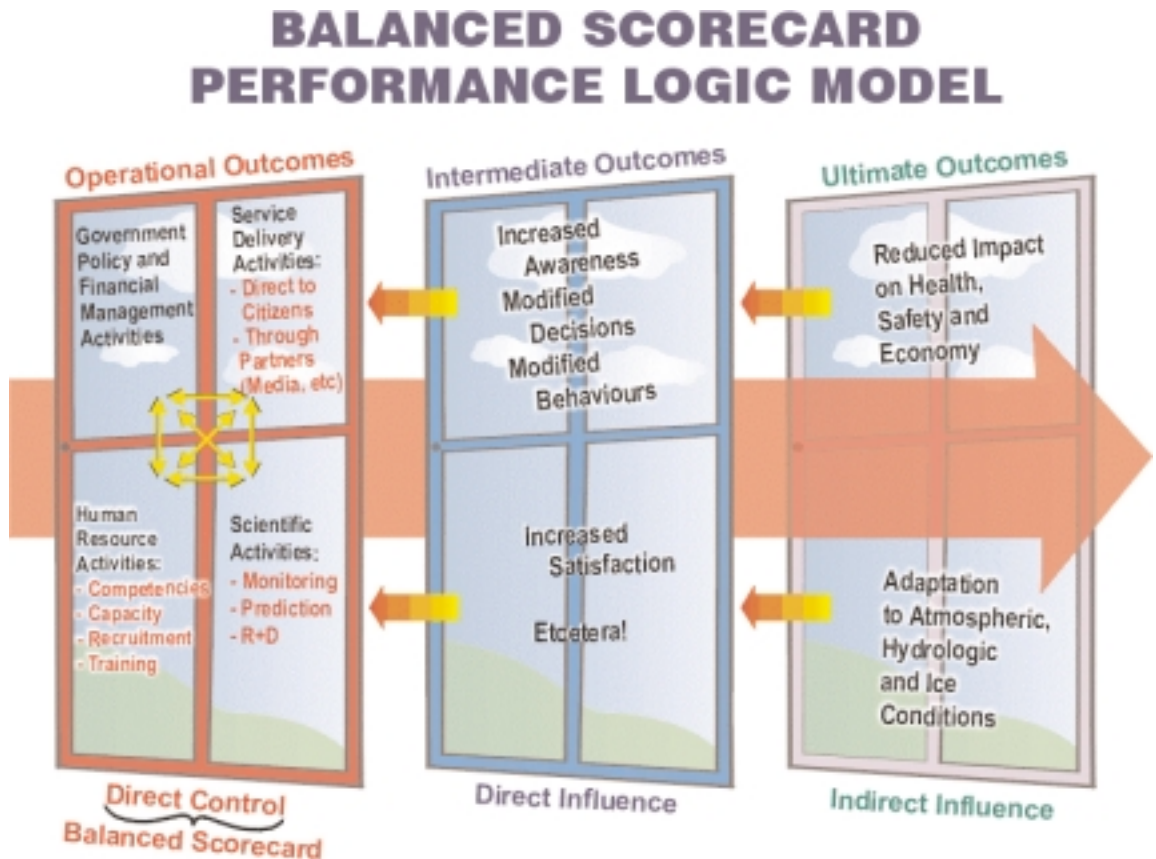

Figure 1

The generic ultimate desired result of a NMS' activities can be described as that of reduced impact of weather and related hazards on health, safety, the economy and the environment. A NMS can only have an indirect influence over such an ultimate outcome of the delivery of its services. A more direct influence is in the area of decision-making and behavioural changes, say, in the form of avoidance of the risks involved or adaptation to them, which come from increased awareness. The questions here are externally focused and deal with client satisfaction, and achievement of intermediate results such as building awareness, improving capacity, and influencing behaviour and actions. A NMS has more direct control over how it manages its human resources, scientific activities, service delivery activities, and its government policy and financial management strategies. The associated questions here focus on internal issues such as how the organization manages these dimensions, with what emphases, trade-offs, etc. Performance needs to be measured in each of these areas with due consideration to their interdependence.

The acceptance of the NMS's products by the public and other users depends on a number of factors. Scientific accuracy is just one of those factors. User-based assessment is about measuring perceptions on a matrix of dimensions important to specific user communities and amongst a diversity of user communities. These perceptions include those about requirements, accessibility, availability, accuracy, timeliness, utility, comprehension, language, sufficiency, and packaging. The user communities range from the individual citizen using the products to make personal decisions, to the media organizations essential for the communication of the product, to government agencies funding the production and delivery of those products. The health of the NMS depends on the perceptions from the full spectrum of these users. This chapter focuses on the characteristics of user-based assessment and methodologies employed. The objective is to ultimately measure performance from the user perspective. This can be done by achieving an understanding on the "logic" underlying the approach to achieving specific results, identifying a limited set of indicators that responds to performance questions for each result, and implementing a data collection plan. Starting with key results (ultimate outcomes) the NMS needs to get consensus and clarity on user strategy (key activities/outputs), define the target groups and the desired influences/changes (intermediate outcomes sought) and focus on the gaps in logic. Such a broad framework can be further broken down through more detailed articulation of specific outcomes and of the nature of influence.

The biggest challenge of user-based assessment is to translate vaguely formulated concerns around ultimate and intermediate outcomes into a well-conceptualised and methodologically sound study. There is a requirement to specify what information is needed, from whom or where the information should be obtained, and how the information will be used. Decisions must be made on how to obtain data from the three possible sources:

1) Documents, records or other existing information
2) Observation, e.g., of actual behaviour of people
3) Questioning.

5.1.1 Characteristics

User-based assessments are focused around the ability to obtain information on specific characteristics of interest through a variety of direct methods such as surveys, focus groups, public opinion monitoring, feedback and response mechanisms, consultations such as users' meetings and workshops, and the collection of anecdotal information. On their own, each of these methods may produce information which is subjective and of questionable reliability. However, taken as a whole, a consistent picture often emerges which is credible. They are the only effective means by which information can be gathered on needs, expectations, satisfaction, etc. More recently, they have also been demonstrated as effective means for getting at the economic value of information such as weather information and forecasts.

5.1.1.1 Subjective

Any perception data is by its very nature subjective. Responding to a question involves four distinct processes:
(1) Respondents must first understand the question.
(2) They must then search their memories to retrieve the requested information.
(3) After retrieving the information, they must think about what the answer to the question might be and how much of that answer they are willing to reveal.
(4) Only then do they communicate an answer to the question.
Cognitive methods provide the means to examine respondents' thought processes as they answer questions. Cognitive testing methods include
• Observation of respondents
• Think aloud interviews
• Focus groups
• Paraphrasing
• Confidence rating.
They are used to find out whether or not respondents understand what the questions mean. In this way, cognitive methods help assess the validity of questions, and identify potential sources of measurement error. Respondents often do not understand the words and concepts the same way as researchers. The researcher must relate to the respondents by using their language and ways of expressing concepts.

5.1.1.2 Perception as Reality

Gauging the perceptions of citizens, direct clients, stakeholders and government agencies is an important component of service evaluation, and all of these 'user communities" must be included in the assessment. The goal of service evaluation is to identify users' needs and to measure the acceptance of the services provided from such dimensions as expectations, understanding, importance, satisfaction, utility, etc. Data on perceptions relative to such parameters is collected through a variety of means.

5.1.1.3 Dimensions: Requirements, Expectations, Understanding, Importance, Satisfaction, Utility, etc.

As previously stated, the perceptions assessed include those about requirements, accessibility, availability, accuracy, timeliness, utility, comprehension, language, sufficiency, and packaging amongst others. Classically, in the design and development of products and services one starts with the assessment of user requirements. That is, what are the needs of the spectrum of end users (the public, stakeholder communities, funding agencies) from the spread of possible services that the NMS has the capacity to provide?

This effort benefits from gaining an understanding of user processes – that is, an understanding of how the information is used in the activity to which it is applied. Frequently, expectations do not line up with actual needs, in which case two alternative paths could be pursued. If the end-user cannot be convinced of the faulty expectations then the survival strategy may be to target on those expectations. In other words, try to provide the information they want, even if you know that it may not be the best information for their purposes. Fortunately, most often with the increasing sophistication of the end-user the result is a realignment of expectations with needs.

A complementary activity with pure user-based assessment is thus that of increasing awareness and user education. The theory is that this process, with iteration, yields improved knowledge of the spread of requirements (stated and implied) that then can be translated into the design of a set of meteorological products and services that cover the degree of requirements that is within the capacity of the NMS to provide. This results in the development of new products and services, and/or the adaptation or refinement of existing products and services, or even in dropping services that are no long needed, to better match the evolving requirements.

5.1.1.4 Economic Value Assessment

Increasingly NMSs are under some pressure to reduce costs of operation and to justify any major upgrades of their services and equipment based on a detailed benefit-cost analysis. NMSs are interested in demonstrating the economic and social benefits of services they provides to the public, industries and organizations. As illustrated in the performance logic model (Figure 1) benefits to society as a whole are commonly perceived as an ultimate outcome of the provision of meteorological services.

For the purpose of this discussion public weather services are generally considered non-rival (if someone uses the service it doesn't stop others from using it) and non-exclusive non-market goods and services. While some services are rivalrous, such as limited capacity telephone-based services, these kinds of services are generally being de-emphasised or commercialised by NMSs.

A variety of research methods in applied economics (environmental, resource, production, information, risk and uncertainty, welfare, etc) can be applied. One of the techniques being increasingly employed is that of contingent valuation whereby respondents, through an iterative process, are asked to indicate their willingness to pay a suggested amount to have access to the services versus having that service withdrawn.

The valuation techniques can be broadly described as being either production based or demand based. The former involves the modelling of the production process while in the latter case direct inferences are made as to the value of non-market services such as public weather services. Economic value assessments range from measuring the value of certain forecast elements to that of estimating the value attributable to the provision of the full set of national services. Reported benefit-cost ratios have been reported as being 2:1 to well over 10:1. Economic value assessments also can be used to determine the justifiability of making investments in research and development into improvements in forecast accuracy. Additionally such assessments can be used to compare the effectiveness of various meteorological service delivery systems. With some measured success some of these techniques have been used to impute the 'social' or non-economic benefits derived from the use of public weather services. Further discussion on the methodologies for the undertaking of such assessments appears further on in this document.

5.2 GUIDING PRINCIPLES FOR METHODOLOGY

There is a need for user-based information for decision-making purposes by individuals, whether office managers or the most senior executives of the NMS. The information is used for day-to-day programme delivery management as well as for longer-term vision and strategic planning. While the information gathered may serve the objectives at a variety of levels within an organization, often the methodology chosen must be specific to the objectives at the organizational level.

5.2.1 Long and Shorter Term Strategic/Tactical Decision Context

The circumstances of planning have changed and the complexities of managing have increased in recent years. The NMS's organisational and decision-making structures have changed. The governmental and departmental planning systems have created new processes and products. The focus on value for money and making the NMS's funds go further has sharpened as budgets have significantly decreased with governmental budget reduction exercises. Performance management has taken on greater prominence with emphasis on frameworks, concrete measures, and continuous improvement. At the same time, in several domains, the programme has expanded from an initial narrow focus on weather, migrating through the larger domain of atmospheric change, to a broader focus on environmental prediction.

User-based assessment needs to be tied closely to performance management, planning and reporting requirements and the links to both operations and long-term strategic results should be clear. A more proactive role can be played by:

- Obtaining direction from senior management on planned user-based assessments to ensure that these assessments will be useful and that there are resources and the management will to take follow-up action once the findings and recommendations are presented
- Working with the organisational units, within the NMS, responsible for implementation of program changes, to advise them of the findings and facilitate follow-up action
- Tracking follow-up actions and reporting back to senior management. Senior management support in terms of commitment and resources to implement change is a key success factor.

Follow-up is essential – if not done, user-based assessment research will have little value.

The kinds of decisions that benefit from the user-based assessment process range from those pertaining to the initiation, continuance or modification of major programmes to specific product lines or programme elements and delivery mechanisms. Within this spectrum is included the range of decision activities as diverse as that regarding investments in research and development, technology for automation, human resource training, and public education or awareness campaigns. Ultimately, within the resource context of the NMS, policies on detailed levels of service can be established.

5.2.2   Multi-year User-based Assessment Strategy

A plan must follow a development process, which accommodates the funding and reporting context that the NMSs find themselves in, and have the following characteristics:
- A limited, manageable number of priorities that reflect the needs of the programme
- A schedule of user-based assessments that supports these priorities while being flexible enough to meet needs arising from unpredictable or opportunistic circumstances
- An approach to communicating findings that promotes sharing information and the development or improvement of products and services.

In developing the schedule of user-based assessment, the areas of research are selected on the basis of programme need, risk management, and commitments in business plans, management frameworks, and performance frameworks. In this multi-year strategy for user-based assessment it is important to cover both product lines and delivery mechanisms and to use consistent questions over years for proper trend-line analysis. Performance measurement, after all, is about the change over time as opposed to the measurement of the state of affairs at a give point in time.

5.2.3   Need to Know Why it Should be Done

The first task in planning a user-based assessment is to specify the objectives as thoroughly as possible. The key to this exercise is to come up with clearly defined concepts and terms. Once the basic objectives have been broken down and defined, the researcher can then proceed to develop operational definitions which indicate who or what is to be observed and what is to be measured. Once operational definitions are developed the researcher can specify the data requirements and decide upon the level of error that is acceptable. Finally, the statement of objectives should indicate the purpose, the areas covered, the kinds of results expected, the users as well as the uses of the data, and the level of accuracy that is desired.

Essentially, a **survey** involves the collection of information about characteristics of interest from some units of a population using well-defined concepts, methods and procedures, and the compilation of such information into a useful summary form. The collection of such information from all units of a population would constitute a census. Surveys are carried out for either one of two purposes: descriptive or analytical. The main purpose of descriptive survey is to estimate certain characteristics or attributes of a population – e.g., awareness of a particular meteorological service. Analytical surveys are generally concerned with testing statistical hypotheses or exploring relationships among the characteristics of a population. An example of an analytical survey would be one that determines whether there is a change in protective behaviours following the introduction of an Ultra Violet Index programme.

There can be many reasons for undertaking a user-based assessment by an NMS. These can include the checking of perceptions against expectations, tracking of trends, seeking feedback to improve existing services, determining requirement for new or different services, assessing perceived effectiveness of overall programme, and the identifying areas where actions can be taken. An NMS's "Service Charter" may dictate the requirement to routinely publish information regarding such dimensions as user satisfaction. Such information can be derived from the administration of a re-useable tracking survey. Subject area surveys can be used to elicit information feedback for the improvement of certain specific surveys or for determining the requirement for new or different services. Large comprehensive surveys can be used for gauging the overall effectiveness of the NMS's total programme.

5.2.4   Credibility and Transparency

There are many considerations that come to mind when wrestling with the concepts of credibility and transparency for user-based assessment. Comments made above, regarding an overall performance management framework and strategy, certainly apply. User-based assessment is an effective and essential component of an organization's "balanced scorecard" giving a comprehensive picture of its health and effectiveness. The adoption of a rigorous approach or methodology based on established theory and practices is essential. The adherence to a multi-year user-based assessment strategy facilitates a co-ordinated and structured approach. Even such simple precepts as undertaking fewer but well planned surveys, focus groups, etc., rather than a large mixture of disconnected ones and following a consistent approach to track trends help. Finally, publicizing the changes triggered by the assessment enhances credibility and transparency.

5.2.4.1 Statistical Significance Issues

With regard to public opinion or stakeholder surveys a focus on sampling and on sampling errors and accuracy can head off credibility and transparency problems.

5.2.4.1.1 Sampling

For a specific subject area relative to the programme of an NMS to be examined one of the first decisions to be made is whether to undertake a *sample survey* or a *census survey*. A *census survey* refers to the collection of information about characteristics of interest from all units of a population. An NMS may want to determine certain characteristics about the redistribution of meteorological products by their domestic media. An NMS may want to determine what ice forecasting services high Arctic marine operators would like to receive. In such cases, for most countries, a census survey may be more appropriate given the very small population under study.

A *sample survey* refers to the collection of information about characteristics of interest from only a part of the population. A survey of the general population's awareness and understanding of a wind-chill programme would be a valid use of a sample survey.

A sample survey is cheaper to do than a census survey. Sampling also reduces data collection and processing time. Sample surveys allow more selective recruiting of interviewers, more extensive training programmes and closer supervision. As well, the smaller scale of operations allows for more extensive follow-up of non-respondents and for a higher level of quality control for such data processing activities as coding and data capture. For these reasons sample surveys can be more accurate than their census counterparts. In some cases where highly trained personnel or specialized equipment is required it would be difficult and expensive to consider a census. Sample surveys inconvenience fewer people meaning reduced respondent burden.

The *target population* is the set to which the survey results are to apply; about which information is sought; to which the sample is intended to represent; and about which one wishes to make inferences based on data collected from a sample. A population has definable characteristics, a specific geographic location and a time period under consideration. The *survey population* is the population that is actually covered which may be different from the target population for practical reasons. For example, in a national survey remote locations are frequently excluded because they are too difficult or costly to enumerate. When a survey population is chosen which differs from the target population, it is necessary to be aware that gap exists between the two populations and recognise that conclusions based on the survey results apply only to the survey population.

Samples can be probabilistic and non-probabilistic.

In non-probability sampling elements are chosen in an arbitrary manner such that there is no way of determining the probability of any one element being included in the sample thus there is no assurance that every element has a chance of being included.

In probability sampling all within the population have a non-zero chance of being selected and inferences are made about the entire population that the sample represents. Probability sampling methods range from simple random selection of members from the population to complex sampling strategies (random, systematic, stratification, and multi-stage).

Stratification is the most common amongst these methods. Stratification is the process of dividing the population into relatively homogeneous groups called strata, and then selecting independent samples. Stratification variables may be geographic or non-geographic (e.g. gender, income, industry, occupation). Reasons for stratification include the desire to acquire estimates at the stratum level. Each stratum requires an adequate sub-sample size to ensure that valid results can be derived that are particular to that stratum.

In random sampling each unit in the population has an equal chance of being included in the sample.

In systematic sampling, units from a list are selected using a selection interval (K), so that every Kth element on the list, following a random start between 1 and K, is included in the sample. If the population size is M and the desired sample size is "n", then K=M/n. Thus systematic sampling requires a sampling interval and a random start.

Multi-stage sampling refers to a process of selecting a sample in two or more successive stages. For the two stage sampling case a number of first stage units are selected - e.g. selected communities, from which second stage units are selected from within the larger units have already been selected, e.g. households within the selected communities. The probability of being selected is $P = P_1 \times P_2$ for the two stage sampling case where P_1 and P_2 represent the probability of being included in the sample at the respective stages.

5.2.4.1.2 Sample Errors and Accuracy

Both sampling and non-sampling errors affect the accuracy of survey results.

Sources of non-sampling errors include non-response, difficulties in establishing precise operational definitions, incorrect information provided by respondents, incorrect interpretation of questions by respondents, and mistakes in processing operations.

Sample error is the difference between the results of a sample estimate and a census, i.e., the population. The size of the sampling error generally decreases as the sampling size increases. The extent of the sampling error also depends on the variability of the characteristics of interest in the population, the sample design, and the estimate method. Thus the size of the sample, population variability, sample design, and the estimation method, are all included as sources of sampling error. The sampling error can be reduced through the development of an efficient sampling plan, where proper use is made of available information in developing the sample design and estimation procedure.

Accuracy refers to the difference between a survey result for a characteristic and the true value of that characteristic of the population. Precision (or reliability) is a measure of the closeness of sample estimates to the results of a census (or

100% enumeration of the population) that is undertaken under identical conditions. The greater the variability in the population, the larger the sample size needed to obtain the specified level of reliability. Complex sampling procedures usually increase the margin of error as they increase the possible sources of errors. Increasing the sample size will lower the margin of error due to non-response but the bias resulting from non-response is not reduced. With respect to the characteristics of interest in the survey, the non-respondents may be different from the respondents.

Confidence interval statements are commonly provided with published survey results. A 95% confidence interval can be described as follows: if sampling is repeated indefinitely with each sample leading to a new confidence interval, then in 95% of the samples the interval will cover the true population value. The size of the confidence interval is usually indicated by the margin of error. For example, if the estimate is 50% and the margin of error is 3% either way (below and above 50%) then the confidence interval is that the "true" percentage falls somewhere between 47% and 53%, 19 in 20 times (i.e., 95% of the time).

The confidence interval does not take into account a margin of additional error that may result from practical difficulties that are involved in conducting a survey. The sources of this type of error include, for example, the way the questions are worded, respondents misunderstanding the questions or answering incorrectly, and non-response. The acceptable level of reliability depends on the estimate under consideration and the intended use of the data, that is, the acceptable level of reliability depends on the level of accuracy required for a particular application. What may be an acceptable margin of error for one estimate may differ from that felt suitable for another estimate.

The determination of sample size involves a process of making practical choices and trade-offs among the conflicting requirements of precision, cost, timeliness and operational feasibility.

5.2.4.2   *Collaboration with Other Relevant Authorities is Desirable*

Working with others can achieve synergies and economies of scale. The process of developing a plan and sharing information on intentions will be more inclusive increasing co-operation, communication, and co-ordination of efforts. Teaming up with others may yield mutual benefits such as reduced costs, increased internal communication, and new ideas. To be successful, this requires communication by all parties. Organisations in the private or not-for-profit sectors can be approached for help in reaching their communities. They may be willing to provide funding or service in kind. Examples include approaches such as co-operation with community support and advocacy organisations for deaf, deafened, and hard of hearing clients.

One of the most common forms of "collaboration" is the use of omnibus surveys that are usually conducted by telephone. In the case of omnibus surveys the NMS buys a portion of a larger survey that may cover several clients. Omnibus surveys are questionnaires consisting of several modules or sections, each dealing with a different topic and each conducted for a separate organization. Organizations are charged on the basis of their level of participation in the omnibus survey. These surveys are routine surveys according to a specific schedule. Frequently, a private survey company will attempt to accommodate the NMS client by pairing the meteorology portion with another on a similar (environmental?) theme. "Piggy-backing" questions on an omnibus has the effect of sharing the cost of the undertaking. They are useful for a research effort where there are only a few questions to be asked. These surveys typically use classification data such as age, gender, region, community size, family income, occupation, education, and mother tongue.

On occasion, a survey company will try to set up a larger one-time survey effort by inviting certain like-minded organizations. These can also provide opportunities for cost reduction. More frequently, with the recent increase in interdisciplinary activities with others in the media, environment and health fields, collaborative efforts result in cross-disciplinary user-based assessments and these assessments often take the form of surveys. One such survey was the Canadian *National Survey on Sun Exposure and Protective Behaviour* into which a section on the Ultra-Violet (UV) Index was added. Major surveys of this nature may be administered every five years to establish trend-line information.

With the NMSs moving towards the provision of broader weather and environmental prediction services there are increasing opportunities for collaborative user-based assessment efforts. Examples of these include air quality and smog forecasting programmes precipitating the need for joint assessment activities between various levels of government and sometimes non-governmental environmental organizations. As a minimum, in-kind resources are offered but more recently actual financial support is provided. With expansion in the road weather forecasting area there could be possibilities for similar collaborative efforts with the transportation sector and other levels of government.

5.2.5   Additional Principles of User-based Assessment Design

5.2.5.1   *Use of Professional Expertise and Independent Administration Authority*

The satisfaction of credibility and transparency concerns is facilitated by the use of external independent expertise as an input to the design and for the administration of the user-based assessments. The use of external accredited consultation expertise can facilitate the free and honest flow of ideas and concerns. Focus group facilitators are essential for the creation of the desired information discussion environment when considering the characteristics of interest to the NMS funding the study. The expertise of a private survey firm, a dedicated government body with the assigned responsibility and appropriate skills, or of an academic (University) professional adds value to the design of a survey instrument. Such expertise and at-arms-length objective

positioning is usually essential for the administration of a survey. Such external expertise will assist in the perception of credibility and in the attainment of statistically valid results from the perspectives of sample size and geographical and geopolitical representation. Indeed, it may be a formal requirement for performance pledge / charter or of quality assurance system to use such expertise.

5.2.5.2   Lack of Professional Advice or Availability of an Independent Capacity Should Not Stop Assessments From Being Done

Although it would be best for an NMS's to use professional advice or some independent capacity, if these are not available, user-based assessments should still be done. It is essential to measure certain basic end-users' understandings and reactions to the services provided. The use of some "best practice" examples of other NMSs providing similar programmes can help. Adaptation of these by in-house staff, and in-house staff administration of such assessments can yield very useful information that can assist in the management and planning of the NMS.

5.2.5.3   Dry Run or Pilot Test the Assessment Instrument

Careful planning and pre-testing the survey or focus group instrument or consultation strategy is essential. Pre-testing will often reveal information on the ability of the proposed question set to deliver on the objective of the survey. Misunderstanding of specific questions and unexpected responses can be detected through such pre-testing. Depending on the objective of the user assessment, pre-testing may reveal the requirement for additional or different questions, faulty skip patterns (skipping to the wrong or unintended subsequent question based on the response to a question just asked), etc. Pre-testing in a demand-based economic valuation exercise will help set the willingness-to-pay (WTP) cost amounts so that eventual responses ideally approximate a normal distribution about that WTP estimate.

5.2.5.4   Information Storage

Determination of and adoption of certain practices for information storage aspects of user-based assessments are essential for both current and future use of the results. Various types of information and sources result from user-based assessments. Audio or video recordings of consultation, workshop and focus group proceedings can be made for future retrieval and "mining" of the information. These recordings should be kept in a safe place. Written transcripts or proceedings or reports of such events can be used in a similar fashion. Reports on analyses of surveys can have a similar use. These should be kept and made accessible in both hard copy and electronic form. Special consideration should be given to the electronic storage of raw survey data in a standardised format be it in a simple flat file, spreadsheet, or statistical format such as SPSS or SAS (commonly used proprietary software packages, available commercially, used for scientific and survey applications).

5.2.6   Communication of Information

To be effective and worth the expenditure of the resources involved the information must be communicated and appropriately used internally within the NMS as well as externally to clients and stakeholders.

5.2.6.1   Accessibility Within the NMS

Increasing the access to user-based assessment results within the NMS is important. Use of this information in both the long and shorter-term strategic/tactical decision context has been discussed above. The results of user-based assessment research need to be made available to managers and employees if they are to be worthwhile. A greater awareness of what has already been done elsewhere could avoid possible duplication. The results could be used by others in various activities such as planning, risk management, briefing note preparation, and tracking issues.

5.2.6.2   Interpretation Reports for Internal and External Consumption

Reporting on the results of user-based assessments can take a variety of forms.

There are the standard statistical reports such as produced by public opinion research firms or in house staff.

Public or stakeholder consultation reports usually summarize the results of the consultation activity along with reporting of the actual dialogue that has taken place. If the consultation process used was that of a workshop then a full proceedings of the workshop is frequently published.

For assessments done by way of focus groups these usually consist of consensus remarks reinforced by some of the dialogue, notable for capturing particular points of view, all done according to the structure implicit in the focus group's questionnaire.

Reports on public opinion surveys usually provide a statistical analysis of the results on a question by question basis within each section. These statistical results can include a variety of descriptive statistics including frequencies and cross-tabulations, custom tables including multiple response tables and tables of frequencies, comparative means, perhaps some linear models, correlations or regressions, perhaps some classification or cluster analysis, and some multiple response analysis. Results are frequently presented in the form of graphical representations (bar, line, pie, area, scatter, etc.). In the case of surveys repeated according to a prescribed schedule, time series analysis and trends analyses may be reported upon. These analytical reports are used by staff to generate issue specific or general summary reports for senior management or for external parties. These summary reports take a variety of forms depending on the purpose intended and the audience.

5.2.6.3 Archive, Publish, Use as Appropriate for Promotion (and Education)

Since user-based assessment is quite costly, it is important to maintain both the reports and raw data in a variety of media, with backup copies, for future use and possible reanalysis. The media range from hard copy to electronic to video or audio. The material can be used for distribution to a variety of users ranging from management for decision making purposes, to staff for internal awareness, to funding authorities for resource justification, to the public or stakeholders for end-user awareness and education, to regulatory bodies for the attainment of approvals, to central agencies to satisfy reporting requirements, etc. It is important that the data is properly indexed and easily retrievable.

5.2.6.4 Targets for Communication of Results

The communication of the results to staff and management will assist in the evolution to a more client-centred organization that can lead to improvement in products, production, efficiency and delivery or even end-user awareness and education thrusts. Communication upwards through higher levels of management will assist in the longer term strategic planning and management for the NMS.

Communication to central agencies may be a defined requirement but can also be used as a justification for resources (current and additional). Communication of the results externally may have the effect of modifying certain practices, such as related to safety, or may encourage or accelerate the development of new services or products within the private sector (Weatheradio units, special services, etc.). Communication to the general public can have the effect of increasing awareness and credibility of the NMS and its offerings.

5.3 METHODS

The information "universe" for assessments is only partially measured by any research technique. Qualitative data can convey detailed information from a few respondents, while quantitative data come from restricted information from many respondents. Qualitative data, such as comes from a few in-person interviews (e.g., data from a rambling in-the-street interview by the media) or focus groups exploring a wide range of dimensions of a particular topic, yield information through restricted observations on a massive domain. Quantitative data, such as that from a sample survey asking a few rigidly structured questions of many people, yield information through a mass of observations on a restricted domain (e.g., data from large sample survey on satisfaction with temperature forecasts). Compared to quantitative data, the meaning of qualitative data is more likely decided after data collection.

The general characteristics of qualitative and quantitative methods are summarized in the table below.

Qualitative techniques are employed when rich contextual program description or new/refined program theory is needed or variations in implementation or process are to be assessed. When causal attribution, incremental effects or resource expenditure assessments are the objective, quantitative methods are more appropriate.

5.3.1 Non-Survey User-based Assessments

While much of the attention in this chapter is given to the design, development and administration of formal surveys, a quantitative technique, it is not the only vehicle for user-based assessment and frequently it is not the best vehicle for specific circumstances. Formal audits, whether mandated or self-imposed, can yield useful information and have the effect of aligning the NMS with overall governmental initiatives.

Focus groups, a qualitative technique, are a very popular means of gathering initial information that may be later used in a formal survey or of acquiring greater in-depth understanding of a particular dimension after a formal survey. Most governments and major corporations monitor their public image through a variety of means and many have formal feedback and response mechanisms. Public and stakeholder consultations are standard means of obtaining input on NMS policies and issues. Most NMSs will undertake operational performance reviews following major meteorological events to assess the effectiveness of their systems. Finally, for more than historical purposes, NMSs collect anecdotal information to be used strategically.

5.3.1.1 Formal Audits

Formal audits, whether mandated or self-imposed, can yield useful information on the operation and effectiveness of the

Dimension	Method	
	Qualitative	Quantitative
Intent/Purpose	Discovery of theory, understanding of phenomena under study	Verification of theory, statistical prediction
Assumption re: origin of meaning	Socially constructed and conferred on objects and acts	Inherent in objects and acts
Scope/Nature of investigation	Holistic, rich in context, emphasizes interactions	Particularistic, guided by program objectives
Sampling	Revealing in nature, population inferences cannot be drawn	Probability, population inferences can be drawn
Data gathering	Semi-structured or unstructured (open-ended) response options, observation	Fixed response options
Analytical techniques	Inductive	Deductive
Generalizing to population	Invalid	Valid
Data collection skills required	On-the-fly processing required	Rigid script

NMS. These also can have the effect of aligning the NMS with overall governmental initiatives. They involve independent auditing of the NMS and its services by an independent party (e.g., government audit agency or consulting company) according to some established or agreed-to criteria. They are usually undertaken according to an established schedule for all or part of the NMS's range of accountabilities. They identify performance improvements achieved and those not adequately achieved and for these they can specify some subsequent reporting of actions taken and associated results on a later date. These audits may be part of an overall quality management system at the service level or across government and its agencies. These should be seen as an opportunity to learn and improve, and perhaps to justify requirement for resources.

5.3.1.2   Focus Groups

Focus groups are a very popular qualitative means of gathering initial information that may be later used in a formal survey. Focus groups are also useful when developing new products or initiatives, to explore needs, understanding and preferences. The user-based assessment process may actually end with the focus groups. An example of one such focus group is one that considers the use and understanding of specific meteorological terminology such as "probability of precipitation".

The focus group participants are selected through a variety of means. Frequently, the NMS wants to collect some qualitative information from certain sectors of society such as professional categories (e.g., mariners), different levels of education, gender, family status (e.g., mothers with children potentially exposed to ultra violet radiation), etc. The contracted or in-house authority could select from a known client list or more or less randomly from sources like phone books to identify potential participants.

Focus group sessions usually last from one to two hours and are typically comprised of 8 to 12 participants seated comfortably around a boardroom type table in a specially designed room.

Frequently, observers can observe the proceedings from behind one-way glass in an adjacent room or via a TV monitor. The proceedings are usually recorded via audio or videotaping but the participants are made aware of this recording activity. The focus group session is usually conducted by a professional facilitator who has been brought up to speed on the subject area. It is critically important that interruptions are avoided and that interference in the conduct of the focus group, by Service personnel, does not occur.

Careful attention must be given to the development of the focus group questionnaire or guide with the facilitator. This is where possible areas of misunderstandings are identified and clarified to the facilitator so that he or she may respond appropriately within the focus group. Unlike a formal survey where a spread of related subjects can be addressed, a maximum of a couple of issues can be addressed by a focus group. For proper treatment of the characteristic of interest several focus group sessions, geographically separated, are desirable.

Care should be taken not to draw statistical inferences from the focus group sessions. The samples are too small for that, but the results can provide useful input to the design of questions for a formal survey. Qualitative data, such as comes from focus groups, may be summarised and synthesized using systematic techniques. Before coding can begin, data often have to be cleaned (i.e., non-relevant or non-codable {incapable of being categorized} material identified and removed) and unitised (broken down into codable units). Meaning is assigned to observations by finding patterns through the processes of integration, differentiation and ordering frequently using a matrix approach.

A formal report on the conclusions of the focus group session is a standard requirement. These reports usually summarize both the central tendencies and significant variations and also make extensive use of verbatim quotes from respondents to illustrate key points.

5.3.1.3   Monitoring Public Opinion and Direct Feedback and Response (Complaints, Compliments, Suggestions) Mechanisms

Many government organizations and major corporations monitor their public image through a variety of means and many have formal feedback and response mechanisms. Many NMSs, or their parent organizations, have designated staff that monitor electronic and printed media reports or purchase media monitoring services for that purpose. Media reports frequently precipitate media interviews of Service personnel that generate further media reports. Just two examples of such occurrences in Canada were the publicity surrounding the windchill issue that spawned over 100 media interviews, and the Ice Storm in January 1998 that generated about 800 media interviews. Such circumstances can be capitalised upon from the perspective of promoting awareness and understanding of service programmes.

It is increasingly common for NMSs to operate feedback and response mechanisms. Some of these systems work via the Internet in conjunction with Web offerings, and others are telephone based, and yet still others operate via regular mail. Specific levels of service regarding initial and final response are generally established. These tend to be very useful input sources of information on the effectiveness and adequacy of service offerings and on the operation of the production and delivery systems. The coding of the information in a database will make it available for future analysis to determine patterns or trends.

5.3.1.4   Consultation

Public and stakeholder consultations are a standard means of obtaining input on NMS policies and issues. These consultations can take various forms. The visiting of user associations such as attendance at their meetings lends a human face to what may otherwise be seen as a faceless bureaucracy producing weather and climate services. Being on "their territory" facilitates the exchange of honest reactions to the services provided by the NMS. User meetings in some neutral ground are also good venues to achieve similar results. User or joint conventions and visiting client sites can be used similarly.

Hosting workshops or other events for the broad user community or for particular clients or client groups is also effective.

5.3.1.5   *Post-Event Review, Case Studies and Debrief*

On the one hand post-event reviews or case studies can be evocative with problems coming to the forefront and becoming more persuasive leading to a motivation to make positive changes, while on the other hand the case may dominate all other information and can be too striking thereby biasing the interpretations. Careful selection of the case is essential. Most NMSs will undertake operational performance reviews following major meteorological events to assess the effectiveness of their systems. One such review was undertaken following the costly "Ice Storm" of January 1998 in Eastern Canada.

Reviews can result in complete end-to-end operational system audits of what worked effectively and what did not. It is common to analyse the accuracy and appropriateness of meteorological products. The effectiveness of the information delivery system is a critical component to be analysed, as is the effectiveness of the NMS's relationship with other agencies involved is disaster management. Surveys of the citizenry and even the local media provide useful information. An assessment of the public "issue management" can lead to improved strategies for future similar situations. Documenting and learning from these situations are key steps towards improvements.

5.3.1.6   *Collection of Anecdotal Information*

Finally, for more than historical purposes, NMSs collect anecdotal information to be used strategically. This involves the collection of stories of lives saved and damage avoided, through effective warnings and forecasts. These "sound bites" can be used strategically for public relations purposes or to defend certain perspectives with clients and partners.

5.3.2   Formal Structured Surveys

5.3.2.1   *Large Aurvey every 4 or 5 Years – Comprehensive*

In most cases survey objectives call for the measurement of many characteristics. In a survey on meteorological services one usually wants to determine more than overall satisfaction or perceptions about weather forecasts. A comprehensive survey may include sets of questions on the general use of weather information, on weather warning information, on regular forecast information, on air quality information, on weather information delivery, demographics, etc. Within these sections of a multi-purpose survey further breakdowns can occur such as under the general topic of weather forecast information one can investigate, on a per season basis perceptions of what is considered accurate for temperature, wind direction/speed, onset of precipitation, probability of precipitation, sky cover conditions (sunny, cloudy), etc. These surveys are usually quite long and demand fairly large sample sizes to facilitate geo-politically based inferences.

To accommodate the measurement of several items within one survey plan, it is likely necessary to make compromises in many areas of the survey design. The method of data collection (telephone, personal interview, mail-out, etc) may be suitable for measurement of some characteristics but not for others. The survey design must be made to properly balance statistical efficiency, time, cost, and other operational constraints. As such, they tend to be rather costly so such base line surveys are usually undertaken once every four or five years. In order to make proper inferences on trends consistency in the design and questions from one baseline survey to the next is necessary. Given the cost, such surveys demand particular senior management discipline and commitment for appropriate long tern execution. An example of such a survey, the 1997 Canadian Goldfarb Survey, forms Appendix 2 of this Technical Document.

5.3.2.2   *More Frequent Tracking Surveys*

One-time or baseline surveys differ from periodic or continuing surveys in many ways. The aim of periodic or continuing surveys is often to study trends or changes in the characteristics of interest over a period of time. Such studies nearly always measure changes in the characteristics of a population with greater precision. Overhead costs of survey development and sample selection can be spread over many surveys and this in turn cuts down the costs. Decisions made in the sample design of periodic or continuing surveys should take into account the possibility of deterioration in design efficiency over time. Designers may elect, for example, to use stratification variables that are more stable, avoiding those that may be more efficient in the short term but which change rapidly over time. Another feature of a periodic or continuing survey is that, in general, a great deal of information is available which is useful for design purposes. If, for example, a Service Charter calls for routine reporting on levels of satisfaction (or another dimension) with regard to certain standard forecast elements, a well designed standard survey instrument can be used repetitively. Recognising the compromises, an omnibus survey vehicle can be used.

An example of a tracking survey is the Hong Kong, China, survey that forms Appendix 3 of the present Technical Document.

5.3.2.3   *Subject Area Surveys*

Subject area surveys offer the potential to delve more deeply into specific characteristics of interest. This can be for the purpose of investigating perceptions regarding key issues of concern to an NMS such as climate change, or even for specific valuation exercises such as estimating the benefits of a specific service provided via a specific delivery mechanism. These surveys are specifically designed to answer a limited set of questions and, as such, all of the design dimensions should be carefully considered. These include the thorough specification of the objectives, the development of operational definitions which indicate who or what is to be observed and what is to be measured, the specification of the

data requirements, an indication of the purpose, the areas covered, the kinds of results expected, the users as well as the uses of the data, and the level of accuracy that is desired.

5.3.2.3.1 Key Issues

As stated above, climate change is an example of an issue area that can be the focus of a subject area survey. Others can be perceptions about air pollution, natural disasters, etc. These issue area investigations are more prevalent in the broader environmental field than in the more narrowly defined scope of meteorological services.

5.3.2.3.2 Product Lines

Public opinion research into product lines is far more common in the meteorology field. User perceptions regarding offered products are popular topics of such surveys. Typically dimensions investigated include the establishment of user requirements, determination of levels of satisfaction or utility, a measurement of the awareness of existence, or the origin or means of accessing of certain products. Also included are an assessment of the level of understanding of the terminology or meaning associated with certain parameters, a determination of the perception of what is accurate relative to individual forecast elements, an assessment of the perceived accuracy or credibility of certain forecast parameters, an assessment of the required frequency for updates to certain forecasts and reports, and an assessment of the timeliness of any one of a variety of warnings.

Some concepts, such as probability of precipitation, windchill and heat indices, represent particular challenges with regard to effective communication leading to appropriate behavioural response. Public opinion research into comprehension and options for effective communication of such more complex parameters is often essential for their design. Given scarce resources, and even more importantly the limited "sound bite" space allowed by dissemination technologies and specifically the media, it is often critical to determine the relative importance of products or weather elements. Decisions on service levels and design are commonly made on the basis of user-based assessments achieved through these means.

5.3.2.3.3 Delivery Systems

As is often stated, critical meteorological information not delivered at all or delivered in an incomprehensible manner or via a medium not receivable by users or specific critical stakeholder communities of users has little or no value. Surveys covering the variety of dissemination technologies such as the popular media (radio, TV, newspapers, etc.), the Internet, weatheradio, telephone, pagers, mobile technologies, and digital radio are frequent targets for subject area surveys.

Specifics analysed can include layout, graphics, colours, duration of a broadcast and length and wording of text. Reach and target audiences of the specific media are other specifics analysed. An example of a delivery system specific survey would be one focused on the acceptance and utility of "crawling" weather warning messages on TV screens thereby interrupting the viewing of programmes and/or commercials.

Information derived from such investigations can be used in presentations made before industry and government authorities in application for licenses etc. Generally, information derived from public opinion research in the service delivery area can lead to decisions on which systems to be utilised for the population as a whole and for specific target audiences, and on specific attributes in terms of product design and delivery.

5.3.2.3.4 Economic Value Estimation

Production-based methods vs. demand-based methods
The value of weather information is a subdivision of the economic literature on the value of information. Two main models or methods have been used for the valuation of meteorological information. Broadly speaking, one can generalise the majority of applications of non-market valuation of weather information services to these two types of methodologies. For either method chosen it is important that the NMS avail itself of professional expertise in the respective economic theory for such estimations.

Production-Based (PB) "Analytical Methods" rely on modelling processes in which the information is used as an input to the production of a consumer product, which is ultimately valued in the marketplace. Thus these prescriptive analysis methods indirectly infer the benefits of the information input as the contribution to the market value of the final product. Typically, the production process is modelled and the added value attributable to the use of meteorological information is estimated at each stage in the production process and aggregated for the entire production process.

The Demand-Based (DB) "Survey or Interview Method" directly infers the benefits of the weather information services via characterisation of the demand for the service, as articulated by users' willingness to pay. Direct descriptive methods rely on modelling the relationship between willingness to pay for a service and the benefits generated by that service in aggregate over the range of users. For this section on user-based assessment the focus will be on methods such as contingent valuation, a widely accepted DB method used by economists to value public goods and services.

Production-Based "Analytical Methods"
Production-based (PB) "Analytical Methods" have been by far the most common approach used in the meteorological literature with a variety of published studies that value weather information in contexts ranging from costs savings to road maintenance, forest fire prevention and fuel load decisions for the aviation industry, irrigation scheduling, and the value of increased accuracy of forecast information to increase production of a variety of agricultural commodities. The assessment is typically not (end) user-based.

Demand-Based "Survey or Interview Method"
The Demand-Based (DB) "Survey or Interview Method" relies on providing the means for users of specific weather information dissemination services to reveal how much they would be willing to pay for the service if they had to do so.

DB "Survey or Interview Methods" assume that the user implicitly knows what the value of the service is to him in the context of his own ability to use it to produce benefits to himself. For business users who use the information as a productive input, the benefits implicitly include the user's understanding of the production process. For household users who use the information for planning recreational activities, the benefits implicitly incorporate a subjective valuation of the increase to household utility from the information.

Different users for the same service likely derive different levels of benefits from it, and these differences would be expected to be reflected in a random sample of all users. The contingent valuation (CV) method (one of a number of survey-based economic valuation techniques that can be employed with the assistance of professional expertise in this economic theory – not to be explained here) directly measures individual willingness to pay (WTP), and can easily differentiate between significant differences in WTP among user groups, provided the sample of each is large enough. The individual WTP for each user group can then be aggregated over the populations of users in each group. The sum of these aggregates is thus the total value of the proposed change in the provision of the service throughout the market.

A demand-based approach is not intended to result in an in-depth analysis of how changes in provision of the service can affect production in a given production process. Typically, production issues are treated qualitatively with additional survey questions that ask each user how they use the information in their own decision-making. On the other hand, demand-based approaches, properly applied, do analyse what, if any, substitutes exist for the service, and then value the service as the marginal value of the service over and above the value of these substitutes.

DB approaches are very specific to the type of weather-information dissemination service considered, and results are not theoretically applicable to other types of services. So while, PB methods value the information itself as a productive input, DB methods are more specific to the means by which the information is delivered because this is the specific good that users employ – a particular bundle of weather information supplied in a particular manner, accessible at particular times, etc. The DB approach assumes that the user of the service knows how they would respond to a price change or quality change in the service by substituting with other sources of the needed information.

In the specific policy context of analysing the impact of alternative weather information delivery systems in a cross-sectoral comparison, DB methods are likely superior to PB methods. In a context requiring an in-depth analysis that models the complexity of means by which a change in the quality of information delivered by any system would affect a particular user group, PB methods are likely superior.

5.3.2.3.5 Current Value Versus Value if Accuracy Increased

Both the PB and DB methods can be used to achieve valuations of both current value of specific services and the degree of increased benefit attributable to improved quality of the services. A Guelph University study used a prescriptive PB approach to value current precipitation forecasts to the Southern Ontario, Canada dry hay industry at CAN\$54 Million while a 50% improvement in those forecasts increased that value to \$58 Million. A descriptive Contingent Valuation (DB) approach was used by Dalhousie University to value the Marine Weather Services in the Canadian Maritime Provinces at more than twice the cost of the provision of the service. One Meteorological Service of Canada Contingent Valuation study demonstrated the ability to select an optimal asking price for services delivered over the telephone for maximisation of cost recovery while another study demonstrated that the benefit of Marine Weather Services delivered via Weatheradio Canada exceeded the anticipated increased cost of provision of that service resulting from to large increases in broadcast tower costs.

5.3.2.4 Questionnaire Design

5.3.2.4.1 Some General Rules for Questionnaire Design and Wording

- It is essential to ensure that the questions and instructions are easy to understand.
- Abbreviations and jargon should be avoided.
- Words and terminology that are too complex should be avoided.
- The frame of reference should be specified. For example if income information is requested then, at a minimum, a time frame should be specified.
- Questions must be as specific as possible.
- The question needs to be understood by all respondents in the same way. To the extent possible, the questions asked should be applicable to all respondents. Clearly, skip patterns (those "go to" type directional statements that determine the next question to be asked based on the response to the question just asked) are defined such that respondents are not required to answer all of the questions.
- The questions should be relevant to the respondent and the respondent should know enough about the subject to answer the question knowledgeably.
- Double-barrelled questions should be avoided. Double-barrelled questions are ones that have two or more questions "nested" within them. Respondents become confused in trying to answer the question, especially when they have different answers for each part. One indicator of the likelihood of a double-barrelled question is the appearance of the conjunction "and" or "or" in the question. The best way to avoid the confusion is to replace double questions with two or more questions.
- Don't try to get two questions answered by way of one question.
- The response categories should be mutually exclusive and exhaustive.
- Care should be taken in developing the wording of the questions so as to avoid the likelihood of drawing invalid inferences from the responses. That is, the questions should not be "leading" or "loaded" i.e. should not suggest that one answer is preferable to another.

5.3.2.4.2 Types of Questions

Open versus closed questions

There are two main types of questions: *open* and *closed questions.* They are sometimes called open-ended and closed-ended questions.

Open questions are answered in the respondent's own words. An open question allows the respondent to interpret the question and answer anyway that he/she wants. The respondent writes the answer or the interviewer records verbatim what the respondent says in answer to the question. Blank spaces are left in the questionnaire after the question for the response to be written in.

Closed questions are answered by means such as by checking a box or circling the proper response from among those that are provided on the questionnaire. A closed question restricts the respondent or interviewer to select from the answers or response options that are specified.

Sometimes a continuum from open to closed questions is employed. This can take the form of a closed question where amongst predetermined optional responses is an option to check off a category such as "other (please specify) _____" followed by a blank space where the respondent writes the answer or the interviewer records verbatim what the respondent says in answer to the question.

Open Questions

Open formats are typically used for qualitative research where "natural" wording is desired; for the provision of the opportunity for self-expression or elaboration; for the attainment of exact numerical data; or simply to add variety to the questionnaire. For the respondent open questions can be more time-consuming and more demanding from the perspective of having to formulate a response. From the researcher's perspective open questions can be costly and difficult to analyse.

Closed Questions

There are many different types of closed questions including two-choice, multiple choice, checklist, ranking format, rating scale, etc. Closed questions provide respondents with definite choices. The respondent indicates which choice is appropriate.

With the two-choice and multiple-choice questions only one choice is allowed.

In a checklist question many choices may be selected but the choices should be non-overlapping.

In a ranking question the respondents are typically asked to rank the choices from highest to lowest according to some criteria. Such questions are often difficult for the respondent to deal with, especially in event of equally ranked items, and the results are difficult to analyse. The order in which the items are listed can influence the results. Difficulties associated with rating scales include the determination of the appropriate number of categories and the tendency for responses to gravitate to the middle area and avoidance of the extremes.

The Thurstone Scale is a composition of two-choice questions where the respondent is presented with a list of statements, each of which he/she is asked to endorse or reject. Each statement should be clear, brief, and easy to understand.

The Likert scale is a composition of multiple-choice questions where the respondent considers each statement and reports how closely it reflects his/her own opinion by indicating not only whether he/she agrees or disagrees, but also how much he/she agrees or disagrees. "Agreement" is not the only response option that can be used. Other response dimensions include "satisfaction", "usefulness", "importance", etc. Degrees of frequency are another possibility.

For a respondent the advantage of closed questions is that they are easier and faster to answer. For the researcher they are easier to code, easier to analyse, generally cheaper to administer and provide consistent response categories. Closed questions are an advantage when you can anticipate all (or most) of the responses and when an exact value is not required.

There are also significant limitations to closed questions. Often, more effort is required to develop closed questions than open questions. A closed question may elicit an answer where no knowledge or opinion exists (including a "Don't know" or "No opinion" response option may help). Closed questions may oversimplify the issue or force answers into an unnatural mold. Closed questions may not be in the same format as the respondent's record-keeping practices. The response categories must be inclusive and non-overlapping.

5.3.2.4.3 Sequencing of Questions

Issues in sequencing include the introduction, the opening questions, the location of sensitive items, the location of demographic items and the flow of items. The order of the questions should be designed to encourage respondents to complete the questionnaire and to maintain their interest in it. The order should facilitate respondent's recall and appear sensible to the respondents. The order should focus on the topic of the survey. It should follow a sequence that is logical to the respondents and should flow smoothly from one question to the next but should not influence the actual response itself.

The introduction should provide the title or subject of the survey and identify the sponsor. It should explain the purpose of the survey and request the respondent's co-operation. Respondents frequently question the value of the information to themselves and to users. Some like to receive feedback about the survey. Therefore it is important to explain why it is important to complete the questionnaire and to ensure that the value of providing information is made clear to respondents. It is helpful to explain how the survey data will be used and how, if possible and/or desirable respondents can access the data. Also, it is important to indicate the degree of confidentiality and any data sharing arrangements.

The opening questions should establish respondents' confidence in their ability to answer the remaining questions. If necessary, the opening questions should establish that the respondent is a member of the survey population. The opening questions should relate to the introduction and the survey objectives. The opening should be applicable to all respondents and be easy and interesting to answer.

The location of sensitive questions is a particular challenge. Sensitive questions (i.e. ones perceived as irritating

or threatening), for example, questions on income and age, tend to get a low response rate and may trigger a refusal by the respondent to co-operate any further. They should not be placed at the beginning of the questionnaire. Introduce them at the point where the respondent is likely to have developed trust and confidence. Locate sensitive questions in a section where they are most meaningful in the context of other questions. It is useful to introduce these gradually by warm-up material that is less threatening. Options or tools that can be employed are self-enumeration (the respondent fills out the questionnaire in private), anonymous questionnaire, careful wording of questions, the use of ranges for response categories and randomised response. In the simplest form of the randomised response technique, the respondent answers one or two randomly selected questions without revealing to the interviewer which question is being answered. One of the questions is on a sensitive topic; the other question is innocuous. Since the interviewer records a "yes" or "no" answer without ever knowing which question has been answered, the respondent should feel free to answer honestly. This can be done, for example, in an in-person interview where the interviewee selects a card (code noted by the interviewer without seeing the side that contains the questions) or is handed one by the interviewer, who notes the respondent's responses to the questions on the card in sequence. Demographic and classification data can be either placed at the end of the questionnaire or inserted into the most relevant sections.

The flow of the items should follow the logic of the respondent. Time reference periods should be clear to the respondent. Similar questions should be grouped together. It is useful to provide titles or headings for each section of the questionnaire. Also, use wording that facilitates movement from one section to the next.

5.3.2.4.4 Layout Considerations for Questionnaires

As a general guideline the questionnaire should appear interesting and easy to complete and respondent-friendly. If done through the mail (regular or electronic) the cover letter and front cover should create a positive initial impression by way of a respondent-friendly introduction. If the questionnaire is administered in person or over the telephone, the questionnaire should be interviewer-friendly. The instructions should be short and clear and the structure should be such that the respondent is guided step-by-step through the questionnaire. The instructions and answer spaces should facilitate proper answering of the questions. Illustrations and symbols (such as arrows and circles) should be used to attract attention and guide respondents or interviewers. It is a good idea for the last page or end of the questionnaire to provide space for additional comments by respondents. Finally, always include an expression of appreciation ("Thank You)".

Typography considerations in organising the printed word on a page include typeface/font (ensure consistency, use **bold face print** or ALL CAPITAL LETTERS to highlight important instructions or words), form titles, section headings, questions and question numbers. Data entry or processing codes should not take precedence over, nor conflict with, the question numbers. The benefits of a respondent friendly questionnaire include improved respondent relations and co-operation, improved data quality, reduced response time and reduced costs.

5.3.2.4.5 Response Errors

A response error is the difference between the true answer to a question and the respondent's answer to it. It can occur anywhere during the question-answer-recording process. There are two types. *Random errors* are variable and tend to cancel out. *Biases* tend to create errors in the same direction.

One of the sources of response error is the questionnaire design. It can come from the wording, the complexity and from the order of the questions. It can also come from the question structure, complicated skip patterns and from the very length of the questionnaire.

Another source of response error is the respondent problems of understanding, recall, judgement, motivation and reporting. Recalling an event or behaviour can be difficult if the decision was made almost mindlessly in the first place, or if the event was so trivial that people have hardly given it a second thought since it occurred. Recalling is also difficult if the question refers to something that happened long ago or if the questions require the recall of many separate events. The resultant errors include the respondent failing to report certain events or failing to report them accurately leading to an underreporting of events. A less frequent memory error is the telescoping error. Here some events may be reported that actually occurred outside the reference period leading to the over-reporting of events. Generally speaking the longer the reference period, the greater is the recall loss while a shorter reference period tends to increase telescoping errors.

Social desirability bias can also emerge. This is the tendency to choose those response options that are most favourable to one's self esteem or most in accord with perceived social norms, at the expense of expressing one's own position.

Finally, the interviewer can be the source of the error.

5.3.2.4.6 Probing for More Information

Probing for more information is a common practice in interviewing whether in the context of a consultation session, a workshop or a focus group session. Indeed, it is the main means of eliciting information and it is the skills of the facilitator that come to advantage here. While it can also be used in in-person one-on-one interviews it is less common in telephone interviews and not possible in mail, Internet or kiosk based interviews. The survey instrument can often be written in such a manner so as to effectively achieve a similar purpose.

5.3.2.4.7 Geographical and Geopolitical Representation

Most national government statistical bodies have developed "standard industrial classifications" that classify industries on the basis of their principal activities and

"standard geographical classifications" for the identification and coding of geographical areas. These "standard geographical classifications" usually correspond to geopolitical boundaries. The objective of the system is to make available a standard set or framework, which can be used to facilitate the comparison of statistics for particular areas. Sample allocation decisions are often made on the basis of these standard classifications.

5.3.2.4.8 Data Coding and Capture

To avoid being faced with a long, expensive error-prone task of manually coding and possibly transcribing data, consideration should be given, at the design stage, to the capture of the data for subsequent processing. It is important to consult early, regularly, and often with the processing staff, to design any formal survey questionnaire for rapid data capture. The best way of ensuring that the concerns of data capture are addressed is to make the individual/organization responsible for this aspect of the survey a permanent member of the team planning and implementing the questionnaire .

If data is to be processed by a computer, which is usually the case, codes for the fields into which answers are to be keyed should appear directly on the questionnaire. These are there to better ensure error-free data entry by interviewers. It is now common to have this process entirely computer resident with the interviewer entering the data into a computer database via a questionnaire data entry screen. The database can be personal computer based utilizing commonly available and relatively inexpensive software. The data can also be analysed using relatively inexpensive spreadsheet software or slightly more costly statistical software packages such as SPSS or SAS.

Chapter 6
CONCLUSIONS

6.1 INTRODUCTION

This Chapter is written especially for those readers who like to read the Introduction to a document, skim through the technical detail in the middle, and jump to the end to find out what the main conclusions were and what, if anything, they should do about it. Here are the answers you seek….

6.2 SUMMARY

Performance assessment should be an essential element of the public weather services programmes of all NMSs. Imagine how it would be if an NMS tried to do forecasting without first gathering observations. Performance assessment is a bit like gathering that basic data – on user requirements, on users' perceptions of services, and on how good the outputs are. Analysis of the data can be used to improve performance.

The purpose of performance assessment is to ensure above all that, as far as possible, the user requirements are being met. It is also used as a check on the operational effectiveness and efficiency of the overall PWS system. Importantly, the information gathered is also very useful for communications with the public and government, which help raise the profile of the NMS and enhance its credibility.

The risk is that a performance assessment programme may be carried out without ever taking any actions based on the results. It is important from the outset to ensure that information is being gathered not to just sit on the shelf, but to be analysed and used for actions which will improve the NMS's performance in the provision of public weather services.

These actions may include improving the products and their delivery, modifying the forecast production system, carrying out needed research and development, and recruiting and training staff, as well as communicating relevant information. Because budgets and resources are always limited, there will of course have to be some prioritisation on what actions will bring the best benefits.

The two essential and complementary aspects of an assessment programme are Verification, and User-Based Assessment. The overall purpose of Verification of forecasts is to ensure that products such as warnings and forecasts are accurate, skilful and reliable from a technical point of view. User-Based Assessment relies on seeking information from people, to obtain a true but subjective reflection of the user perception of products and services provided by the NMS, as well as qualitative information on desired products and services.

6.3 HOW TO GET STARTED ON A PERFORMANCE ASSESSMENT PROGRAMME

For those NMSs which don't currently have a performance assessment programme, now is the time to get started on that first step (always the hardest!).

6.3.1 Planning

Since performance assessment involves a range of functions within the NMS, the first step should be to set up a team to develop a programme plan. This team should be large enough to involve the main functions – in particular, forecasting, computing systems, marketing (or whatever this function is called) – but also small enough so that it does not become unwieldy. Commitment from senior management is essential, and preferably at least one senior manager should be on the team.

The first task of the team should be to reach agreement on the purposes and objectives of the performance assessment programme. What is the most important information you want to discover? Do you need particular information for reporting purposes? Have there been many complaints about a particular forecast? Have you asked the users recently whether the products are meeting there needs? A review of this Technical Document should provide lots of clues and cues for the kind of information you might want to gather.

Planning should then proceed on how best to gather that information, how it is going to be analysed and used and communicated, and who is going to be responsible for ensuring that actions are actually taken based on the results. Since this will all involve work, it is important to "keep it simple" and not embark on an overly ambitious programme to start with. Communicate widely within the NMS as this planning takes place, and seek feedback from people who are interested. Forecasters, amongst others, will undoubtedly have something useful to contribute.

6.3.2 User-based Assessment

In the area of User-Based Assessment, the questionnaire from the Hong Kong Observatory in Appendix 3 is a good example of a simple, focussed questionnaire. This gathers some basic information on the public's use of weather forecasts, how they access them, and what their perceptions are of their accuracy.

You might wish to use this as the basis of a similar questionnaire for your NMS. But, before doing so, think very carefully about how the information gathered will be used by you. Some of the information in this sample questionnaire is clearly designed for "tracking performance" – this is useful for reporting purposes and also for suggesting remedial action if the performance is perceived to be very poor in some areas. Other information about the delivery channels can be used for re-prioritising the effort put into different products for the different channels. You should also consider how the questions should be modified to fit your own circumstances, and needs to information to communicate and make decisions on.

6.3.3   Verification

Temperatures

A simple first step into verification is to verify maximum temperature forecasts. These are provided by most NMSs, and just about everyone cares about temperatures. The example in Section 4.3.1 shows many measures of reliability, accuracy and skill which can be used to verify these. Perhaps the first questionnaire you use can also ask the public what they consider to be an "accurate" maximum temperature forecast. Is within 2°C accurate? Within 3°C?

As statistics accumulate, you can see how skilful the forecasts are compared to benchmarks, which could include statistical forecasts based on numerical model output. Do the manual forecasts have a worthwhile improvement over model forecasts? Are they both poor? Is it worth considering a research and development programme to improve the guidance? Do the forecasters need more information available on temperature climatology, and on case studies of unusually hot or cold temperatures?

Precipitation

A typical second step into verification would be to verify forecasts of precipitation. In most parts of the world this is of significant interest to the public - but maybe you should check this as part of your first questionnaire?

Verification of "yes" or "no" for precipitation is covered in some detail in Section 4.3.2, and the example in Appendix 1 shows how a simple spreadsheet can be used to compute various scores. You can ask yourself the same kinds of questions as for maximum temperatures above. If in some climates a simple "yes" or "no" may not suffice – the three category example of "dry", "showers" and "wet" described in Section 4.3.4 could be used instead.

Severe Weather Warnings

Given the importance of forecasts of severe weather, these could form the third part of an initial Verification programme. It is critical for these forecasts to have a well-defined criteria, or else verification will be difficult. For example, the criterion used in New Zealand for issuing (and verifying) a warning of heavy rainfall is for more than 100 mm in 24 hours, over a widespread area (more than 1000 km^2). Such forecasts can be verified using the scores in Section 4.3.2.

6.3.4   Ongoing Assessment

A Performance Assessment Programme is not something that you just set up, and let run. It will need ongoing development, and adjustment, and fine tuning. In fact, you should be assessing the Assessment Programme itself. Many of the methods described in Chapter 5 can be used with your *internal* customers in the NMS to make sure that the programme is meeting *their* needs, and to improve it.

6.4   FINAL WORDS

Performance Assessment is the key to ensuring an effective, efficient and sustainable Public Weather Services programme. We trust that the guidelines provided in this Technical Document will be of value to you in establishing or developing your own Programme, and wish you well in that endeavour.

REFERENCES

Brier, G.W., 1950: Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78, 1-3.

Epstein, E.S., 1969: A scoring system for probability forecast of ranked categories. *Journal of Applied Meteorology*, 8, 985-987.

Gordon, N.D., 1982: Evaluating the skill of categorical forecasts. *Monthly Weather Review*, 110, 657-661.

Hanssen, A.W., and W.J.A. Kuipers, 1965: On the relationship between the frequency of rain and various meteorological parameters. *KNMI Meded. Verhand.*, 81, 2-15.

Murphy, A.H., 1997: Forecast verification. In *Economic Value of Weather and Forecasts*, ed. R.W. Katz and A.H. Murphy, 19-74. Cambridge: Cambridge University Press.

Patton, M., 1990: *Qualitative Evaluation and Research Methods*, (2nd Edition). Newbury Park, California: Sage Publications.

Platek, R., F.K. Pierre-Pierre and P. Stevens, 1985: *Development and Design of Survey Questionnaires*. Statistics Canada

Purves, Glenn T., 1997: *Economic Aspects of AES Marine Weather Services in Marine Applications, A Case Study of Atlantic Canada*. Dalhousie University.

Rollins, Kimberly, J. Shaykewich, 1997: *Cross-Sector Economic Valuation of Weather Information Dissemination Services: Two Applications Using the Contingent Valuation Method*. University of Guelph.

Satin, A., W. Shastry, 1983: *Survey Sampling: A Non-Mathematical Guide*. Statistics Canada.

Stanski, H.R., L.J. Wilson and W.R. Burrows, 1989: *Survey of common verification methods in meteorology*. WWW Technical Report No. 8 (WMO/TD 358), 114 pp.

Turner, Jason R., 1996: *Value of Weather Forecast Information for Dry Hay and Winter Wheat Production in Ontario*. University of Guelph.

Wilks, D.S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.

Appendix 1
EXAMPLE OF MONTHLY RAINFALL VERIFICATION

The following table shows an example (using a simple spreadsheet) of rain / "no rain" verifications.

RAINFALL VERIFICATION

LOCATION: Auckland
MONTH: July   **YEAR:** 1999

Enter either R (for Rain) or N (for No rain)

Day	Forecast	Observed
1	N	R
2	N	N
3	R	R
4	R	R
5	R	R
6	N	N
7	R	R
8	R	R
9	R	N
10	N	R
11	N	N
12	N	N
13	R	R
14	N	N
15	R	R
16	R	R
17	R	R
18	R	R
19	R	R
20	R	N
21	R	N
22	R	R
23	R	R
24	N	N
25	R	R
26	R	R
27	R	R
28	R	R
29	R	N
30	R	R
31	R	R

The following area on the spreadsheet shows various skill scores which can be computed from the 2 by 2 contingency table resulting from these data. The scores are defined in Section 4.3.2, and the 2 by 2 contingency table is the same as used for an example in that section.

SUMMARY

		Observed	
		Yes	No
Forecast	Yes	19	4
	No	2	6

% correct all forecasts	81%	
% correct for rain forecast:	83%	
% correct for no rain forecast:	75%	
Bias:	110%	
Rain POD:	90%	
Rain FAR:	17%	
Rain Threat Score or CSI:	0.76	
Rain hits expected by chance:	15.6	
Heidke Skill Score:	0.46	
Equitable Threat Score:	0.36	
Hanssen-Kuipers Skill Score:	0.50	
No-Rain hits expected by chance:	2.6	
% correct expected by chance:	59%	
Skill of % correct over chance:	53%	

FORMULAE:

		Observed	
		Yes	No
Forecast	Yes	A	B
	No	C	D

$(A+D)/(A+B+C+D)$
$PC=A/(A+B)$
$D/(C+D)$
$(A+B)/(A+C)$
$A/(A+C)$
$B/(A+B)$
$A/(A+B+C)$
$CHA=(A+B)*((A+C)/(A+B+C+D))$
$(A-CHA)/(A+B-CHA)$
$(A-CHA)/(A+B+C-CHA)$
$A/(A+C)+D/(B+D)-1$
$CHD=(B+D)*((C+D)/(A+B+C+D))$
$CHPC=(CHA+CHD)/(A+B+C+D)$
$(PC-CHPC)/(1-CHPC)$

Appendix 2
ENVIRONMENT CANADA'S ATMOSPHERIC PRODUCTS AND SERVICES 1997 NATIONAL PILOT SURVEY

Administered by:
Goldfarb Consultants

for:
The Program Evaluation Group of the Policy, Program and International Affairs Directorate

Good morning/afternoon/evening. My name is _____ of Goldfarb Consultants, a national survey and opinion research firm. We are conducting a survey on behalf of Environment Canada today. The results of this study will be used to help design and modify existing programs and services to better meet your needs. We are not selling anything. We are simply interested in your attitudes and opinions. Can you spare some time to answer some questions for me? THANK YOU.

A. May I please speak with the male/female **[ROTATE]** in the household age 18 or over whose birthday comes next? **[IF THE RESPONDENT IS NOT AVAILABLE, GET PERSON'S NAME, MARK AS "ARNA", AND ARRANGE FOR A CALL BACK.]**

   [REINTRODUCE IF NECESSARY]

B. Respondent is...

   Male ☐
   Female ☐

   [WATCH QUOTAS – TERMINATE IF NECESSARY]

C. I would just like to confirm that you are over the age of 18.

   Yes, respondent is over 18 ☐
   Respondent is under 18 ☐ **TERMINATE**

D. We are interested in people's occupations. Do you or does anyone in your household work for...

   A radio or television station ☐
   A newspaper or magazine ☐
   A public relations firm ☐
   An advertising agency ☐
   A market research firm ☐

   IF "YES" TO ANY OF THE ABOVE, TERMINATE.

SECTION ONE: USE OF WEATHER INFORMATION

1. We would like to talk to you about the types of news that you hear or look at. During a typical day, how likely are you to look at or hear news on each of the following topics? Are you very likely, somewhat likely, not very likely or not likely at all to get news on... **[ROTATE]**

	Very Likely	Somewhat Likely	Not Very Likely	Not Likely At All
Local events and politics	☐	☐	☐	☐
Entertainment	☐	☐	☐	☐
Weather	☐	☐	☐	☐
Traffic	☐	☐	☐	☐
Sports	☐	☐	☐	☐

2a) We'd like to focus more on weather information for the remainder of this interview. First of all, on a typical day, how many times would you say that you specifically make a point of actually looking at or listening to weather forecasts? Would it be... **[READ LIST]**

More than four times a day	☐
Three times a day	☐
Two times a day	☐
Once a day	☐
Less often than once a day	☐

2b) If you are in need of a weather forecast, how often is it available to you? Is it available…   **[READ LIST]**

Always	☐
Most of the time	☐
About half of the time	☐
Less than half of the time	☐
Rarely or never	☐

2c) Compared to two years ago, would you say that you are using weather forecasts more often today, the same, or less often than you were two years ago?

More often	☐
The same	☐
Less often	☐

2d) Compared to two years ago, how satisfied are you with your access to weather information or forecasts?
[READ. CHECK ONE]

Much more satisfied now	☐
A little more satisfied now	☐
Just about as satisfied now as then	☐
A little less satisfied now	☐
Much less satisfied now	☐

3a) We are interested in where you get your weather information from. From what main source are you most likely to get your daily weather information? **[DO NOT READ. CHECK ONE ONLY. CLARIFY "TELEVISION" AND "TELEPHONE" RESPONSES.]**

3b) What other sources do you get weather information from? **[DO NOT READ. CHECK AS MANY AS APPLY.]**

	3a) Primary source	3b) Secondary source
Television – General mention	☐	☐
Television – Weather network	☐	☐
Television – Local Environment Canada cable channel	☐	☐
Radio	☐	☐
Newspaper	☐	☐
Internet Access	☐	☐
WeatherRadio Canada	☐	☐
WeatherCopy Canada	☐	☐
Contact Environment Canada weather office	☐	☐
Telephone – General mention	☐	☐
Telephone – 1-800 number	☐	☐
Telephone – 1-900 number	☐	☐
Environment Canada recorded tape	☐	☐
Family member	☐	☐

3a) Other Primary : _____

3b) Other Secondary: _____

4. On a typical day, when do you make a point of trying to look at or hear weather forecasts? **[PROBE]** Are there any other times? **[DO NOT READ. CHECK ALL THAT APPLY.]**

Morning – General mention	☐
Morning – Wake-up	☐
Morning – While dressing/dressing kids	☐
Morning – With news	☐
Morning – Drive to work	☐
Afternoon- General mention	☐
Evening – General mention	☐
Evening – Drive home	☐
Evening – With news	☐
Evening – Before bed	☐
Evening – Before work	☐
Other	☐

5a) We would like to know if the information provided in weather forecasts is sufficient enough for you to make decisions on plans or actions that you would take, on a typical day. That is, do you feel that weather forecasts always provide you with enough information to make decisions, sometimes provide you with enough information, rarely provide you with enough information or never provide you with enough information to make decisions?

Always	☐	
Sometimes	☐	**ASK QUESTION 5B**
Rarely	☐	**ASK QUESTION 5B**
Never	☐	**ASK QUESTION 5B**

5b) What other information would you require to make decisions? **[DO NOT READ. PROBE. CHECK ALL THAT APPLY.]**

Temperature – GM	☐	Wind speed	☐
High/Maximum	☐	Direction of wind	☐
Low/Minimum	☐	Whether it will be gusty	☐
		Significance of wind-chill	☐
Humidity level	☐		
Humidex	☐	Visibility information	☐
		Amount of sun	☐
Precipitation/Rain/Snow	☐	UV Index	☐
Amount of rain/snow	☐	Air quality	☐
Type of precipitation (rain/snow/hail)	☐		
When precipitation will start	☐	Expected weather changes	☐
When precipitation will end	☐	Storm expectations	☐
Whether precipitation will be heavy/light	☐		
Probability of precipitation	☐	Historical information	☐

Other:_____

_____

6. We'd now like you to think specifically about Environment Canada for a moment. Can you think and tell me the types of weather-related services Environment Canada provides and performs? **[PROBE AND CLARIFY]**

_____

7. Now, how often does your work or job require you to make decisions based on the weather? Is it... **[READ LIST]**

Always	☐
Sometimes	☐
Rarely	☐
Never	☐ **GO TO QUESTION 10**
Don't work	☐ **GO TO QUESTION 10**

8. What parts of the weather forecast do you need for you to make work-related decisions? **[DO NOT READ. PROBE. CLARIFY. CHECK ALL THAT APPLY.]**

Temperature-GM	☐	Wind speed	☐
High/Maximum	☐	Direction of wind	☐
Low/Minimum	☐	Whether it will be gusty	☐
		Significance of wind-chill	☐
Humidity level	☐		
Humidex	☐	Visibility information	☐
		Amount of sun	☐
Precipitation/Rain/Snow	☐	UV Index	☐
Amount of rain/snow	☐	Air quality	☐
Type of precipitation (rain/snow/hail)	☐		
When precipitation will start	☐	Expected weather changes	☐
When precipitation will end	☐	Storm expectations	☐
Whether precipitation will be heavy/light	☐		
Probability of precipitation	☐	Historical information	☐

Other:_____

_____

9a) What is your main source of weather information for work-related decisions? **[DO NOT READ. CHECK ONE ONLY]**

9b) From what other sources do you get work-related weather information? **[DO NOT READ. CHECK AS MANY AS APPLY.]**

	9a) Primary source	9b) Secondary source
Television – General mention	☐	☐
Television – Weather network	☐	☐
Television – Local Environment Canada cable channel	☐	☐
Radio	☐	☐
Newspaper	☐	☐
Internet Access	☐	☐
WeatherRadio Canada	☐	☐
WeatherCopy Canada	☐	☐
Contact Environment Canada weather office	☐	☐
Telephone – General mention	☐	☐
Telephone – 1-800 number	☐	☐
Telephone – 1-900 number	☐	☐
Environment Canada recorded tape	☐	☐
Family member	☐	☐
Directly from employer	☐	☐

9a) Other Primary : _____

9b) Other Secondary: _____

10a) We would like you to think of the four seasons. On a scale of 1 to 10, where 10 means "very important" and 1 means "not important at all", how important are weather forecasts to you for each of the following seasons? **[START RANDOMLY, AND THEN PROCEED IN ORDER.]**

	Not Very important									Very important
Spring	1	2	3	4	5	6	7	8	9	10
Summer	1	2	3	4	5	6	7	8	9	10
Fall	1	2	3	4	5	6	7	8	9	10
Winter	1	2	3	4	5	6	7	8	9	10

10b) Now we would like you to think of the changes between seasons. On a scale of 1 to 10, where 10 means "very important" and 1 means "not important at all", how important are weather forecasts to you for each of the following change of seasons? **[START RANDOMLY, AND THEN PROCEED IN ORDER.]**

	Not Very important									Very important
Change from Spring to Summer	1	2	3	4	5	6	7	8	9	10
Change from Summer to Fall	1	2	3	4	5	6	7	8	9	10
Change from Fall to Winter	1	2	3	4	5	6	7	8	9	10
Change from Winter to Spring	1	2	3	4	5	6	7	8	9	10

11. Say you are planning a vacation six months from now to an area of Canada that you've never been to. Would the kind of weather you'd likely experience in six months from now in that location be very important, somewhat important not very important or not important at all to you in planning your holiday?

Very important ☐
Somewhat important ☐
Not very important ☐
Not important at all ☐

12. If you did need this kind of weather information now for your trip in six months, from where do you think you could get this type of information? **(DO NOT READ – CHECK ALL THAT APPLY.)**

     Weather Office   ☐
     Library   ☐
     Atlas   ☐
     CAA   ☐
     Travel Agent   ☐
     Travel Books   ☐
     Television – General mention   ☐
     Weather Network – Specific mention   ☐
     Radio   ☐
     Newspaper   ☐
     Internet Access – The Web (WWW)   ☐
     WeatherRadio Canada   ☐
     WeatherCopy Canada   ☐
     Environment Canada recorded tape   ☐
     Contact Environment Canada
       weather office   ☐
     Family member   ☐
     Other   ☐
     Don't know   ☐

13. Besides vacation planning, have you ever obtained this kind of long term weather information for other purposes?

     Yes   ☐
     No   ☐ **GO TO NEXT SECTION**
     Don't know ☐

14. For what use?

_____

SECTION TWO: WEATHER WARNING INFORMATION

We would like to talk to you about weather warnings a specific type of weather forecast that Environment Canada provides to all Canadians …

1. First of all, what do you think of when you see or hear the words "Weather Warning" as part of a weather report? What does a "Weather Warning" mean to you? **[PROBE AND CLARIFY]** Anything else?

2a) From what source are you most likely to receive a "Weather Warning"? **[DO NOT READ LIST. CHECK ONE]**

2b) From what other sources are you likely to receive "Weather Warnings"? **[DO NOT READ LIST. CHECK ALL THAT APPLY]**

	2a) Primary source	2b) Secondary source
Television – General mention	☐	☐
Television – Weather network	☐	☐

Television – Local Environment Canada cable channel ☐ ☐
Radio ☐ ☐
Newspaper ☐ ☐
Internet Access ☐ ☐
WeatherRadio Canada ☐ ☐
WeatherCopy Canada ☐ ☐
Contact Environment Canada weather office ☐ ☐
Telephone – General mention ☐ ☐
Telephone – 1-800 number ☐ ☐
Telephone – 1-900 number ☐ ☐
Environment Canada recorded tape ☐ ☐
Family member ☐ ☐
Directly from employer ☐ ☐

[ROTATE SUMMER AND WINTER WARNING SECTIONS RESPONDENT TO RESPONDENT. IF CONDUCTING SUMMER WARNINGS, START BELOW. IF CONDUCTING WINTER WARNINGS, GO TO QUESTION 9.]

We would like you to think of a summer weather situation in which you hear that a Weather Warning is in effect for an approaching summer storm.

3. Of all the times that you have heard a summer storm warning for your area, how often does the summer storm actually occur in your area? Would you say that it occurs...

Always ☐
Most of the time ☐
About half of the time ☐
Less than half the time ☐
Rarely ☐
Never ☐
[DON'T READ]
Don't know / No answer ☐

4. How often would you say that you receive enough notice in order to properly react to a warning about a summer storm heading toward your area?

Always ☐
Most of the time ☐
About half of the time ☐
Less than half the time ☐
Rarely ☐
Never ☐
[DON'T READ]
Don't know / No answer ☐

5. We would like to know how clear and well-communicated various aspects of a summer storm warning are presented to you. Based on what you know or have experienced, are the following communicated very well, somewhat well, not very well or not well at all? **[ROTATE]**

	Very well	Somewhat well	Not very well	Not at all well	Don't Know
The area that the summer storm is going to affect	☐	☐	☐	☐	☐
The severity of the summer storm	☐	☐	☐	☐	☐
When the summer storm will be in your area	☐	☐	☐	☐	☐
How long the summer storm will last in your area	☐	☐	☐	☐	☐
The type of damage expected from the summer storm	☐	☐	☐	☐	☐

What actions to take to ensure the safety of
   yourself, your family and your property?   ☐   ☐   ☐   ☐   ☐

6.   What other type of information do you feel you need to hear as part of the warning message in order to properly prepare and respond to a summer storm warning? **[PROBE AND CLARIFY]**

_____

7a) When you hear a summer storm warning for your area, how much advance notice do you need in order to ensure your safety? Would you need... **[READ LIST]**

Less than five minutes   ☐
5 minutes to under 15 minutes   ☐
15 minutes to under 30 minutes   ☐
30 minutes to under 1 hour   ☐
1 hour or more   ☐

[DO NOT READ] Don't Know   ☐

7b) What is the minimum a mount of time that you would accept in order to prepare for a summer storm warning for your area? Would you say it is... [READ LIST]

Less than five minutes   ☐
5 minutes to under 15 minutes   ☐
15 minutes to under 30 minutes   ☐
30 minutes to under 1 hour   ☐
1 hour or more   ☐

[DO NOT READ] Don't Know   ☐

8a) Based on what you can recall and your own experience over the last two years with summer storm warnings, generally did you have enough time to respond?

Yes   ☐ **GO TO NEXT SECTION**
No   ☐

DON'T READ Don't Know   ☐

8b) How much more time did you require? Would you require... **[READ LIST.]**

Less than five minutes   ☐
5 minutes to under 15 minutes   ☐
15 minutes to under 30 minutes   ☐
30 minutes to under 1 hour   ☐
1 hour or more   ☐

[DO NOT READ] Don't Know   ☐

Now, we would like you to consider a winter weather situation, and you hear that a winter storm warning is in effect for an approaching winter storm.

9. Of all the times that you have heard a Winter Storm warning in your area, how often does this winter storm occur? Would you say that it occurs …

Always ☐
Most of the time ☐
About half of the time ☐
Less than half the time ☐
Rarely ☐
Never ☐

[DO NOT READ] Don't Know ☐

10. How often would you say that you have received enough notice in order to properly react to a warning about a winter storm warning in your area?

Always ☐
Most of the time ☐
About half of the time ☐
Less than half the time ☐
Rarely ☐
Never ☐

[DO NOT READ] Don't Know ☐

11. We would like to know how clear and well communicated various aspects of a winter storm warning are presented to you. Based on what you know and have experienced, are the following communicated very well, somewhat well not very well or not well at all? **[ROTATE]**

	Very well	Somewhat well	Not very well	Not at all well	Don't Know
The area that the winter storm is going to affect	☐	☐	☐	☐	☐
The severity of the winter storm	☐	☐	☐	☐	☐
When the winter storm will be in your area	☐	☐	☐	☐	☐
How long the winter storm will last in your area	☐	☐	☐	☐	☐
The type of damage expected from the winter storm	☐	☐	☐	☐	☐
What actions to take to ensure the safety of yourself, your family and your property	☐	☐	☐	☐	☐

12. What other type of information do you feel you need to hear as part of the warning message in order to properly prepare and respond to a Winter storm Warning? **[PROBE AND CLARIFY]**

_____

13a) When you hear a winter storm warning for your area, how much advance notice do you need in order to ensure your safety? Would you say you need... **[READ LIST. CHECK ONE ONLY.]**

Less than one hour ☐
One to three hours ☐
Over three hours to six hours ☐
Over six hours to 12 hours ☐
Over 12 hours to 24 hours ☐
Over 24 hours to 48 hours ☐
Over 48 hours ☐

13b) What is the <u>minimum</u> amount of time that you would accept in order to prepare for a winter storm warning for your area? Would you say it is... **[READ LIST. CHECK ONE ONLY.]**

Less than one hour ☐
One to three hours ☐
Over three hours to six hours ☐
Over six hours to 12 hours ☐
Over 12 hours to 24 hours ☐
Over 24 hours to 48 hours ☐
Over 48 hours ☐

[DO NOT READ] Don't Know   ☐

14a) Based on what you can recall and your own experience with winter storm warnings, generally did you have enough time to respond?

Yes   ☐ **GO TO NEXT SECTION**
No   ☐

DON'T READ Don't Know ☐

14b)   How much more time did you require? Would you require.. **[READ LIST. CHECK ONE ONLY]**

Less than one hour ☐
One to three hours ☐
Over three hours to six hours ☐
Over six hours to 12 hours ☐
Over 12 hours to 24 hours ☐
Over 24 hours to 48 hours ☐
Over 48 hours ☐

[DO NOT READ] Don't Know   ☐

SECTION 3A: WEATHER FORECAST INFORMATION

SUMMERTIME SCENARIO

We would like to know your opinions about the accuracy of various types of weather forecasts. Consider a summer forecast that you hear in July for your area.

1a) So, let's say that this forecast states that the anticipated high for the day would be 25 degrees. Suppose the actual high is not 25, but is some temperature <u>less</u> than 25 degrees. At what <u>temperature below 25</u> would you consider the forecast <u>inaccurate</u>?

[WRITE IN]   **[DON'T READ]** Don't Know ☐

1b) Now suppose the actual high is not 25, but is some temperature <u>more</u> than 25 degrees. At what t<u>emperature above 25</u> would you consider the forecast <u>inaccurate</u>?

[WRITE IN]   **[DON'T READ]** Don't Know ☐

2a) Say the forecast states that the anticipated overnight low would be 20 degrees. Suppose the actual low is not 20, but is some temperature <u>less</u> than 20 degrees. At what <u>temperature below 20</u> would you consider the forecast <u>inaccurate</u>?

[WRITE IN]   **[DON'T READ]** Don't Know ☐

2b) Now suppose that the actual overnight low is not 20, but is some temperature <u>more</u> than 20 degrees. At what <u>temperature above 20</u> would you consider the forecast <u>inaccurate</u>?

[WRITE IN] **[DON'T READ]** Don't Know ☐

3a) Say the forecast mentioned that the anticipated wind speed would be 30 kilometers per hour. Suppose that the actual wind-speed is not 30, but is at some speed <u>less</u> than 30. At what <u>speed below 30</u> would you consider the forecast <u>inaccurate</u>?

[WRITE IN] **[DON'T READ]** Don't Know ☐

3b) Now suppose that the actual wind speed is not 30, but is at some speed <u>more</u> than 30. At what <u>speed above 30</u> would you consider the forecast <u>inaccurate</u>?

[WRITE IN] **[DON'T READ]** Don't Know ☐

4. Say the forecast mentioned that the *wind would be coming from the west.* Would you consider the forecast to be accurate or not accurate if the wind <u>actually</u> came from… **[READ LIST. START RANDOMLY AND CONTINUE IN ORDER]**

	Accurate	Not Accurate	Don't Know
the South	☐	☐	☐
the Southwest	☐	☐	☐
the Northwest	☐	☐	☐
the North	☐	☐	☐

5. Say the forecast said *"rain beginning in the afternoon".* Would you consider the forecast to be accurate or not accurate if the rain <u>actually</u> began… **[READ LIST. START RANDOMLY AND CONTINUE IN ORDER]**

	Accurate	Not Accurate	Don't Know
In the morning	☐	☐	☐
Around noon	☐	☐	☐
Mid afternoon	☐	☐	☐
In the late afternoon	☐	☐	☐
In the evening	☐	☐	☐
If no rain occurred throughout the day or evening	☐	☐	☐

6. Say the forecast says *"Sunny with afternoon cloudy periods".* Would you consider the forecast accurate or not accurate if it was… **[ROTATE. READ LIST]**

	Accurate	Not Accurate	Don't Know
Sunny all day	☐	☐	☐
Cloudy all day	☐	☐	☐
Cloudy in the morning and sunny in the afternoon	☐	☐	☐

7. Say that *heavy rain* with over 50 millimeters of rainfall over the next 24 hours is forecast. Would you consider the forecast to be accurate or not accurate if *actually…* **[READ LIST. ROTATE]**

	Accurate	Not Accurate	Don't Know
The ground was slightly wet with 5mm of rainfall	☐	☐	☐
There are some puddles with 15 mm of rainfall	☐	☐	☐
A lot of water has accumulated with 30 mm of rainfall	☐	☐	☐
Basements have been flooded with over 55 mm of rainfall	☐	☐	☐

8. Say the forecast said the probability of precipitation was 70% for today. When you hear that the *probability of precipitation for today is 70%,* what does that mean to you?   [READ LIST. ROTATE. READ NUMBERS. CHECK ONE ONLY.]

    1   Rain was expected to occur for 70% of the day   ☐

    2   There is a 70% chance that the rain will occur at
         a particular geographic point in the forecast area today   ☐

    3   There is a 70% chance that rain will occur somewhere in the forecast area today   ☐

    4   70% of the forecast area is expected to receive some rain today   ☐

    [DON'T READ] Don't know / No answer   ☐

9. *And continue to think about the summer...* Which forecast do you use most to plan for special activities, events or weekends?
    [READ LIST. CHECK ONE ONLY.]

    The forecast for that particular day   ☐
    The forecast for TWO DAYS in advance   ☐
    The forecast for THREE OR MORE days in advance   ☐

    [DON'T READ] Don't Know   ☐

10. We would like to know how useful various parts of a summer weather forecast are to you. On a scale of 1 to 10, where 10 is "extremely useful" and 1 is "not useful at all" how useful are each of the following parts of a weather forecast and other summer weather information... [READ LIST. ROTATE]

	Not Useful At All									Extremely Useful
The overnight low temperature	1	2	3	4	5	6	7	8	9	10
The daytime high temperature	1	2	3	4	5	6	7	8	9	10
If it is going to rain	1	2	3	4	5	6	7	8	9	10
Whether the rain is going to be light or heavy	1	2	3	4	5	6	7	8	9	10
The amount of rain expected	1	2	3	4	5	6	7	8	9	10
When the rain will start and when it will end	1	2	3	4	5	6	7	8	9	10
The probability of precipitation	1	2	3	4	5	6	7	8	9	10
The amount of sun or cloud expected	1	2	3	4	5	6	7	8	9	10
The humidity level	1	2	3	4	5	6	7	8	9	10
The UV index	1	2	3	4	5	6	7	8	9	10
If a change in the weather is expected	1	2	3	4	5	6	7	8	9	10
The wind direction	1	2	3	4	5	6	7	8	9	10
The wind speed	1	2	3	4	5	6	7	8	9	10
A reduction of visibility due to fog	1	2	3	4	5	6	7	8	9	10

11. Now we would like to know how accurate summer weather forecasts are on each of the following weather measures. In your experience, on a scale of 1 to 10, where 10 is "extremely accurate" and 1 is "not accurate at all" how accurate are each of the following parts of a weather forecast and other summer weather information... [READ LIST. ROTATE]

	Not Accurate At All									Extremely Accurate	Don't Know
The overnight low temperature	1	2	3	4	5	6	7	8	9	10	☐
The daytime high temperature	1	2	3	4	5	6	7	8	9	10	☐
If it is going to rain	1	2	3	4	5	6	7	8	9	10	☐
Whether the rain is going to be light or heavy	1	2	3	4	5	6	7	8	9	10	☐

The amount of rain expected	1	2	3	4	5	6	7	8	9	10	☐
When the rain will start and when it will end	1	2	3	4	5	6	7	8	9	10	☐
The probability of precipitation	1	2	3	4	5	6	7	8	9	10	☐
The amount of sun or cloud expected	1	2	3	4	5	6	7	8	9	10	☐
The humidity level	1	2	3	4	5	6	7	8	9	10	☐
The UV index	1	2	3	4	5	6	7	8	9	10	☐
If a change in the weather is expected	1	2	3	4	5	6	7	8	9	10	☐
The wind direction	1	2	3	4	5	6	7	8	9	10	☐
The wind speed	1	2	3	4	5	6	7	8	9	10	☐
A reduction of visibility due to fog	1	2	3	4	5	6	7	8	9	10	☐

SECTION 3B: WEATHER FORECAST INFORMATION

FALL/SPRING TIME SCENARIO

We would like to know your opinions about the accuracy of various types of weather forecasts. Consider a fall or spring forecast that you hear in October or March for your area.

1a) So, let's say that this forecast states that the anticipated high for the day would be plus one. Suppose the actual high is not plus one, but is some temperature <u>less</u> than plus one. At what <u>temperature below plus one</u> would you consider the forecast <u>inaccurate</u>?
[CONFIRM PLUS OR MINUS WITH RESPONDENT]

PLUS　　　　　　　　**[WRITE IN]**　　MINUS　　　　**[WRITE IN]**　　Don't Know ☐

1b) Now suppose the actual high is not plus one, but is some temperature <u>more</u> than plus one. At what <u>temperature above plus one</u> would you consider the forecast <u>inaccurate</u>?

PLUS　　　　　　　　**[WRITE IN]**　　**[DON'T READ]** Don't Know ☐

2a) Say the forecast states that the anticipated overnight low would be minus five degrees. Suppose the actual low is not minus five, but is some temperature <u>less</u> than minus five. At what <u>temperature below minus five</u> would you consider the forecast <u>inaccurate</u>?

MINUS　　　　　　　**[WRITE IN]**　　**[DON'T READ]** Don't Know ☐

2b) Now suppose that the actual overnight low is not minus five, but is some temperature <u>more</u> than minus five. At what <u>temperature above minus five</u> would you consider the forecast <u>inaccurate</u>?
[CONFIRM PLUS OR MINUS WITH RESPONDENT.]

PLUS　　　　　　　　**[WRITE IN]**　　MINUS　　　　**[WRITE IN]**　　Don't Know ☐

3a) Say the forecast mentioned that the anticipated wind speed would be 30 kilometers per hour. Suppose that the actual wind-speed is not 30, but is at some speed <u>less</u> than 30. At what <u>speed below 30</u> would you consider the forecast <u>inaccurate</u>?

　　　　　　　　[WRITE IN]　　**[DON'T READ]** Don't Know ☐

3b) Now suppose that the actual wind speed is not 30, but is at some speed <u>more</u> than 30. At what <u>speed above 30</u> would you consider the forecast <u>inaccurate</u>?

[WRITE IN]   **[DON'T READ]** Don't Know ☐

4. Say the forecast mentioned that the *wind would be coming from the west.* Would you consider the forecast to be accurate or not accurate if the wind <u>actually</u> came from… **[READ LIST. START RANDOMLY AND CONTINUE IN ORDER]**

	Accurate	Not Accurate	Don't Know
the South	☐	☐	☐
the Southwest	☐	☐	☐
the Northwest	☐	☐	☐
the North	☐	☐	☐

5. Say the forecast said "*wet snow developing in the afternoon*". Would you consider the forecast to be accurate or not accurate if the wet snow <u>actually</u> began… **[READ LIST. START RANDOMLY AND CONTINUE IN ORDER]**

	Accurate	Not Accurate	Don't Know
In the morning	☐	☐	☐
Around noon	☐	☐	☐
Mid afternoon	☐	☐	☐
In the late afternoon	☐	☐	☐
In the evening	☐	☐	☐
If no wet snow occurred throughout the day or evening	☐	☐	☐

6. Say the forecast says "*Sunny with afternoon cloudy periods*". Would you consider the forecast accurate or not accurate if it was… **[ROTATE. READ LIST]**

	Accurate	Not Accurate	Don't Know
Sunny all day	☐	☐	☐
Cloudy all day	☐	☐	☐
Cloudy in the morning and sunny in the afternoon	☐	☐	☐

7. Say that *freezing rain* is forecast. Would you consider the forecast to be accurate or not accurate if the precipitation was <u>actually</u>… **[READ LIST. ROTATE]**

	Accurate	Not Accurate	Don't Know
Just rain	☐	☐	☐
Just snow	☐	☐	☐
Mix of snow and rain	☐	☐	☐
Freezing rain	☐	☐	☐
Freezing drizzle	☐	☐	☐
No precipitation occurred at all	☐	☐	☐

8. Say the forecast said the probability of precipitation was 70% for today. When you hear that the probability of precipitation for today is 70%, what does that mean to you? **[READ LIST. ROTATE. READ NUMBERS. CHECK ONE ONLY.]**

1   Rain was expected to occur for 70% of the day   ☐

2   There is a 70% chance that the rain will occur at
    a particular geographic point in the forecast area today   ☐

3   There is a 70% chance that rain will occur somewhere in the forecast area today   ☐

4   70% of the forecast area is expected to receive some rain today   ☐

[DON'T READ] Don't know / No answer   ☐

9. *And continue to think about the fall and/or spring...* Which forecast do you use most to plan for special activities, events or weekends? Would it be... **[READ LIST. CHECK ONE ONLY.]**

The forecast for that particular day   ☐
The forecast for TWO DAYS in advance   ☐
The forecast for THREE OR MORE days in advance   ☐

[DON'T READ] Don't Know   ☐

10. We would like to know how useful various parts of a fall or spring weather forecast are to you. On a scale of 1 to 10, where 10 is "extremely useful" and 1 is "not useful at all" how useful are each of the following parts of a weather forecast and other fall or spring weather information... **[READ LIST. ROTATE]**

	Not Useful At All									Extremely Useful
The overnight low temperature	1	2	3	4	5	6	7	8	9	10
The daytime high temperature	1	2	3	4	5	6	7	8	9	10
When the temperature will cross the zero degree Celsius mark	1	2	3	4	5	6	7	8	9	10
If there is going to be some precipitation	1	2	3	4	5	6	7	8	9	10
Whether the precipitation is going to be light or heavy	1	2	3	4	5	6	7	8	9	10
What the precipitation type will be	1	2	3	4	5	6	7	8	9	10
The amount of precipitation expected	1	2	3	4	5	6	7	8	9	10
When the precipitation will start and when it will end	1	2	3	4	5	6	7	8	9	10
The probability of precipitation	1	2	3	4	5	6	7	8	9	10
The amount of sun or cloud expected	1	2	3	4	5	6	7	8	9	10
The humidity level	1	2	3	4	5	6	7	8	9	10
The wind-chill	1	2	3	4	5	6	7	8	9	10
If a change in the weather is expected	1	2	3	4	5	6	7	8	9	10
The wind direction	1	2	3	4	5	6	7	8	9	10
The wind speed	1	2	3	4	5	6	7	8	9	10
The amount of snow currently on the ground	1	2	3	4	5	6	7	8	9	10
A reduction of visibility due to fog	1	2	3	4	5	6	7	8	9	10

11. Now we would like to know how accurate spring and/or fall weather forecasts are on each of the following weather measures. In your experience, on a scale of 1 to 10, where 10 is "extremely accurate" and 1 is "not accurate at all" how accurate are each of the following parts of a weather forecast and other fall or spring weather information... **[READ LIST. ROTATE]**

	Not Accurate At All									Extremely Accurate	Don't Know
The overnight low temperature	1	2	3	4	5	6	7	8	9	10	☐
When the temperature will cross the zero degree Celsius mark	1	2	3	4	5	6	7	8	9	10	☐
If there is going to be some precipitation	1	2	3	4	5	6	7	8	9	10	☐

Whether the precipitation is going to be light or heavy	1	2	3	4	5	6	7	8	9	10	☐
What the precipitation type will be	1	2	3	4	5	6	7	8	9	10	☐
The amount of precipitation expected	1	2	3	4	5	6	7	8	9	10	☐
When the precipitation will start and when it will end	1	2	3	4	5	6	7	8	9	10	☐
The probability of precipitation	1	2	3	4	5	6	7	8	9	10	☐
The amount of sun or cloud expected	1	2	3	4	5	6	7	8	9	10	☐
The humidity level	1	2	3	4	5	6	7	8	9	10	☐
The wind-chill	1	2	3	4	5	6	7	8	9	10	☐
If a change in the weather is expected	1	2	3	4	5	6	7	8	9	10	☐
The wind direction	1	2	3	4	5	6	7	8	9	10	☐
The wind speed	1	2	3	4	5	6	7	8	9	10	☐
The amount of snow currently on the ground	1	2	3	4	5	6	7	8	9	10	☐
A reduction of visibility due to fog	1	2	3	4	5	6	7	8	9	10	☐

SECTION 3C: WEATHER FORECAST INFORMATION

WINTER TIME SCENARIO

We would like to know your opinions about the accuracy of various types of weather forecasts. Consider a winter forecast that you hear in January for your area.

1a) So, let's say that this forecast states that the anticipated high for the day would be minus 5 degrees Celsius. Suppose the actual high is not minus 5, but is some temperature <u>less</u> than minus 5. At what <u>temperature below minus 5</u> would you consider the forecast <u>inaccurate</u>?

MINUS   **[WRITE IN]**   **[DON'T READ]** Don't Know ☐

1b) Now suppose the actual high is not minus 5, but is some temperature <u>more</u> than minus 5. At what <u>temperature above minus 5</u> would you consider the forecast <u>inaccurate</u>? **[CONFIRM PLUS OR MINUS WITH RESPONDENT]**

PLUS   **[WRITE IN]** MINUS   **[WRITE IN]**   Don't Know ☐

2a) Say the forecast states that the anticipated overnight low would be minus 20 degrees Celsius. Suppose the actual low is not minus 20, but is some temperature <u>less</u> than minus 20. At what <u>temperature below minus 20</u> would you consider the forecast <u>inaccurate</u>?

MINUS   **[WRITE IN]**   **[DON'T READ]** Don't Know ☐

2b) Now suppose that the actual overnight low is not minus 20, but is some temperature <u>more</u> than minus 20. At what <u>temperature above minus 20</u> would you consider the forecast <u>inaccurate</u>? **[CONFIRM PLUS OR MINUS WITH RESPONDENT.]**

PLUS   **[WRITE IN]** MINUS   **[WRITE IN]**   Don't Know ☐

3a) Say the forecast mentioned that the anticipated wind speed would be 30 kilometers per hour. Suppose that the actual wind-speed is not 30, but is at some speed <u>less</u> than 30. At what <u>speed below 30</u> would you consider the forecast <u>inaccurate</u>?

[WRITE IN]   **[DON'T READ]** Don't Know ☐

3b) Now suppose that the actual wind speed is not 30, but is at some speed <u>more</u> than 30. At what <u>speed above 30</u> would you consider the forecast <u>inaccurate</u>?

[WRITE IN]   **[DON'T READ]** Don't Know ☐

4. Say the forecast mentioned that the wind would be coming from the west. Would you consider the forecast to be accurate or <u>not</u> accurate if the wind <u>actually</u> came from… **[READ LIST. START RANDOMLY AND CONTINUE IN ORDER]**

	Accurate	Not Accurate	Don't Know
the South	☐	☐	☐
the Southwest	☐	☐	☐
the Northwest	☐	☐	☐
the North	☐	☐	☐

5. Say the forecast said *"snow beginning in the afternoon".* Would you consider the forecast to be accurate or not accurate if the wet snow <u>actually</u> began… **[READ LIST. START RANDOMLY AND CONTINUE IN ORDER]**

	Accurate	Not Accurate	Don't Know
In the morning	☐	☐	☐
Around noon	☐	☐	☐
Mid afternoon	☐	☐	☐
In the late afternoon	☐	☐	☐
In the evening	☐	☐	☐
If no snow occurred throughout the day or evening	☐	☐	☐

6. Say the forecast says *"Sunny with afternoon cloudy periods".* Would you consider the forecast accurate or not accurate if it was… **[ROTATE. READ LIST]**

	Accurate	Not Accurate	Don't Know
Sunny all day	☐	☐	☐
Cloudy all day	☐	☐	☐
Cloudy in the morning and sunny in the afternoon	☐	☐	☐

7. Say that *heavy snow* is forecast. Would you consider the forecast to be accurate or not accurate if the precipitation was <u>actually</u>… **[READ LIST. ROTATE]**

	Accurate	Not Accurate	Don't Know
The ground was slightly covered	☐	☐	☐
There is some snow on the ground	☐	☐	☐
There is snow on the streets that needs to be cleaned	☐	☐	☐
Snow has piled up significantly	☐	☐	☐
People are stranded because of the extreme amount of snow	☐	☐	☐
No precipitation occurred at all	☐	☐	☐

8.   Say the forecast said the probability of precipitation was 70% for today. When you hear that the *probability of precipi-
     tation for today is 70%,* what does that mean to you?   [**READ LIST. ROTATE. READ NUMBERS. CHECK ONE
     ONLY.**]

   1   Snow was expected to occur for 70% of the day   ☐

   2   There is a 70% chance that snow will occur at
        a particular geographic point in the forecast area today   ☐

   3   There is a 70% chance that snow will occur somewhere
        in the forecast area today   ☐

   4   70% of the forecast area is expected to receive some snow today   ☐

   [**DON'T READ**] Don't know / No answer   ☐

9.   *And continue to think about the winter...* Which forecast do you use most to plan for special activities, events or weekends?
     Would it be... [**READ LIST. CHECK ONE ONLY.**]

   The forecast for that particular day   ☐
   The forecast for TWO DAYS in advance   ☐
   The forecast for THREE OR MORE days in advance   ☐

   [**DON'T READ**] Don't Know   ☐

10. We would like to know how useful various parts of a winter weather forecast are to you. On a scale of 1 to 10, where 10 is
     "extremely useful" and 1 is "not useful at all" how useful are each of the following parts of a weather forecast and other winter
     weather information... [**READ LIST. ROTATE**]

	Not Useful At All								Extremely Useful	
The overnight low temperature	1	2	3	4	5	6	7	8	9	10
The daytime high temperature	1	2	3	4	5	6	7	8	9	10
If it is going to snow	1	2	3	4	5	6	7	8	9	10
Whether the snow is going to be light or heavy	1	2	3	4	5	6	7	8	9	10
The amount of snow expected	1	2	3	4	5	6	7	8	9	10
When the snow will start and when it will end	1	2	3	4	5	6	7	8	9	10
The probability of precipitation	1	2	3	4	5	6	7	8	9	10
The amount of sun or cloud expected	1	2	3	4	5	6	7	8	9	10
The humidity level	1	2	3	4	5	6	7	8	9	10
The wind-chill	1	2	3	4	5	6	7	8	9	10
If a change in the weather is expected	1	2	3	4	5	6	7	8	9	10
The wind direction	1	2	3	4	5	6	7	8	9	10
The wind speed	1	2	3	4	5	6	7	8	9	10
The amount of snow currently on the ground	1	2	3	4	5	6	7	8	9	10
A reduction of visibility due to blowing snow	1	2	3	4	5	6	7	8	9	10

11. Now we would like to know how accurate winter weather forecasts are on each of the following weather measures. In your
experience, on a scale of 1 to 10, where 10 is "extremely accurate" and 1 is "not accurate at all" how accurate are each of the follow-
ing parts of a weather forecast and other winter weather information... [**READ LIST. ROTATE**]

	Not Accurate At All									Extremely Accurate	Don't Know
The overnight low temperature	1	2	3	4	5	6	7	8	9	10	☐
The daytime high temperature	1	2	3	4	5	6	7	8	9	10	☐
If it is going to snow	1	2	3	4	5	6	7	8	9	10	☐
Whether the snow is going to be light or heavy	1	2	3	4	5	6	7	8	9	10	☐
The amount of snow expected	1	2	3	4	5	6	7	8	9	10	☐
When the snow will start and when it will end	1	2	3	4	5	6	7	8	9	10	☐
The probability of precipitation	1	2	3	4	5	6	7	8	9	10	☐
The amount of sun or cloud expected	1	2	3	4	5	6	7	8	9	10	☐
The humidity level	1	2	3	4	5	6	7	8	9	10	☐
The wind-chill	1	2	3	4	5	6	7	8	9	10	☐
If a change in the weather is expected	1	2	3	4	5	6	7	8	9	10	☐
The wind direction	1	2	3	4	5	6	7	8	9	10	☐
The wind speed	1	2	3	4	5	6	7	8	9	10	☐
The amount of snow currently on the ground	1	2	3	4	5	6	7	8	9	10	☐
A reduction of visibility due to blowing snow	1	2	3	4	5	6	7	8	9	10	☐

SECTION FOUR: AIR QUALITY INFORMATION

We would like you to now think about the environment in your area

1a) Do you consider your local area to have an air pollution problem?

     Yes   ☐
     No   ☐ **GO TO QUESTION 2**

1b) What air pollution or air quality problems do you feel your area has?

2a) Two different types of air-quality information messages could be provided to you. First, <u>anticipated</u> or <u>expected</u> levels of pollution for the day could be provided, or information on the <u>actual</u> pollution levels as they are presently occurring could be provided. Would you prefer to have information on the anticipated pollution levels, on the current levels as they're happening, or on both?

     Anticipated or expected levels   ☐
     Actual levels   ☐
     Both   ☐

3a) Are you aware of any air quality or air pollution information sources available for your area that reflect the current conditions?

     Yes   ☐
     No   ☐ **GO TO QUESTION 6**

4. How often do you make a point of checking for information on the current levels of air pollution in your area?

    Several times a day   ☐
    Once a day   ☐
    Several time a week   ☐
    Once a week   ☐
    Less often than once a week   ☐
    Never   ☐

5. On a scale of 1 to 10, 1 being "Not at all satisfied" and 10 being "Extremely satisfied", how satisfied are you with all the infor-mation you see or hear now about the levels of air pollution in your area? **[CIRCLE ONE]**

    Not at all   Extremely
    satisfied   satisfied
    1   2   3   4   5   6   7   8   9   10

6. If you heard a message indicating high levels of air pollution, how likely are you to do each of the following?

	Very Likely	Somewhat Likely	Not Very Likely	Not Likely At All
Reduce time spent outdoors	☐	☐	☐	☐
Reduce car use	☐	☐	☐	☐
Carpool	☐	☐	☐	☐
Avoid using gas-powered equipment (lawnmowers, BBQs, etc..)	☐	☐	☐	☐

SECTION FIVE: ENVIRONMENT CANADA DELIVERY SERVICES

We would like to talk to you about various weather services that are available to you either by phone or electronically.

Free Recorded Local Weather Message

In most major urban centres, Environment Canada provides a free 24 hour recorded local weather forecast *accessible only over the telephone*. Callers in the local dialing area do not pay any charges. However, those calling from outside the local area must pay long distance charges to hear about weather that affects their area.

1. Are you aware of this *Environment Canada* ~~24 hour~~ recorded ~~local~~ weather forecast ~~service~~ *message only accessible over the telephone*?
    (Words in italics were added to the questionnaire during the field work, on March 5, 1997 – after a review of preliminary data seemed suspect)

    Yes   ☐
    No   ☐   **GO TO QUESTION 8**

2. Have you ever used it?

    Yes   ☐
    No   ☐   **GO TO QUESTION 8**

3. How often do you use it? **[READ LIST. CHECK ONE ONLY]**

   More than once a day ☐
   Once a day ☐
   Two or more times per week ☐
   Once a week ☐
   Two or more times a month ☐
   Once a month ☐
   Less often than once a month ☐

4. How often do you try to call this weather line and receive a busy signal? **[READ LIST]**

   Always ☐
   Most of the time ☐
   About half of the time ☐
   Less than half of the time ☐
   Rarely or never ☐

5. On a scale of 1 to 10, where 10 is "extremely satisfied" and 1 is "not satisfied at all", how satisfied are you with the type of information provided through this service?

Not at all satisfied								Extremely satisfied	
1	2	3	4	5	6	7	8	9	10

6. On a scale of 1 to 10, where 10 is "extremely satisfied" and 1 is "not satisfied at all", how satisfied are you with the accessibility of weather information provided by this service?

Not at all satisfied								Extremely satisfied	
1	2	3	4	5	6	7	8	9	10

7. On a scale of 1 to 10, where 10 is "extremely satisfied" and 1 is "not satisfied at all", how satisfied are you with the format and the presentation of the weather information provided by this service?

Not at all satisfied								Extremely satisfied	
1	2	3	4	5	6	7	8	9	10

8. For budgetary reasons, Environment Canada cannot provide such a service free of long distance charges uniformly across Canada to smaller centres. Do you think that Environment Canada should… **[READ AND ROTATE]**

   Require everyone to pay, even if someone calls from within their local area ☐

   or keep it as it currently is … that is callers from the local calling area are not charged, but callers from outside the area are charged long distance ☐ **GO TO QUESTION 10**

   [DO NOT READ] No charge/free/1-800 number ☐ **GO TO QUESTION 10**

   [DO NOT READ] Don't Know ☐ **GO TO QUESTION 10**

9a) Would you prefer to pay a fixed fee per call or a charge per minute?

Fixed fee   ☐
Charge per minute   ☐ **GO TO QUESTION 9C**

[DO NOT READ]   Both   ☐
[DO NOT READ]   Neither   ☐ **GO TO QUESTION 10**

9b) How much would you be willing to pay <u>per call</u>? Would it be... **[READ LIST]**

Under $1.00   ☐
$1.00 – $1.99   ☐
$2.00 – $2.99   ☐
$3.00 – $3.99   ☐
$4.00 – $4.99   ☐
$5.00 or more   ☐

[DON'T READ]   Nothing   ☐ **GO TO QUESTION 10**
[DON'T READ]   Don't Know   ☐ **GO TO QUESTION 10**

IF CHARGE PER MINUTE ABOVE...

9c) How much per minute would you be willing to pay for this service? Would it be... **[READ LIST]** **(IF ASKED, THE AVERAGE LENGTH IS 3 MINUTES)**

50 cents per minute   ☐
$1 per minute   ☐
$2 per minute   ☐
$3 per minute   ☐

[DON'T READ]   Nothing   ☐
[DON'T READ]   Don't Know   ☐

10. So that Environment Canada does not charge all users for this service, commercial advertising needs to be played on this line. Do you think this is .. **[READ LIST]**

An excellent idea   ☐
A good idea   ☐
A fair idea   ☐
A poor idea   ☐

[DON'T READ]   Don't know   ☐

<u>**Environment Canada's New 1- 900 User-Pay Telephone Weather Services**</u>

Environment Canada has recently launched a new national service, a 1-900 user-pay telephone weather service called "Weather Menu" which provides up-to-date weather and environmental bulletins.

(** If asked ..The phone number is 1-900-565-5000 in English/ 1-900-565-4000 in French called "*Meteo à la carte*")

11. Are you aware of this 1-900 User Pay Telephone service?

Yes   ☐
No   ☐ **GO TO QUESTION 14**

12. Have you ever used it?

Yes ☐
No ☐ **GO TO QUESTION 14**

13. How often do you use it? **[READ LIST. CHECK ONLY ONE)**

More than once a day ☐
Once a day ☐
Two or more times per week ☐
Once a week ☐
Less than once a week ☐

14. The cost for this type of service is 95 cents per minute. Do you think this is… **(READ LIST. CHECK ONE)**

Just right ☐
Too low ☐
Too high ☐

WeatherRadio

WEATHERADIO is an Environment Canada Service that broadcasts weather information 24 hours a day in many areas across Canada. A special radio must be purchased to receive these weather broadcasts.

(**If asked one can purchase a special receiver at major electronics retailers like RADIO SHACK)

15. Were you aware of Environment Canada's WEATHERADIO service?

Yes ☐
No ☐ **GO TO QUESTION 21**

16. Have you ever used it?

Yes ☐
No ☐ **GO TO QUESTION 21**

17. How often do you use it? **[READ LIST. CHECK ONE]**

More than once a day ☐
Once a day ☐
Two or more times per week ☐
Once a week ☐
Less than once a week ☐

18. On a scale of 1 to 10, where 10 is "extremely satisfied" and 1 is "not satisfied at all", how satisfied are you with the type of information provided on the WeatheRadio Broadcasts?

    (CODE ONLY <u>ONE</u>)

    1   2   3   4   5   6   7   8   9   10

19. On a scale of 1 to 10, where 10 is "extremely satisfied" and 1 is "not satisfied at all", how satisfied are you with the format and presentation of information on the WeatheRadio Broadcasts?

    (CODE ONLY <u>ONE</u>)

    1   2   3   4   5   6   7   8   9   10

20. On a scale of 1 to 10, where 10 is "extremely timely" and 1 is "not timely at all", how timely do you consider the 20 minute cycle for the WeatheRadio Broadcasts?

    (CODE ONLY <u>ONE</u>)

    1   2   3   4   5   6   7   8   9   10

<u>INTERNET "WEB" PAGES</u>

Environment Canada has a World Wide Web Internet site providing weather and environmental information.

[If they ask for the Universal Resource Locator, i.e. the URL, it is: http://www.ec.gc.ca/]

21. Were you aware of Environment Canada's Information centre on the INTERNET.

    Yes   ☐
    No   ☐ **GO TO DEMOGRAPHICS**

22. Do you use it to obtain weather information and/or forecasts?

    Yes   ☐
    No   ☐ **GO TO DEMOGRAPHICS**

23. How often do you use it for weather information or forecasts? **[READ LIST. CHECK ALL THAT APPLY.]**

    More than once a day   ☐
    Once a day   ☐
    Two or more times per week   ☐
    Once a week   ☐
    Less than once a week   ☐

24. On a scale of 1 to 10, where 10 is "extremely satisfied" and 1 is "not satisfied at all", how satisfied are you with the type of weather information provided on Environment Canada's Internet Pages?

    (CODE ONLY <u>ONE</u>)

    1   2   3   4   5   6   7   8   9   10

25. On a scale of 1 to 10, where 10 is "extremely satisfied" and 1 is "not satisfied at all", how satisfied are you with the format and presentation of weather information in Environment Canada's Internet Pages?

    (CODE ONLY <u>ONE</u>)

    1   2   3   4   5   6   7   8   9   10

G. DEMOGRAPHICS

THE FOLLOWING QUESTIONS ARE FOR CLASSIFICATION PURPOSES ONLY. YOUR ANSWERS ARE STRICTLY CONFIDENTIAL, AND WILL ONLY BE USED IN COMBINATION WITH OTHER RESPONSES.

1a) In which of the following age categories do you belong?

18 – 24	☐
25 – 34	☐
35 – 49	☐
50 – 64	☐
65 and over	☐

1a) Are you...

Married, or living common-law	60-1
Single	2
Divorced	3
Widowed	4
Separated	5

1b) How many people, **including yourself**, live in your household?

1	59-1	**SKIP TO QUESTION 3**
2	2	
3	3	
4	4	
5	5	
6 or more	6	

2a) Do you have any children living in your household under the age of 18?

Yes 61-1
No   2   **GO TO QUESTION 3**

2b) What ages are the children under the age of 18 that live in your household.
      [CHECK ALL THAT APPLY]

0 – 2 yrs old	62-1
3 – 5 yrs old	2
6 – 10 yrs old	3
11 – 15 yrs old	4
16 – 17 yrs old	5

3.   What is the highest level of education that you have attained?

Some elementary school	63-1
Completed elementary school	2
Some secondary school	3
Completed secondary school	4
Some post-secondary (community college, university)	5
Completed a post-secondary program (community college, university)	6

4a) Please indicate which of the following best describes your current status.

Working full-time outside the home   66-1
Working part-time outside the home   2
Working full or part time in your home   3
Unemployed/looking for work   4   **GO TO QUESTION 5a)**
Retired   5   **GO TO QUESTION 5a)**
Student   6   **GO TO QUESTION 5a)**

4b). What is your occupation? _____

5a) How many cars, trucks and vans are owned or leased by you or all members of your household?

None   68-1
1   2
2   3
3   4
4   5
5 or more   6

5b) And finally, in which category does your total annual household income fall before income taxes?

Under $25,000 per year   71-1
$25,000 to $49,999 per year   2
$50,000 to $74,999 per year   3
$75,000 to $99,999 per year   4
$100,00 or more per year   5

Refused   6

THANK

Finally, may I have your first name in case my supervisor needs to verify that I conducted this interview with you?

NAME:

PHONE:

Appendix 3
HONG KONG OBSERVATORY SURVEY

MAIN QUESTIONNAIRE

Q1 Do you usually read, watch or listen to weather reports ?
1. Yes   Go to Q2
2. No   End of questionnaire

Q2 From where do you usually obtain weather information of Hong Kong? Do you obtain from radio, television, newspaper, weather hotline, internet, pagers / mobile phones, or other sources? Any other? (up to 3 sources)

(For "weather hotline", probe : Is it Hong Kong Observatory's Dial-a-Weather hotlines 1878-200, 1878-202 and 1878-066, or Hong Kong Observatory's Information Enquiry System 2926-1133 or Hong Kong Telecom's 18-501 and 18-503, 18-508?)

(For "internet", probe : Is it Hong Kong Observatory's Homepage or other homepages?)

1. Radio
2. Television
3. Newspaper
4. Hong Kong Observatory's Dial-a-Weather hotlines (1878-200 / 202 / 066)
5. Information Enquiry System (2926-1133)
6. Hong Kong Telecom's 18 501 / 3 / 8
7. Observatory's Home Page
8. Other homepages
9. Pagers / Mobile Phones
10. Other sources (please specify)

Q3a Do you consider the weather forecasts of the Hong Kong Observatory over the past several months accurate or inaccurate? (Probe the degree)
1. Very accurate
2. Somewhat accurate
3. Average
4. Somewhat inaccurate
5. Very inaccurate
6. Don't know / no comment

Q3b What percentage of weather forecasts of the Hong Kong Observatory over the past several months do you consider accurate ?
1. _____ per cent
2. Don't know / No comment

Q4 Do you consider the following aspects of weather forecasts of the Hong Kong Observatory over the past several months accurate or inaccurate?

	Inaccurate	Accurate	Don't know/ No comment
Temperature			
Fine / Cloudy			
Rain storm forecasts / warning			
Typhoon prediction / warning			

Q5 How do you compare weather forecasts nowadays with those from the past 3 to 4 years ago? Is it more accurate, less accurate or about the same?
1. More accurate
2. About the same
3. Less accurate
4. Don't know / no comment

Q6 How satisfied are you with the services provided by the Hong Kong Observatory? If you rate on a scale of 0 to 10, with "5" being the passing mark and "10" being "excellent service", how many marks will you give?

End of Questionnaire