

Министерство природных ресурсов и экологии Российской Федерации
Федеральная служба по гидрометеорологии и мониторингу окружающей среды
ГУ «Сибирский научно-исследовательский гидрометеорологический институт»
(ГУ «СибНИГМИ»)

ГУ «Новосибирский центр по гидрометеорологии и мониторингу
окружающей среды с функциями регионального специализированного
метеорологического центра Всемирной службы погоды
(ГУ «Новосибирский ЦГМС-РСМЦ»)

УДК 556.5.048 556.5.06

№ госрегистрации 01200964773
Инв. №



ОТЧЁТ

О НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

СОЗДАТЬ МОДЕЛЬ И АВТОМАТИЗИРОВАТЬ ДОЛГОСРОЧНЫЙ ПРОГНОЗ
ПРИТОКА ВОДЫ ВО 2-3 КВАРТАЛЕ К НОВОСИБИРСКОМУ
ВОДОХРАНИЛИЩУ И РАСХОДОВ ВОДЫ ПО СТВОРУ ОБЬ-БАРНАУЛ


(заключительный)
Шифр темы 8-78 (раздел 2)


Руководитель темы,
гл. научн. сотрудник ГУ «СибНИГМИ»
д.ф.-м.н.


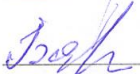
Л.Н. Романов

Новосибирск 2010

СПИСОК ИСПОЛНИТЕЛЕЙ

Ответственный исполнитель от ГУ "СибНИГМИ":
Главный научный сотрудник, д.ф.-м. н.  Л.Н. Романов

Ответственный исполнитель от ГУ "Новосибирский ЦГМС-РСМЦ":
Начальник отдела гидрологических прогнозов  В.Ф. Богданова

Исполнители:
Ведущий научный сотрудник ГУ "СибНИГМИ"  Г.М. Виноградова
Старший научный сотрудник ГУ "СибНИГМИ"  Е.Г. Бочкарева

Нормоконтроль,
зав. группой НТИ  Т.П. Панькова

РЕФЕРАТ

Отчет 28 с., Табл. 8., Илл.2 , Ист. 7.

МОДЕЛИРОВАНИЕ, ДОЛГОСРОЧНЫЙ ПРОГНОЗ, РАСХОД ВОДЫ, ПРИТОК, ВРЕМЕННОЙ РЯД, ЭМПИРИЧЕСКАЯ ВЕРОЯТНОСТЬ, СРЕДНИЙ РИСК.

Объектом исследования является приток воды в Новосибирское водохранилище с месячным разрешением.

Цель работы – разработка и создание модели прогноза притока и расходов воды во 2-3 кварталах на 2-3 месяца вперед.

В процессе выполнения темы были подготовлены данные и проведены многочисленные эксперименты по установлению прогностических и диагностических связей притока и расхода воды с различными совокупностями гидрометеорологических и аэрокосмических характеристик. Полученные результаты легли в основу общей конструкции прогностической модели. На материале 2004-2008 годов были проведены испытания модели, давшие обнадеживающие результаты и позволяющие надеяться на успешное использование модели в практике УГМС.

Испытание модели в оперативном режиме планируется провести в 2010 году.

СОДЕРЖАНИЕ

Введение	5
Раздел 1. Исходные данные и постановка задачи	8
Раздел 2. Моделирование притока и расхода воды с помощью непрерывных функций.....	12
Раздел 3. Модель расхода и притока воды	17
Раздел 4. Результаты испытания модели и выводы	22
Заключение	27
Список использованных источников	28

ВВЕДЕНИЕ

Гидрологические исследования с целью прогноза основных гидрологических характеристик неизбежно сопряжены с изучением крупномасштабных процессов в атмосфере, даже если эти характеристики связаны с относительно небольшим регионом. Об этом в частности свидетельствуют работы в области гидрологических прогнозов, а также богатый опыт моделирования расходов и притоков воды в естественные и искусственные водохранилища [2,5]. Таким образом, задача прогноза поведения рек и водоемов автоматически наследует трудности исследования крупномасштабных процессов в атмосфере и долгосрочного прогнозирования. Кроме того, возникают новые трудности, связанные с тем, что вода, как основной объект при изучении поведения рек и водоемов, может присутствовать в процессе погодообразования в различных фазовых состояниях, что ограничивает или делает практически невозможным применение при моделировании каких либо известных законов, управляющих атмосферой. В таких условиях моделирование стоков и притоков воды может осуществляться лишь на основании опыта прошлых лет, путем построения эмпирических моделей на основе доступных данных, которые потенциально могли бы быть задействованы в процессе. В условиях же, когда гидрологическому исследованию подлежат объекты искусственного происхождения, задача еще более усугубляется, поскольку в этом случае временные ряды данных для построения модели катастрофически коротки. В таких условиях приходится решать проблему в упрощенном варианте, оценивая, например, лишь тенденцию, или превышение некоторого критического уровня развития процесса, при минимальных требованиях к оправдываемости прогнозов.

Одна из главных оценок метода представляет собой эмпирическую вероятность ошибки прогноза, не превышающую установленную для данной предсказываемой переменной величину. Определяется такая оценка как процентное отношение числа случаев с соответствующими ошибками к общему числу прогнозов. При этом такая оценка может меняться не только в зависимости от прогнозируемого параметра, но и в зависимости от условий прогнозирования. Так, например, требования к гидрологическим прогнозам в мае существенно отличаются от аналогичных требований к прогнозам в июле.

Принятая система оценок прогнозов в значительной степени определяет подход, который может использоваться при построении модели, поскольку в данном случае не столько важна (или даже совсем не важна) величина ошибки, не превосходящей заданную допустимую величину δ , сколько число ошибок, превосходящих эту величину. Таким образом, не столько важна средняя по выборке ошибка прогнозов, сколько наличие больших ошибок, или ошибок, превосходящих некоторую допустимую величину. Это означает, что мы при построении модели можем постараться уменьшить число больших ошибок, за счет увеличения ошибок маленьких. Заметим, что методы восстановления зависимостей с помощью линейной или нелинейной регрессии, основаны на минимизации именно средней по выборке ошибки прогнозов.

Минимизация числа больших ошибок могла бы быть реализована сведением задачи к линейному программированию, где проблема решается путем минимизации некоторого функционала при линейных ограничениях [1]. Но в этом случае мы можем добиться лишь минимума некоторой суммарной ошибки, превосходящей заданную критическую величину, Однако в нашем случае требуется минимизация именно числа таких ошибок.

Таким образом, принятая система оценивания гидрологических прогнозов и, в частности, прогнозов притока воды в водохранилище, в значительной степени предопределяет подход, который должен использоваться при моделировании.

Однако оценивание гидрологических прогнозов может состоять не только в определении вероятности больших ошибок. Средняя по выборке ошибка также имеет значение, поэтому в практике прогнозов используются также и другие оценки. В некоторых случаях для сравнения качества различных моделей используется отношение средней квадратичной погрешности модели к дисперсии прогнозируемой гидрологической характеристики. В главе 4 настоящего отчета описана модель для прогноза притока и расхода воды, которая минимизирует как число больших ошибок, так и среднюю ошибку по ситуациям, для которых ошибка не превосходит критическую величину. Такая модель предполагает, в качестве первого этапа ее построения, аппроксимацию зависимости с помощью непрерывной функции. Далее, в зависимости от результата аппроксимации, строится разделяющая гиперплоскость, позволяющая выделить критическую область в многомерном пространстве, в которую попадают ситуации с большими ошибками.

Работы по созданию модели долгосрочного прогноза притока воды в Новосибирское водохранилище были выполнены коллективом сотрудников ОСМП в составе Л.Н.

Романова, Е.Г. Бочкаревой, Г.М. Виноградовой. Подготовка гидрологической информации, а также оценка и анализ результатов осуществлялись группой сотрудников отдела гидрологических прогнозов ЦГМС во главе с начальником отдела ГП ГМЦ В.Ф. Богдановой.

РАЗДЕЛ 1

ИСХОДНЫЕ ДАННЫЕ И ПОСТАНОВКА ЗАДАЧИ

Для моделирования поведения уровня воды в Новосибирском водохранилище использовался многолетний архив ситуаций, созданный ранее в лаборатории для построения моделей прогноза погоды различных масштабов. Основные параметры, которые были извлечены из этого архива, составляют данные, которые использовались ранее для построения моделей долгосрочного прогноза температуры и осадков. В эти данные входили среднемесячная температура, суммарные за месяц осадки для 25 станций Западной Сибири, информация об общей циркуляции атмосферы, солнечной активности, эфемериды планет, данные о Луне и др.

Для создания моделей гидрологического прогноза долгосрочный архив был пополнен гидрологическими характеристиками, которые нужны для создания моделей долгосрочного прогноза расхода и притока воды в Новосибирское водохранилище. В архив были включены среднемесячные и декадные расходы воды по створу реки Обь, среднемесячные и декадные данные по притоку воды в водохранилище, среднемесячные расходы воды по шести рекам, впадающим в Обь.

Кроме этого, архив был пополнен данными о максимальных запасах воды в снеге, выпавшем за зиму в горной части р. Обь, бассейне Оби от г. Барнаула до ГЭС, бассейнах рек Бия, Катунь, Алей, Чарыш, Чумыш, Бердь.

В таблице 1 параметры перечислены вместе с указанием года, начиная с которого были сняты соответствующие данные. Как видно из таблицы, длина временного ряда, который можно составить из этих данных, крайне невелика.

В таблице 2 параметры представлены в порядке их расположения в общем архиве ситуаций. Данные измерений, фигурирующие под номерами от 1-го до 400-го, также использовались при моделировании стока и расхода воды, однако в этой таблице они не приводятся.

Гидрологические параметры для моделирования расходов воды к створу р. Обь-Барнаул и притока воды в НВДХ. СНЕГ - обозначает максимальный запас воды в снеге, P – приток, S - расход. D1, D2, D3-данные за 1, 2, 3 декады месяца соответственно. Последний столбец обозначает год начала временного ряда

NN	Река- Пункт	Параметр	год
1	Обь - НВДХ	P	1936
2	Обь- г. Барнаул	S	1922
3	Бия- г. Бийск	S	1901
4	Катунь- с.Сростки	S	1936
5	Чарыш- свх. Чарышский	S	1948
6	Алей - г. Алейск	S	1954
7	Чумыш - р.п. Тальменка	S	1943
8	Вердь- ст. Искитим	S	1960
9	горы (мм)	СНЕГ	1936
10	горы (коэффициент)	СНЕГ	1936
11	Бия	СНЕГ	1948
12	Катунь	СНЕГ	1948
13	Чарыш	СНЕГ	1949
14	Алей	СНЕГ	1948
15	Обь до Барнаула	СНЕГ	1948
16	Чумыш	СНЕГ	1948
17	Вердь	СНЕГ	1948
18	Обь до ГЭС	СНЕГ	1948
19	Обь-г. Барнаул	D1S	1936
20	Обь-г. Барнаул	D2S	1936
21	Обь-г. Барнаул	D3S	1936
22	Обь - НВДХ	D1P	1936
23	Обь - НВДХ	D2P	1936
24	Обь - НВДХ	D3P	1936

Приведенные параметры были сняты за все месяцы года и все они были задействованы в той или иной степени при моделировании. Используя эти параметры, предполагалось построить модели долгосрочного прогноза расхода воды в створе Обь- Барнаул, а также модель долгосрочного прогноза притока воды в Новосибирское водохранилище. Заблаговременность при этом предполагалась двум-трем месяцам.

Для каждого из прогнозируемых месяцев (апрель-сентябрь) модель строится отдельно, поскольку условия, при которых формируется водный режим региона, существенно различаются. При этом требуется строить отдельное правило прогнозирования не только для отдельных месяцев, но и для каждого из двух прогнозируемых элементов (расход, приток). В условиях, когда нет никакой другой информации, кроме исходных данных измерений, построение может быть реализовано лишь с помощью обучающих систем, точнее говоря, с помощью обучения с учителем.

Номера гидрологических характеристик
в порядке их расположения в архиве

ПРИТОК	Обь - НВДХ 402			
РАСХОД	Обь-Барнаул 403	Бия 404	Катунь 405	Чарыш 406
РАСХОД	Алей 407	Чумыш 408	Бердь 409	
СНЕГ	Горы (мм) 410	Горы (коэф.) 411	Бия 412	Катунь 413
СНЕГ	Чарыш 414	Алей 415	Обь до Барн. 416	Чумыш 417
СНЕГ	Бердь 418	Обь до Гэс 419		
РАСХОД (декадники)	Барнаул D1S 420	Барнаул D2S 421	Барнаул D3S 422	
ПРИТОК (декадники)	НВДХ D1P 423	НВДХ D2P 424	НВДХ D3P 425	

Такая возможность представляется нам благодаря наличию в архиве значений расхода и притока воды, которые при построении модели должны выступать в качестве известных значений прогнозируемых гидрологических параметров.

Если материал для обучения представить в виде ряда ситуаций $x=(x^1, \dots, x^N)$, и ряда скаляров y , представляющих собой соответствующие значения прогнозируемого элемента,

$$x_1, \dots, x_N \quad (2.1)$$

$$y_1, \dots, y_N,$$

то задача будет состоять в том, чтобы построить такую функцию $f(x)$ по этому материалу, чтобы эта функция удовлетворяла определенным статистическим критериям. В нашем случае эта функция должна удовлетворять критерию минимума функционала

$$J = \sum_i p(x^i) f(x^i), \quad (2.2)$$

называемому средним риском. Здесь $p(x) \leq 1$ - известная функция координат, а минимум ищется по всем непрерывным функциям $f(x)$ заданного класса на всем исходном множестве ситуаций (2.1). Таким образом, минимум функционала (2.2) определяет не только вид функции f , но и совокупность параметров, от которых она должна зависеть.

Непрерывным условием постановки задачи обучения с целью определения неизвестной зависимости является случайность и независимость поступления ситуаций, осуществляемые согласно некоторому фиксированному распределению. Однако на практике, текущие ситуации не поступают случайно и независимо, а берутся из некоторого доступного временного интервала, причем длина такого интервала может быть крайне ограничена. Рассматривая модели для каждого прогнозируемого месяца отдельно, мы тем самым еще более сокращаем используемый ряд, что не может не оказывать отрицательных последствий на результат моделирования. Однако подобная детализация позволяет существенно упростить выборочное распределение, и таким образом, существенно ограничить область поиска аппроксимирующей функции.

РАЗДЕЛ 2

МОДЕЛИРОВАНИЕ ПРИТОКА И РАСХОДА ВОДЫ С ПОМОЩЬЮ НЕПРЕРЫВНЫХ ФУНКЦИЙ

Простейшей моделью, которая может быть построена по данным, - это линейная модель. Такая модель использует для предсказания линейную функцию

$$f(x) = a_0 + a_1 x_1 + \dots + a_n x_n,$$

обладающую свойством на исходном множестве ситуаций (2.1) минимизировать функционал

$$\sum_{i=1}^N G_i (y_i - f(x))^2.$$

Формальное построение такой функции не составляет большого труда. Трудности возникают тогда, когда встает вопрос о выборе аргументов, ее определяющих. При построении моделей стоков и расходов воды в регионе Новосибирского водохранилища, выбор параметров, или аргументов функции, осуществлялся с помощью скользящего контроля [3], который по существу есть один из подходов, используемых для минимизации среднего риска. Суть этого подхода состоит в следующем. Обозначим через $f_l(x)$ функцию, полученную путем минимизации функционала

$$\sum_{i=1}^N G_i (y_i - f_l(x))^2$$

на множестве ситуаций (2.1), из которого исключена ситуация с индексом l . Путем поодиночного исключения из множества (2.1) всех N ситуаций, таким образом, может быть получено N функций

$$f_0(x), \dots, f_{N-1}(x)$$

каждую из которых определяет $N-1$ ситуация.

Для организации процедуры с целью получения оценки $f(x)$ вычислим значения

$$\delta_i = y_i - f_0(x) \quad (i=1, \dots, N),$$

которые, в соответствии с терминологией [6], будем называть псевдозначениями скользящего контроля. Совокупность псевдозначений, или вектор-столбец $\delta = \{\delta_1, \dots, \delta_N\}$ определит среднеквадратичную ошибку скользящего контроля

$$\Delta = \frac{1}{N} \sqrt{\sum_i \delta_i^2},$$

которая может служить характеристикой качества аппроксимации неизвестной зависимости с помощью функции $f(x)$. Вычисляя, таким образом, Δ для всех подмножеств из исходных параметров, мы можем выбрать такое подмножество, которое соответствует наименьшему значению Δ , или значению скользящего контроля.

На рис.1 изображена кривая поведения ошибок скользящего контроля в зависимости от размерности упорядоченного пространства исходных параметров. Для сравнения приведена кривая поведения средней квадратичной ошибки, вычисленной по всей выборке. Такая кривая всегда располагается ниже на графике и всегда монотонно убывает вместе с увеличением размерности пространства. Кривая же ошибок скользящего контроля, как правило, имеет один минимум, который и должен определять оптимальную размерность пространства. В данном конкретном случае минимум достигается при числе параметров, равном 5. Данные, по которым были получены кривые на рис.1, использовались для построения правила предсказания расхода воды в мае с заблаговременностью в 4 месяца. Однако аналогичное поведение соответствующих кривых наблюдается и на других месяцах, причем число информативных параметров, как правило, не превосходит 10-12 единиц, а в среднем оно колеблется около 6.

На рисунке 2 изображены аналогичные кривые, построенные для случая, когда прогнозировался приток в Новосибирское водохранилище. Модель строилась для прогноза притока в январе на четыре месяца вперед. Здесь также имеется достаточно ярко выраженный минимум скользящего контроля, который и определяет оптимальную размерность 8.

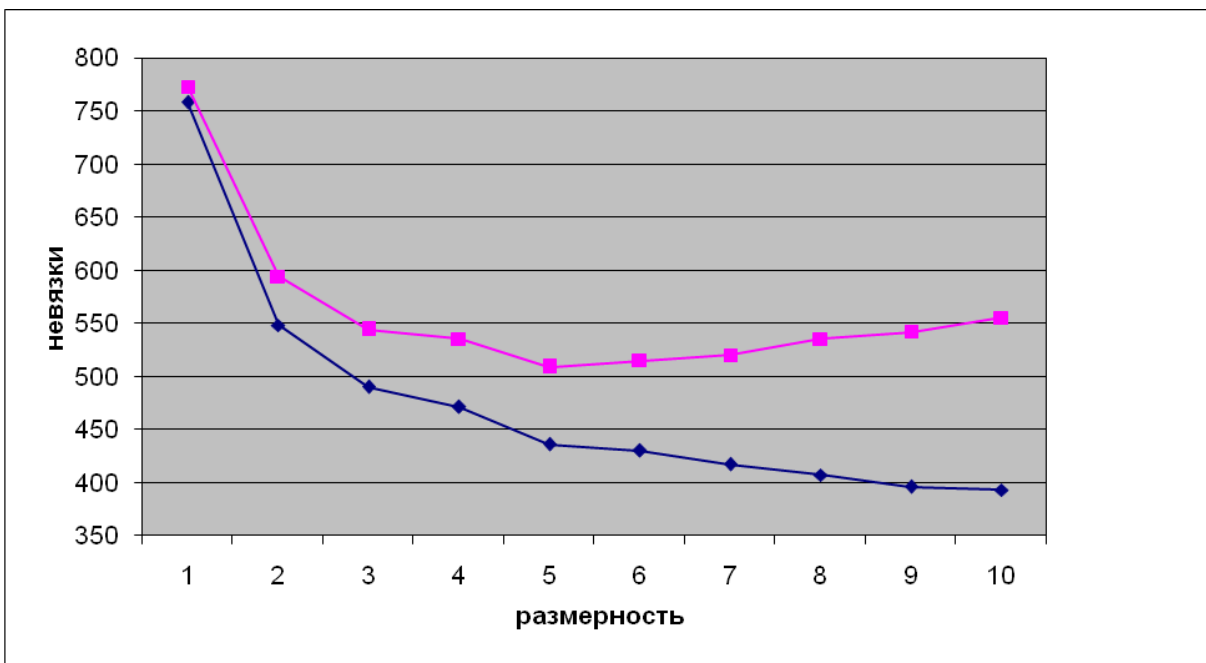


Рис.1 Кривые поведения ошибок при моделировании **расхода** воды. Нижняя кривая соответствует среднеквадратичным ошибкам для соответствующих размерностей, верхняя, – ошибкам скользящего контроля

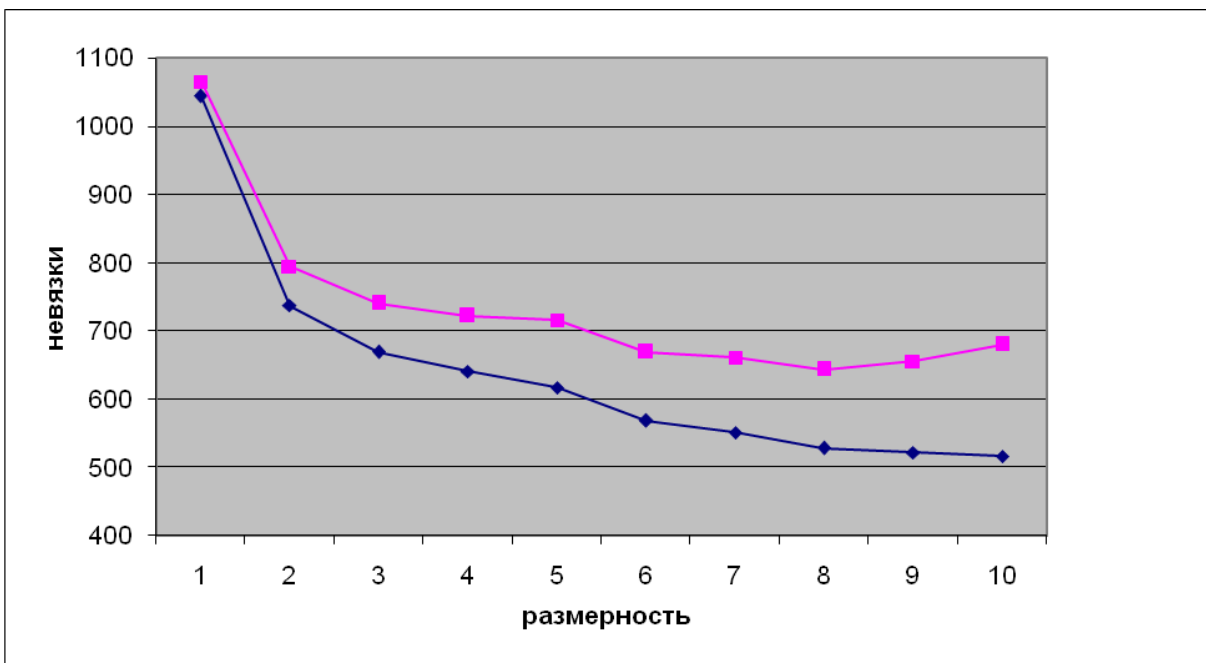


Рис.2 Кривые поведения ошибок при моделировании **притока** воды. Нижняя кривая соответствует среднеквадратичным ошибкам для соответствующей размерности, верхняя – ошибкам скользящего контроля

Учитывая сложность процессов, протекающих в атмосфере, от которых зависит гидрологический прогноз, надежда получить какой-либо положительный результат, основываясь только на линейных методах, чрезвычайно слаба. Поэтому все подходы при прогнозе гидрологических характеристик реализуются либо с помощью кусочно-линейной аппроксимации, либо с помощью полиномиальной аппроксимации.

Для поиска решения в случае нелинейной аппроксимации следует осуществить преобразование

$$\{x^1, \dots, x^n\} \rightarrow \{z^1, \dots, z^m\}, \quad (3.1)$$

где слева расположены исходные параметры, а справа - преобразованные. В частном случае z^1, \dots, z^m могут быть значениями одночленов полинома некоторой степени, взятых от переменных x^1, \dots, x^n . В последнем случае предполагается, что некоторые степени переменных могут принимать и нулевые значения. Последующее применение линейного восстановления к преобразованным с помощью (3.1) параметрам позволяет определить нелинейную зависимость прогнозируемой величины y от некоторой информативной совокупности параметров из исходного множества параметров x^1, \dots, x^n .

Задачу поиска неизвестной функции в классе полиномов путем минимизации среднего риска можно по существу интерпретировать как разложение этой функции в степенной ряд, с последующим определением числа членов этого разложения. Если при этом удастся достигнуть глобального минимума оценки среднего риска, то соответствующее полиномиальное разложение будет иметь наибольшую скорость сходимости, и окончательное представление функции после отбрасывания неинформативных слагаемых будет иметь наиболее простой вид. Если обозначить функцию, подлежащую восстановлению через $F(x)$, то полином можно записать в виде

$$F(x) \approx a_0 + \sum_i \lambda_i q_i,$$

где λ_i - коэффициенты полинома, а q_i - одночлены, степени которых могут принимать и нулевые значения.

Нелинейный подход с целью прогноза может быть реализован также с помощью кусочно-линейной аппроксимации. Одну из возможностей реализации такой аппроксимации представляет собой кусочно-линейная двухуровневая модель, предназначенная для прогноза непрерывно изменяющихся метеорологических элементов. Такая модель строится следующим образом: на первом этапе строится несколько линейных функций для аппроксимации зависимости при различных заблаговременностях прогноза. Число таких заблаговременностей зависит от длины исходного ряда и должно определять в конечном итоге максимально допустимый интервал предыстории процесса. Далее, на втором уровне, исходные данные преобразуются с помощью полученных на первом этапе функций в некоторые обобщенные параметры, которые затем используются для построения функции второго уровня, определяющей окончательный прогноз. При этом все линейные функции, как первого, так и второго уровня, строятся путем минимизации среднего риска.

Главным достоинством такой модели является возможность с ее помощью не только определять интервал предыстории процесса, но и проследить развитие процесса, предшествовавшего прогнозируемой погодной ситуации. Другим важным достоинством модели является ее способность усваивать практически неограниченное количество регулярных данных, поступающих через равные временные интервалы.

Двухуровневая кусочно-линейная модель была реализована для целей прогноза расхода воды в створе р. Обь-г.Барнаул и притока воды в Новосибирское водохранилище, и уже несколько лет используется в практике СибУГМС в качестве вспомогательной. В рамках выполнения темы 8.78 раздела 2 модель была модифицирована, и вновь была подвергнута испытаниям с участием дополнительных гидрологических параметров.

РАЗДЕЛ 3

МОДЕЛЬ РАСХОДА И ПРИТОКА ВОДЫ

При прогнозировании расхода и притока воды приходится исходить из двух критериев. С одной стороны, хотелось бы, чтобы ошибка прогноза была в каждом конкретном случае невелика, с другой - чтобы число ошибок, превосходящих некоторую конкретную величину, была как можно меньше. Применение таких взаимно противоречивых критериев существенно усложняет как построение модели, так и саму модель. Кроме того, применение двойного критерия делает задачу построения логически согласованной модели [2], при достаточно представительной исходной выборке данных, практически нереализуемой. Однако, если выборка не велика, можно попытаться построить модель, которая будет близка к логически согласованной модели. Напомню, что логически согласованное построение модели - это когда критерий построения модели совпадает с критерием ее оценивания. Так, например, если модель строится с помощью регрессии, то оценивать на независимом материале надо среднюю квадратичную ошибку. Если же нужно минимизировать среднюю величину больших ошибок, то и такую же среднюю величину больших ошибок надо минимизировать и при построении модели.

Заметим, что в некоторых случаях удается построить логически согласованную модель и при наличии сложных критериев. Например, если при прогнозировании некоторой непрерывной величины требуется минимизировать сумму ошибок, превосходящих некоторое ее критическое значение, то такая модель может быть построена с помощью симплекс-метода [1]. Однако число «больших» ошибок не обязательно при этом будет минимальным. Для минимизации же числа больших ошибок требуется другой подход, принципиально отличный от того, который используется для минимизации суммарной ошибки, поскольку в этом случае функция, подлежащая минимизации, не является непрерывной.

Если требуется минимизировать только число больших ошибок, без каких-либо дополнительных требований, то в этом случае можно попытаться выделить некоторую область метрического пространства, в которую попадают все ошибочные ситуации, и построить для этих ситуаций некоторое, отличное от первоначального, правило прогнозирования.. При этом исходная выборка для второго правила должна быть статистически значима, т.е. ошибочных ситуаций должно быть по крайней мере не меньше,

чем размерность результирующего пространства. В этом случае, начиная процесс упорядочения для построения второго правила прогнозирования, мы можем надеяться получить соотношение числа параметров и ситуаций по крайней мере как один к двум, или один к трем.

Пусть, например, имеется выборка исходных ситуаций x_1, x_2, \dots, x_N и вектор значений $y = y_1, y_2, \dots, y_N$, представляющих собой соответствующие известные значения прогнозируемой величины y . Используя эту выборку, построим функцию $\tilde{\varphi}(x)$, удовлетворяющую условию

$$\min_{\varphi} \sum_{i=1}^N (\varphi(x_i) - y_i)^2, \quad (4.1)$$

где минимум функционала в общем случае ищется по всем классам задаваемых функций и всем подгруппам из исходно задаваемых параметров. Однако, для простоты написания будем предполагать, что минимум (4.1) ищется по коэффициентам исходно заданной линейной функции. Иначе говоря, найденная функция $\tilde{\varphi}$ есть линейная функция тех же исходных параметров.

Пусть далее величина δ обозначает верхнее значение ошибки, выше которого прогноз с помощью модели (4.1) считается неоправдавшимся. В этом случае всю исходную выборку можно разделить на две части: первая часть - это все ситуации X (назовем эту выборку A), для которых ошибка прогноза мала ($\varepsilon_i \leq \delta$), и вторая часть, - это когда все ситуации из X (назовем их множеством B), превосходят по величине δ ($\varepsilon_i \geq \delta$).

Если число ситуаций множества B , достаточно велико, то по этим ситуациям также может быть построена аппроксимирующая функция ϕ , минимизирующая средний квадрат ошибки

$$\min_{\phi} \sum_{i \in B} (\phi(x_i) - y_i)^2. \quad (4.2)$$

Таким образом, будем иметь следующее правило для прогнозирования: если ситуация x_p принадлежит множеству A , то

$$y = \tilde{\varphi}(x_p), \quad x_p \in A,$$

а если множеству B , то

$$y = \tilde{\phi}(x_p), \quad x_p \in B$$

Разбивая таким образом исходные множества на два подмножества и строя аппроксимирующие функции на каждом из этих подмножеств, мы сможем избежать больших ошибок, по крайней мере на исходной выборке ситуаций. По существу, мы тем самым строим два цилиндра в многомерном пространстве

$$(y - \tilde{\varphi}(x))^2 = \delta^2$$

$$(y - \tilde{\phi}(x))^2 = \delta^2$$

так, что все точки исходного множества (или их подавляющее большинство) находятся внутри этих цилиндров. Цилиндры, хотя и могут пересекаться, тем не менее общих точек из исходного множества они не имеют. Первый цилиндр по самому построению покрывает все точки множества A , второй же цилиндр покрывает далеко не все точки этого множества.

Для того чтобы знать каким правилом воспользоваться при поступлении очередной ситуации x , не принадлежащей исходному множеству, разделим исходное множество на два класса с помощью гиперплоскости $(ax) = c$, причем так, чтобы множество точек A находилось по одну его сторону, а множество точек B – по другую. Иначе говоря, найдем такой вектор a и коэффициент c , что для любого x_i имеют место неравенства

$$\begin{aligned} (ax_i) &\leq c && \text{если } x_i \in A, \\ (ax_i) &\geq c && \text{если } x_i \in B. \end{aligned}$$

Если существует гиперплоскость $(ax) = c$, разделяющая множества A и B , то мы можем естественным образом поставить в соответствие этим множествам классы точек. Например, к первому классу отнесем точки полупространства $(ax) \leq c$, а ко второму классу точки, попадающие в полупространство $(ax) \geq c$.

В нашем конкретном случае к первому классу будут относиться точки, на которых ошибки первой регрессии малы ($\varepsilon_i \leq \delta$), ко второму – точки, на которых эти ошибки достаточно велики ($\varepsilon_i > \delta$). Заметим однако, что при таком построении ошибка регрессии, соответствующая первому классу будет смещенной, поскольку среднее значение для первой регрессии оценивается по всем точкам $A+B$. Для того, чтобы исключить это несоответствие, надо скорректировать функцию φ , взяв в качестве исходной выборки только ситуации множества A . Результат от этого может только улучшиться, по крайней мере в смысле минимума среднеквадратического отклонения, поскольку из исходного множества удаляются все соответствующие большим ошибкам ситуации.

Резюмируя, можно выделить следующие основные этапы построения модели.

1. Построение функции $\varphi(x)$ по всем исходным ситуациям.
2. Разделение исходной выборки на две группы, A и B , в зависимости от величины соответствующих им ошибок.
3. Построение функции $\phi(x)$ по ситуациям группы B .
4. Построение гиперплоскости, разделяющей множества A и B .
5. Коррекция функции $\varphi(x)$ путем исключения из исходной выборки ситуаций, соответствующих большим ошибкам.

Текущее же использование модели предполагает лишь две операции.

1. Для вновь поступившей ситуации x_k вычисляем скалярное произведение (αx_k) , и определяем, по какую сторону от гиперплоскости $(\alpha x) = c$ эта ситуация расположена.
2. В зависимости от принадлежности к классу текущей ситуации x_k вычисляем одно из двух значений $\varphi(x_k)$, или $\phi(x_k)$, которое и определяет прогноз.

Рассмотренная конструкция не обязательно должна приводить к существенному повышению точности прогнозов: все зависит в данном случае от исходного распределения ситуаций и от исходной выборки ситуаций. Однако никакие наши действия не должны были привести к ухудшению прогнозов, в сравнении с прогнозами с помощью одной функции, поскольку все эти действия были направлены на локализацию больших ошибок, и если нам удалось это сделать хотя бы в какой-то степени, то это уже должно давать эффект.

Если же этого нам сделать не удалось, т.е. не удалось выделить некоторую пространственную область (цилиндр), в котором располагаются ситуации с большими ошибками, то автоматически функционирует лишь первый этап построения, когда решение принимается лишь на основе одной аппроксимирующей функции, построенной по всему материалу.

Описанная в настоящем разделе модель фактически использует три уровня преобразования исходных данных: на первом строится регрессия, на втором разделяющая гиперплоскость, и далее по выделенным критическим ситуациям вновь строится регрессия. Естественно было бы классифицировать такую модель как трехуровневую кусочно-альтернативную модель. В дальнейшем же для краткости будем называть эту модель кусочно-альтернативной (модель 2), в отличие от кусочно-линейной двухуровневой модели, которую будем называть кусочно-линейной моделью (модель 1).

Описанная модель, использующая для прогноза разделяющую гиперплоскость, достаточно сложна для построения. Однако, только таким образом мы, оставаясь в рамках чисто формального подхода, могли учитывать те специфические климатические и географические условия, которые используются синоптиками при неформальном прогнозировании.

РАЗДЕЛ 4

РЕЗУЛЬТАТЫ ИСПЫТАНИЯ МОДЕЛИ И ВЫВОДЫ

С использованием описанных выше инструментов были проведены многочисленные эксперименты по восстановлению зависимостей с целью построения моделей долгосрочного прогноза притока и расхода воды в районе Новосибирского водохранилища. При этом использовались в качестве исходной информации многолетние ряды данных, пополненные новыми физическими параметрами, призванными наиболее полно отражать гидрологические процессы в регионе.

По этим данным были построены модели притока и расхода воды в регионе для различных прогнозируемых месяцев, и были проведены сравнения с прежними результатами, полученными в условиях отсутствия в архиве важных гидрологических характеристик. Другим важным фактором, который должен был повлиять на качество прогнозов, является расширение области поиска аппроксимирующей функции, описанное в предыдущем разделе.

Для этой цели использовались временные ряды данных, начиная 1936 года и до 2008 года. При этом каждый временной слой ряда включал в себя не только параметры таблицы 2, но и группу параметров из общего архива, который ранее использовался для целей долгосрочного прогноза других элементов погоды.

В табл. 3 приведены оценки оправдываемости прогнозов притока и расходов воды в регионе для кусочно-линейной модели (модель 1) и кусочно-альтернативной модели (модель 2). Оценка прогнозов проводилась в соответствии с «Наставлением по службе прогнозов», разд. 3, ч. 1 (Ленинград, 1962). Прогнозы по модели 1 получены при использовании модели в оперативной практике в 2004-2008 годах. Здесь следует отметить, что нулевая оправдываемость прогнозов в 2008 году была связана с ошибкой при переводе модели на новую структуру базы данных. Прогнозы по модели 2 получены в результате авторских испытаний. Как видно из таблицы 3, оправдываемость прогнозов по модели 2 превосходит оправдываемость прогнозов по модели 1. При этом преимущество сохраняется как при прогнозе расхода воды в створе Барнаул-Обь, так и при прогнозе притока воды в Новосибирское водохранилище. Для сравнения в таблице 4 приведены оправдываемости официальных прогнозов притока и расходов воды, полученные на том же материале. Как

видно из таблицы 3 и таблицы 4, оправдываемость прогнозов ГМЦ выше оправдываемости прогнозов по модели 1 и несколько ниже оправдываемости прогнозов по модели 2.

Таблица 3

Оправдываемость (%) прогнозов расходов воды р. Обь - г. Барнаул и притока воды в Новосибирское водохранилище с помощью кусочно-линейной модели (Модель 1) и прогнозов с помощью кусочно-альтернативной модели (Модель 2) за апрель-сентябрь 2004-2008гг.

Год	Модель 1		Модель 2	
	Расход воды р. Обь-Барнаул	Приток воды в Новосибирское водохранилище	Расход воды р. Обь-Барнаул	Приток воды в Новосибирское водохранилище
2004	67	67	83	83
2005	83	83	83	100
2006	50	50	83	67
2007	67	50	67	67
2008	0	0	83	83
Среднее	53	50	80	80

Таблица 4

Оправдываемость (%) прогнозов расходов воды р. Обь – г. Барнаул и притока воды в Новосибирское водохранилище с помощью кусочно-линейной модели (Модель 1), и прогнозов ГМЦ, за апрель-сентябрь 2004-2008гг

Год	Модель 2		ГМЦ	
	Расход воды р. Обь-Барнаул	Приток воды в Новосибирское водохранилище	Расход воды р. Обь-Барнаул	Приток воды в Новосибирское водохранилище
2004	83	83	67	83
2005	83	100	83	83
2006	83	67	83	67
2007	67	67	50	67
2008	83	83	67	83
Среднее	80	80	70	77

Подробную картину оправдываемости прогнозов по модели 2 и прогнозов ГМЦ для каждого из 6 прогнозируемых месяцев можно проследить по таблицам 5-8.

Таблица 5

Оправдываемость (%) прогнозов **притока воды** в Новосибирское водохранилище
(**модель 2**)

Год	апрель	май	Июнь	июль	август	сентябрь	Среднее
2004	100	0	100	100	100	100	83
2005	100	100	100	100	100	100	100
2006	0	100	100	100	100	0	67
2007	100	100	0	100	100	0	67
2008	100	100	100	0	100	100	83
Среднее	80	80	80	80	100	60	80

Таблица 6

Оправдываемость (%) прогнозов **притока воды** в Новосибирское водохранилище
(**прогнозы ГМЦ**)

Год	Апрель	май	Июнь	июль	август	Сентябрь	Среднее
2004	100	0	100	100	100	100	83
2005	100	100	0	100	100	100	83
2006	100	0	0	100	100	100	67
2007	0	100	0	100	100	100	67
2008	100	100	100	100	100	0	83
Среднее	80	60	40	100	100	80	77

Таблица 7

Оправдываемость (%) прогнозов **расходов воды** в створе реки Обь-Барнаул
(модель 2)

Год	Апрель	май	июнь	Июль	август	Сентябрь	Среднее
2004	100	0	100	100	100	100	83
2005	100	100	100	100	100	0	83
2006	100	100	100	100	100	0	83
2007	100	100	0	0	100	100	67
2008	100	100	100	100	100	0	83
Среднее	100	80	80	80	100	40	80

Таблица 8

Оправдываемость (%) прогнозов **расхода воды** в створе реки Обь-Барнаул
(прогнозы ГМЦ)

Год	Апрель	май	июнь	Июль	август	Сентябрь	Среднее
2004	0	0	100	100	100	100	67
2005	100	100	0	100	100	100	83
2006	100	0	100	0	100	100	83
2007	0	100	0	0	100	100	50
2008	0	100	100	100	100	0	67
Среднее	40	60	60	80	100	80	70

Детального анализа, что и в какой мере послужило причиной повышения качества прогнозов с помощью модели 2 по сравнению с прогнозами с помощью модели 1, не

проводилось. Однако было очевидно, что информация о снеге была очень важна при прогнозе в период половодья. Параметры, отражающие эту информацию, отбирались среди первых при моделировании, как притока, так и расходов воды в мае и июне месяцев.

Как видно из таблиц 5-8, ошибки прогнозов с помощью модели 2 и ошибки официальных прогнозов часто совпадают. Это можно объяснить двумя причинами: первое - это недостаточность «опыта», обусловленная короткими временными рядами; и второе - это неполное описание исходных ситуаций, по которым осуществляется прогнозирование. Если первая причина еще долгие годы может представлять серьезное препятствие на пути создания надежных прогнозов, то вторую причину можно в значительной степени преодолеть. Для этого надо только более детально и аккуратно использовать уже существующие исходные данные.

Наибольшая оправдываемость прогнозов с помощью модели имеет место в августе, наименьшая - в сентябре. Стопроцентная оправдываемость для всех пяти лет сохраняется только для августа месяца.

Один из важных выводов, который может быть сделан из экспериментов по аппроксимации неизвестных функций, состоит в том, что при прогнозе стока полиномиальная аппроксимация не дает никаких преимуществ по сравнению с линейной. Причина этого может заключаться в том, что при прогнозе на большие сроки, и к тому же по характеристикам, осредненным по большим временным интервалам, возможно лишь грубое описание происходящих при этом процессов, даже если данные при этом достаточно точны, и имеется достаточное их количество. По этой причине описание процессов, в условиях применения критерия минимума среднего риска, упрощается, и, как правило, сложная нелинейная зависимость вырождается в простую линейную. Однако это не означает, что следует пренебрегать использованием аппарата нелинейной аппроксимации; при восстановлении диагностических связей, даже в условиях осредненных по времени параметров, полезные нелинейные зависимости могут восстанавливаться достаточно точно, Это частично подтверждается экспериментами по восстановлению диагностических зависимостей притока и расходов воды. Дальнейший же прогресс в области использования полиномиальной аппроксимации может быть связан лишь с включением данных прямого измерения, или данных с меньшим интервалом пространственного или временного осреднения.

ЗАКЛЮЧЕНИЕ

В результате выполнения научно-исследовательских работ по теме 8.78 (раздел 2) была создана модель долгосрочного прогноза притока воды к Новосибирскому водохранилищу и расходов воды в створе р. Обь – г. Барнаул для 2-3 кварталов. Проведены авторские испытания модели на материале 2004-2008 годов. Средняя оправдываемость прогнозов составила 80% и при одинаковой заблаговременности превысила оправдываемость оперативных прогнозов ГМЦ.

На Техническом Совете ГУ "Новосибирского ЦГМС-РСМЦ" было принято решение о проведении оперативных испытаний метода в 2010 году.

В то же время, результаты, изложенные в настоящем отчете, ни в коей мере не претендуют на полноту исследования проблемы, поскольку резервы повышения точности прогнозов полностью не исчерпаны. Эти резервы сохраняются не только в возможностях увеличения объема исходных данных или улучшения их качества, но и за счет улучшения качества аппроксимации путем достижения более глубокого экстремума функционала качества, называемого средним риском. Обе эти возможности могут реализовываться как независимо, так и в сочетании друг с другом. Однако наибольшего эффекта можно ожидать, когда обогащение потенциальной модели новым физическим содержанием гармонично сочетается с достижением более глубокого минимума оценки среднего риска.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Дуда Р., Харт П. Распознавание образов и анализ сцен. М: Мир, 1976.
2. Корень В.И., Бельчиков В.А. Методические указания по использованию методов краткосрочных прогнозов ежедневных расходов (уровней) воды для речных систем на основе математических моделей. Л.: Гидрометеиздат, 1989. 176 с.
3. Романов Л.Н. Минимизация риска и восстановление пропусков в атмосферных данных // Сиб. журн. вычисл. математики. 2009. Т.12, № 2.
4. Романов Л.Н. О выборе моделей для статистического прогноза // Труды ЗапСибНИГМИ. 1990. Вып. 93.
5. Руководство по гидрологическим прогнозам. Выпуск 2. Краткосрочные прогнозы расхода и уровня воды на реках. Л.: Гидрометеиздат, 1989. 245 с.
6. Наставление по службе прогнозов, разд.3, часть 1. Ленинград, 1962.
7. Gray H.L., Schucany W.R. The generalized jackknife statistic. N.Y., M. Dekker, Inc., 1972.